Check for updates

Received: 25 March 2021 | Revised: 31 January 2022 | Accepted: 21 February 2022

DOI: 10.1002/asi.24636

RESEARCH ARTICLE

JASIST WILEY

Free access to scientific literature and its influence on the publishing activity in developing countries: The effect of Sci-Hub in the field of mathematics

Kilian Buehling[1] | **Matthias Geissler**[2] | **Dorothea Strecker**[3]

[1]Research Group Knowledge and Technology Transfer, Technische Universität Dresden, Dresden, Germany

[2]Digitalization and Innovation Section, Rationalisierungs- und Innovationszentrum der Deutschen Wirtschaft e. V. RKW Kompetenzzentrum, Eschborn, Germany

[3]Berlin School of Library and Information Science, Humboldt University Berlin, Berlin, Germany

Correspondence
Kilian Buehling, Research Group Knowledge and Technology Transfer, Technische Universität Dresden, Muenchner Platz 2-3, D-01187 Dresden, Germany.
Email: kilian.buehling@tu-dresden.de

Funding information
Bundesministerium für Bildung und Forschung, Grant/Award Number: 01PU17007B

Abstract

This paper investigates whether free access to scientific literature increases the participation of under-represented groups in scientific discourse. To this end, we aggregate and match data tracing access to Sci-Hub, a widely used black open access (OA) repository or *shadow library*, and publication data from the Web of Science (WoS). We treat the emergence of Sci-Hub as an exogenous event granting relatively unrestricted access to publications, which are otherwise hidden behind a paywall. We analyze changes in the publication count of researchers from developing countries in a given journal as a proxy for general participation in scientific discourse. Our results indicate that in the exemplary field of mathematics, free access to academic knowledge is likely to improve the representation of authors from developing countries in international journals. Assuming the desirability of greater international diversity in science (e.g., to generate more original work, reproduce empirical findings in different settings, or shift the research focus toward topics that are overlooked by researchers from more developed countries), our findings lend evidence to the claim of the OA movement that scientific knowledge should be *free* and widely distributed.

1 | INTRODUCTION

Very much akin to trade, communication, and travel, academic science has become an increasingly globalized endeavor. International co-authorship, for example, has been on the rise in all scientific disciplines since the 20th century (Coccia & Wang, 2016; Davidson Frame & Carpenter, 1979; Fortunato et al., 2018).[1] Generally speaking, increased internationalization in academia may be desirable because collaborations on a global scale can tap into a wider creativity pool (Huang et al., 2012) and produce more impactful research (Persson et al., 2004; Wagner et al., 2019). However, a closer look reveals that countries representing central nodes of global collaboration networks continue to be from Europe and North America, except for a few emerging economies, most notably China, India, and Brazil (Beauchesne, 2011; Gui et al., 2019). Overall, the representation of authors from developing countries in internationally published scholarly literature remains low (Dahdouh-Guebas et al., 2003; Demeter, 2020; Kutlača et al., 2014).

Aside from authorship networks, inequalities on a global scale also exist regarding access to scientific literature. For decades, scholarly findings were distributed

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
© 2022 The Authors. *Journal of the Association for Information Science and Technology* published by Wiley Periodicals LLC on behalf of Association for Information Science and Technology.

internationally via specialized publishers who organized the peer-review process and invested in a distribution network in exchange for subscription fees often paid by libraries (Brown, 2015; Ware & Mabe, 2015). This *closed access* model has been criticized because it may give rise to club good situations (excludability in access to a good, which is non-rivalrous in consumption). These are likely to favor certain organizations that have the resources to pay for access (James, 2020), although theory usually conceives of academic knowledge as a public good (Stiglitz, 1999). This situation may distort research agendas, create regional or global lock-ins in research trajectories, or lead to an inefficient allocation of resources. Furthermore, promising research areas, theories, or solutions to scientific problems may remain unexplored if large parts of a global scientific community are prevented from contributing to current discussions by putting the explicit knowledge and research results of their peers behind a paywall. Similar to other club good situations, the system is likely to contribute to the continuous reproduction of existing structures and distribution of scientific activity, but also to the direction of possible scientific advances. This may lead "outsiders," most notably developing countries, to be trapped in an everlasting process of "catching up."

With the advent of digital technology in general and the internet in particular, criticism of the prices and market structures for producing and distributing scientific knowledge has been growing over the last two decades (Hoey & Todkill, 2002; Houghton, 2001; Van Noorden, 2013a). While the costs of physical distribution of research articles declined to a negligible amount (Ware & Mabe, 2015), new costs arose for publishers, for example, to establish digital infrastructure or paywalls (Grossmann & Brembs, 2019). However, Grossmann and Brembs (2019) estimate that the pure costs of publication only amount to 5–25% of the charged subscription fees per publication in paywalled journals (depending on a journal's quality control measures and rejection rates, excluding management costs). Although proponents of the traditional closed access model (with users charged for access) have put forward favorable arguments (notably quality control), some researchers and politicians are stressing that scientific knowledge should be available to a wide audience for the greater good of society.[2] The open access (OA) movement has gained a lot of momentum, especially since funding organizations have started to explicitly demand this mode of publication (e.g., the National Science Foundation (2016), the Japan Ministry of Education, Culture, Sports, Science and Technology (2012), or the European Commission (2017)).[3] While the exploration of suitable models is in full gear, a common promise of OA is the reduction of club good problems by unconditionally granting all interested individuals (researchers and others) the ability to access knowledge derived from state-of-the-art research.

However, the promise of OA to free knowledge, increased diversity and efficiency in knowledge production, and that it leads to a more even distribution of scientific activities can be challenged from several perspectives. First, types of OA relying on author payment schemes may simply shift the timing of compensation for the publishers from "pay-when-reading" to "pay-when-publishing", therefore excluding the same individuals, who were previously excluded from accessing scholarly literature, from publishing (Papin-Ramcharan & Dawe, 2006; Pourret et al., 2021; Sengupta, 2021). Second, knowledge production, especially basic research, is a risky and costly endeavor associated with large investments in infrastructure, human capital, and material in several fields of research. Free access to scientific knowledge may have little further consequences if researchers who were previously financially unable to access the latest literature do not possess the complementary assets to undertake cutting-edge research themselves. Third, seminal work by Polanyi (1958) and subsequent work on tacit dimensions of knowledge production indicate that disclosure of a result or discovery, howsoever detailed, will usually not give a full account of the means to reproduce it or readily put it into use (Agrawal, 2006). Moreover, research at the current knowledge frontier is a highly specialized task that usually requires personal investment and individual training. Individuals not possessing the necessary absorptive capacity (Cohen & Levinthal, 1990; Forero-Pineda & Jaramillo-Salazar, 2002) will not be able to understand publications, even if they are freely available. This may also include very basic obstacles, such as language barriers for researchers from non-English speaking countries (Blicharska et al., 2017).

Although progress has been made over the last several years, publishers and scientists alike remain relatively slow in adopting the new publication models developed by the OA movement (Hampl et al., 2021; Himmelstein et al., 2018; Van Noorden, 2013b). In the meantime, illegal practices, known as black OA, literature pirating, or shadow libraries, have emerged in reaction to the closed access publishing system and as an act of impatience in the face of diffusion of green and gold OA, which is perceived to be both slow and complicated (Björk, 2017; Bodó, 2016). The most prominent example is Sci-Hub, which provides free access to publications behind paywalls through the fraudulent use of existing licensee accounts. The use of Sci-Hub within the academic community (Bohannon, 2016; Greshake, 2016), the finding that two-thirds of downloads in the field of medical research stem from low-and lower-middle-income countries (Sagemüller et al., 2021; Till

et al., 2019), and its widespread use and global availability allow tentative answers to *what-if* questions relating to the free-of-charge provision of academic knowledge as a public good: To what extent can free access alone contribute toward a leveling of the playing field? Does an increase in access allow more or other scientists, or even institutions, to participate in cutting-edge discussions? Does it affect the diversity of authors and locations in which research is conducted?

Moving toward the answers to these questions, we match Sci-Hub access data with entries from WoS and derive diversity measures concerning the geographical location of published mathematics scholars at the journal level. Furthermore, we aim to explain the growing (or shrinking) representation of top downloading countries (via Sci-Hub) at the journal level from the year 2010 (pre-Sci-Hub) to the year 2016.

The paper proceeds with a literature review on scholarly findings regarding the access of developing countries to academic literature in section 1, continues with a summary of developments in OA and shadow libraries in section 2, hypotheses are developed in section 3. Section 4 presents the data sample and the methodological approach. Results are shown in section 5, followed by a discussion in the concluding section 6.

2 | PARTICIPATION OF DEVELOPING COUNTRIES IN RESEARCH AND OA

2.1 | Developing countries' access to scholarly literature and their participation in scientific research

Representation of developing countries (i.e., the number of publishing authors that are affiliated with an institution in such countries) in the international peer-reviewed literature is low compared to the representation of developed countries. Demeter (2020) calculates that more than 70% of authors publishing in Scopus-indexed humanities or social science journals are affiliated with institutions located in North America or Western Europe. Moreover, even within relatively under-represented areas, such as Southeast Europe, it is the more developed countries that dominate regional publication output, for instance in mathematics and physics (Kutlača et al., 2014). Paradoxically, the representation of authors from developing countries is also low in journals that directly address these countries (e.g., in the journal Development Studies) (Cummings & Hoebink, 2017), that discuss general problems in these countries (Rafols et al., 2019), and in studies regarding research that is conducted in less-developed countries (Dahdouh-Guebas et al., 2003). This underrepresentation could be detrimental in areas that are likely to affect developing countries to a substantial degree in the future, including research in climate change (Blicharska et al., 2017; Karlsson et al., 2007) and tropical infectious diseases (Forero-Pineda & Jaramillo-Salazar, 2002).

A reason for this North-South publication gap (Blicharska et al., 2017) may lie in developing countries' inability to access international research publications, as illustrated by a World Health Organization (WHO) study revealing that the majority of medical institutions in developing countries subscribed to a maximum of five international journals (Aronson, 2004). Furthermore, Boudry et al. (2019) present the case of the WHO's Hinari Access to Research for Health programme (Aronson & Long, 2003) for institutional access to biomedical literature, affirming that access to scholarly literature still differs vastly between institutions and regions. Moreover, financial resources are needed to access the latest research, and a lack of funding constitutes a common bottleneck for developing countries (Adcock & Fottrell, 2008; Arunachalam, 2003; Boudry et al., 2019; Gordon, 1979; Tennant et al., 2016). In sum, insufficient access is likely to be one of developing countries' major obstacles to conducting impactful research in the first place (Arunachalam, 2003; Gordon, 1979; James, 2020).

2.2 | Shadow libraries, Sci-Hub, and Open Access

The OA movement seeks to remove barriers to accessing scholarly literature through measures such as free licensing, the establishment and operation of a publication infrastructure, and the development of alternative funding models. There are various categories or variations of OA, including publication in an OA journal (gold OA) or availability on a preprint server (green OA) (Suber, 2012). Up to 50% of scholarly articles have been published OA in recent years (Hampl et al., 2021; Himmelstein et al., 2018; Van Noorden, 2013b), with the share of publications in purely OA journals (gold OA) amounting to 12–13% (Hampl et al., 2021; Schimmer et al., 2015). According to Crawford (2019), article processing fees were charged for the majority (65%) of these contributions.

In addition to gold and green OA, the literature sometimes considers the semi-institutionalized distribution of paywalled articles while ignoring the copyrights of the respective publishers to be black OA (Björk, 2017; Bodó, 2016; Bohannon, 2016; Greshake, 2016). Possible activities in this regard include requests for articles via social

networks such as Twitter (#icahhazpdf) or Reddit (r/scholar), which are then forwarded by individuals with (legal) access to these documents, resulting in peer-to-peer sharing activities (Gardner & Gardner, 2017). Shadow libraries constitute a more automatized form of copyright circumvention based on online databases of unauthorized copies of books or articles (Karaganis, 2018). The largest shadow libraries, Library Genesis (focusing on books) and Sci-Hub (focusing on scholarly articles), partly incorporate the content posted to social media platforms (Cabanac, 2016) and were most likely originally developed in post-Soviet states, where covert, self-organized sharing of copied literature has a long tradition (Bodó, 2018).[4]

Although shadow libraries emerged "at the peripheries of the globalizing world of scholarly publishing," they also solved problems at the "center" by reducing financial and other constraints (Bodó, 2016, p. 8). In 2011, researcher and software engineer Alexandra Elbakyan launched Sci-Hub, further automating the ingestion of new content. If a user-requested article is not in its database, Sci-Hub uses library accounts to insource it from publisher websites (Elbakyan, 2019). Initially, the service was mainly used by Russian researchers, but quickly gained popularity in other countries (Elbakyan & Bozkurt, 2021). In 2015, Sci-Hub was used globally (Bohannon, 2016; Greshake, 2016), and in 2017, the service provided access to almost 70% of journal papers registered in Crossref and 85% of papers published in paywalled journals (Himmelstein et al., 2018). Two-thirds of all downloads of medical science literature via Sci-Hub happen in low-and lower-middle-income countries (Till et al., 2019), supposedly reflecting their inability (or unwillingness) to pay subscription charges. The platform is also accessed by individuals affiliated with research institutions from more developed countries that presumably pay subscription fees to offer legal access, which can be explained by Sci-Hub's user-friendly design (Bohannon, 2016). Moreover, Correa et al. (2022) highlight that the number of Sci-Hub downloads is a robust predictor of future article citations in the fields of economics, consumer research, and neuroscience, as well as for multi-disciplinary journals, hinting at Sci-Hubs' function as an illegal, yet efficient disseminator of scientific knowledge.

Sci-Hub, as well as other shadow libraries, are often not considered *real* OA, because their modes of conduct are not compatible with widely accepted definitions of OA (such as the Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities (Max Planck Society, 2003)). Machin-Mastromatteo et al. (2016), for example, argue that large parts of the OA community reject the idea that Sci-Hub is OA because the articles are not available under an open licensing scheme such as Creative Commons. Instead, publications are merely pirated at the end of a closed-access process. In addition, the OA community worries that classifying Sci-Hub as OA will harm the movement by associating it with illegal practices and alienating parts of the academic world. On a more fundamental level, Saleem et al. (2017) take the stance that for science to be true to its own ethics, illegal practices should not be pursued even for the sake of progress. On the contrary, Travis (2016) presents the results of a survey of 11,000 scientists that show 88% of the respondents do not condemn the downloading of illegally sourced articles. Bendezú-Quispe et al. (2016) further illustrate the dilemma for practicing physicians in low-and-middle-income countries, who are often in a predicament between acting illegally and keeping up with state-of-the-art knowledge to offer the best treatment to their patients. Elbakyan (2016), the founder of Sci-Hub, argues that Sci-Hub offers truly *open* access to scholarly literature because it completely removes barriers for virtually anyone. Strielkowski (2017) points out parallels between Sci-Hub and the illegal music downloading platform Napster in the 1990s. Just like Napster kicked off a profound change in the music industry's business model, Sci-Hub may be a catalyst for change in the academic publishing business. In conclusion, Sci-Hub does not fall under common definitions of OA, but it does provide access to otherwise costly literature (Boudry et al., 2019). In that sense, there may be a possible future scenario of a true scholarly commons that serves the interests of researchers and the public all over the world (Lawson, 2017).

3 | HYPOTHESES

One of the promises of OA is that it allows research to be conducted efficiently on a global scale. Researchers whose institutions might be unable to afford the license fees to access articles behind a paywall could easily participate in global debates in their field, and research findings would diffuse faster. Tennant et al. (2016) show in their literature review that OA publications yield a higher citation count, highlighting the potential increase in the impact of accessible publications. On the contrary, Moksness and Olsen (2020) indicate that Norwegian researchers perceive publication in non-OA journals to be better for their scientific career compared to OA journals. This can result in a self-fulfilling prophecy with disadvantages for those who cannot afford access to paywalled articles.

Journal subscription fees are costly, so that even renowned institutions, such as the University of

California or the German Max Planck Society, are canceling and renegotiating subscription deals with major publishers.[5] Although the acceptance of huge investments in journal subscriptions is dwindling, publishing authors may be actively or involuntarily discriminated against by their peers if they cannot access state-of-the-art research publications in a specific field or journal when trying to publish their own research findings (Gordon, 1979; James, 2020). Even worse, their research endeavors may be outdated without them even being aware of it.

Under the assumption that scientific "genius" is distributed equally around the world, a shift from closed access to OA is one of the factors that could lead to a discernable increase in diversity (e.g., regarding the origin of authors) in journals. Unconditional access to a journal allows readers to increase familiarity with the research questions a scientific community is tackling. It also fosters referencing of prior work via citations, which can act as a signal to editors that authors are familiar with the scope of a journal, but also with debates and research results of the community the journal is targeting. Furthermore, the results of Correa et al. (2022) underline the impact of Sci-Hub on the dissemination of scientific knowledge. Therefore, if access to existing scientific knowledge is a major factor in successful scholarly work on the aggregate (country) level, access to publications from a certain journal should lead to an observable increase in publications from a specific country. Hence:

> **H1a.** Increased access for researchers of a specific country to publications in a certain journal increases the overall representation of researchers from the respective country in this journal.

Researchers in less-developed countries often lack the resources to pay license fees for access to (paywalled) journals in the first place (Adcock & Fottrell, 2008; Gordon, 1979; Tennant et al., 2016). Free access should be of particular relevance for them because they would benefit relatively more from unrestricted access than more developed countries.

> **H1b.** Free access increases the representation of authors from less-developed countries in a journal to a greater degree than the representation of authors from more developed countries.

A fundamental problem with knowledge is the impossibility of determining its value *ex-ante*, making it unsuitable for purely market-based transactions and causing disincentives to invest in its production (Arrow, 1962). This observation is also applicable to the investment in scholarly papers by institutions and individual researchers (Forero-Pineda & Jaramillo-Salazar, 2002). At the same time, there is both an increasing number of scholarly journals (Ware & Mabe, 2015) and a greater demand for article availability, as scholars nearly doubled their yearly reading between 1977 and 2005 to keep up with the research in their field (Tenopir et al., 2009). Although access to cutting-edge research may be a prerequisite to participate in and contribute to current academic discussions, publications in top-tier scientific journals may still require additional investments and learning. The possibility for research articles to be published in these journals is further complicated by a ranking-driven rush to the limited space in highly cited journals (Wouters, 2019). However, recognition and familiarity with cutting-edge research and probably subsequent citation of state-of-the-art publications from top-tier journals may affect publication likelihood in low-ranked scientific journals as well. Hence:

> **H2.** Increased access to research published in top journals of a subfield increases the representation of a country's authors in journals belonging to this subfield.

4 | DATA, VARIABLES, AND METHODOLOGICAL APPROACH

We match data on papers available via Sci-Hub (Sci-Hub listed Digital Object Identifiers (DOIs), published March 2017; see Hahnel (2017) or Elbakyan (2018), respectively) and Sci-Hub server logs for the year 2015/2016 (Bohannon & Elbakyan, 2016) with entries from the WoS database. The Sci-Hub access dataset allows us to index WoS papers that have been requested via Sci-Hub at least once in the period of 2011 (emergence of Sci-Hub) to 2017. The Sci-Hub server logs allow downloaders to identify the countries of origin via IP addresses (Bohannon & Elbakyan, 2016).

Because server logs for the entire period are not available (yet), we have to assume that the proportions of country-specific requests are representative of the whole period under investigation. Furthermore, we focus on mathematics as a research discipline with relatively low capital intensity to counter concerns of endowment effects or infrastructure, which could exist in other basic research disciplines such as physics or engineering. Laverde-Rojas and Correa (2019) show that scientific productivity in mathematics has a positive impact on a country's economic complexity, which in turn is a useful predictor for competitiveness in the global economy.[6]

Scientific output in general (Solarin & Yen, 2016) and scientific output in the field of mathematics (among others) are correlated with economic growth (Jaffe et al., 2013). WoS classified 4,172 journals as belonging to the field of mathematics in 2018. Because we are interested in changes in publications that have been made accessible via Sci-Hub, we eliminate all journals that did not exist in 2010 (the year before the creation of Sci-Hub). Furthermore, we exclude all journals that did not have at least one Sci-Hub request in the period under investigation, because we cannot reliably establish whether the journal is simply of lower interest or generally unavailable via Sci-Hub. After applying these restrictions, we are left with 486 journals, including DOIs of every article listed in WoS. Country-level publication variables are derived by matching WoS-listed publications with a list of 249 country names issued by the International Organization for Standardization (ISO 3166), gross domestic product data published by the World Bank, and the education index (as part of the United Nations Human Development Index). Countries that have never been observed publishing in one of the mathematics journals or downloading papers from Sci-Hub are excluded from the data, resulting in 164 countries total in our dataset.

Matching the WoS-listed mathematics publications with the Sci-Hub access dataset (listed DOIs, 2011–2016), we find an overlap of roughly 80%, which is in line with prior evidence (Himmelstein et al., 2018) and suggests that four out of five publications in mathematics have been requested at least once via Sci-Hub between the years 2011 and 2016 (see also Greshake (2017) for the average age of requested papers). Additionally, we use the dataset from Bohannon and Elbakyan (2016)[7] to determine the country with the most download requests on the journal level in 2015–2016 by matching DOIs.

Table 1 summarizes country representation in both 2016 and 2010 and indicates a high skewness (distribution is right-skewed). Therefore, we construct our dependent variable from the logarithmic ratio of representations of authors from all of the countries in the dataset, such that:

$$\begin{aligned}&\mathrm{Log(representation-ratio)}\\&=\log\left(\frac{1+\text{country representation}_{2016}}{1+\text{country representation}_{2010}}\right)\\&=\begin{array}{l}\log\left(1+\dfrac{\#\text{ papers from country}_{2016}}{\text{Total \# of papers in journal}_{2016}}\right)\\-\log\left(1+\dfrac{\#\text{ papers from country}_{2010}}{\text{Total \# of papers in journal}_{2010}}\right).\end{array}\end{aligned}$$

The papers from each country are counted via fractional counting (Hooydonk, 1997), assigning each of the N authors of a paper 1/N credits. An author's institutional origin is assigned by his or her first affiliation in the WoS.

Our first explanatory variable is "Downloads," which is the number of requests from every country between September 2015 and March 2016 (Sci-Hub server logs) for papers published in each of the journals and serves as a proxy for all downloads in the period under investigation.[8]

Our second explanatory variable is "Top5Downloads_weighed." This variable measures how often Sci-Hub users

TABLE 1 Descriptive statistics of the regression variables

	N	Mean	SD	Min	Max	Level
Log (Representation-Ratio)	79,704	0.000	0.015	−0.372	0.364	Journal/Country
Country Representation$_{10}$	79,704	0.006	0.033	0.000	0.991	Journal/Country
Country Representation$_{16}$	79,704	0.006	0.032	0.000	0.927	Journal/Country
Downloads	79,704	2.775	31.935	0.000	3363	Journal/Country
Top1Downloads_weighed	79,704	7.120	56.680	0	2948.488	Journal/Country
Top5Downloads_weighed	79,704	28.018	154.381	0	5648.877	Journal/Country
Share Not Available	79,704	0.317	0.279	0.000	1	Journal
Expected Citations$_{2010}$	79,704	4.357	3.241	0.450	28	Journal
Gini Country$_{2010}$	79,704	0.841	0.124	0.034	0.960	Journal
Gini Keyword$_{2010}$	79,704	0.993	0.025	0.500	1	Journal
HDI-Class$_{\text{VeryHigh}}$	74,358	0.392	0.488	0	1	Country
HDI-Class$_{\text{High}}$	74,358	0.288	0.453	0	1	Country
HDI-Class$_{\text{MedLow}}$	74,358	0.320	0.467	0	1	Country
ΔGDP	74,844	11.468	27.362	−66	109	Country
ΔEducation Index	72,900	6.182	6.758	−23	46.073	Country

from each country downloaded papers published in the top five journals, which focus on a similar subfield of mathematics as the focal journal. To determine which journals are top journals in a subfield, the keywords of all papers of every journal were retrieved from WoS. After using Porter's stemming algorithm implemented in the "tm" package in R (Feinerer, 2013), all keyword vectors are added to a document-term matrix containing the term frequencies of every keyword used in every journal.[9] This allows for an analysis of cosine similarity between journals based on their respective keyword vectors.[10] The five most similar top journals are determined for each journal, and the download counts of each country in each of the top journals (weighed by their respective cosine similarity) are calculated. In this context, a journal qualifies as a top journal by attaining an average citation rate in the topmost quartile between 2005 and 2010.

On the country level, we use the United Nation's Human Development Index (HDI) of 2016 to establish three groups of countries resulting in the two variables "HDI-$Class_{High}$" and "HDI-$Class_{M\&L}$," with countries having a very high HDI acting as the group of reference.[11]

We added several controls on the journal level and the country level. "Share not Available" is the percentage of articles that have never been requested via Sci-Hub (until 2017), which is a proxy for the overall attractiveness or availability of a journal. "Expected $Citations_{2010}$" is the mean journal paper citations between 2005 and 2010, according to WoS. These are usually treated as "expected" citations (of 2010 in our case) and control for the general attractiveness of the journal. Furthermore, we include two concentration measures because our argumentation is largely based on diversity aspects. Both could be interpreted as baseline effects before the introduction of Sci-Hub. "Gini $Country_{2010}$" is a proxy for journal diversity regarding authors' affiliations in 2010 computed from WoS data. "Gini $Keyword_{2010}$" proxies for the diversity of article keywords within journals in 2010 and serves as a measure for general openness and topical variety. To control for the growth of the overall economic activity in a country, the change in GDP per capita between 2010 and 2016 is introduced as "ΔGDP." The variable "ΔEducation Index" shows the development of each country's United Nations Education Index score (a subindex of the HDI) in the aforementioned timeframe and is used to control for the change in the quality of countries' education systems, as this is likely to impact overall academic performance.[12]

5 | RESULTS

First descriptive indicators for the regression covariates (see Table A2) reveal a negative correlation between "Share not Available" and "Expected $Citations_{2010}$" irrespective of the country of the top downloader. Furthermore, the share of requested papers and Gini coefficients of authors' country of origin (as a proxy for diversity of contributing authors) are positively correlated in 2010. A slightly positive correlation is also found for the share of requested papers and Gini coefficients for keywords (as a proxy for diversity of research topics). These results indicate that it is the more reputable journals that are accessed (presumably the ones that are also more expensive to access), and that the availability increases slightly with diversity in authors and topics.

To test the hypotheses outlined above, an OLS regression on the "Log(Representation-Ratio)" for authors from downloading countries between 2010 and 2016 was estimated (see Table A3). To account for the fact that publication count data was collected on the journal level, the models are also estimated with cluster-robust standard errors and reported in Table 2. Model (I) relates the change in the representation of the downloading country before and after the emergence of Sci-Hub to the number of downloads and controls for the share of never requested papers (which is a measure of journal interest for Sci-Hub users) and the expected citations in 2010 (journal quality or impact). The explanatory variable "Downloads" has a significant positive effect showing that countries downloading a journal more often have an increasing representation in that same journal, supporting H1a. "Share Not Available" (never downloaded) turns out to be insignificant as opposed to "Expected $Citations_{2010}$," suggesting the quality effect already apparent in the correlations (see above). Adding the concentration measures in Model (II) reveals that the "Representation-Ratio" of a country decreases with a higher concentration of countries publishing in a journal and a narrower topical focus. Model (II) also includes the "HDI-Class" variables for high and middle/low developed countries according to the Human Development Index instead of the "Downloads," and Model (III) includes both. The results are very similar to each other, showing a significant positive relationship between the lower development status of countries and the share of representation in the respective journal. This result supports H1b, as it implies that developing countries increase their representation share with increased downloads via Sci-Hub. More interestingly, the results indicate that there is a nonlinear relationship. Countries that are relatively more developed and closer to the group of the very highly developed ones can increase their representation significantly more (Wald test for difference of coefficients of "HDI-$Class_{High}$" and "HDI-$Class_{M\&L}$" is significant at $p < .001$). A possible reason may be that these countries are more likely to have already established institutions and infrastructure allowing for internationally competitive research. Model

TABLE 2 OLS Regressions with cluster-robust standard errors

	(I)	(II)	(III)	(IV)	(V)	(VI)
(Intercept)	−0.094	0.017	1.885	4.767*	12.975	4.949
	(0.244)	(1.952)	(2.928)	(2.291)	(8.424)	(4.468)
Downloads	0.193***		0.185**		0.087	
	(0.070)		(0.069)		(0.060)	
Share Not Available	0.650	−0.459	0.457	−0.172	0.191	−0.378
	(0.399)	(0.350)	(0.446)	(0.409)	(0.595)	(0.471)
Expected Citations_{2010}	−0.099*	0.034	−0.066	0.013	0.230*	0.258**
	(0.047)	(0.033)	(0.055)	(0.064)	(0.092)	(0.093)
Gini Country_{2010}		−4.001***	−4.745***	−4.445***	−1.338	−2.300
		(0.843)	(0.957)	(0.909)	(1.886)	(1.422)
Gini Keyword_{2010}		−4.778**	−6.266*	−5.794**	−19.284*	−12.071**
		(1.656)	(2.764)	(2.119)	(8.524)	(4.655)
$\text{HDI-Class}_{\text{High}}$		19.128***	18.562***			9.667***
		(1.541)	(1.505)			(1.292)
$\text{HDI-Class}_{\text{M\&L}}$		9.626***	9.608***			2.573**
		(0.946)	(0.947)			(0.895)
ΔGDP				0.177***	0.200***	0.180***
				(0.024)	(0.024)	(0.024)
ΔEducation Index				0.477***	0.318***	0.294***
				(0.064)	(0.056)	(0.051)
ΔGDP:Downloads				0.007**		
				(0.002)		
$\text{Expected Citations}_{2010}$:Downloads				−0.007		
				(0.008)		
Top5Downloads_weighed					0.114***	−0.081
					(0.021)	(0.051)
$\text{Expected Citations}_{2010}$: Top5Downloads_weighed					−0.009*	−0.007*
					(0.004)	(0.003)
$\text{HDI-Class}_{\text{High}}$:Top5Downloads_weighed						0.225***
						(0.054)
$\text{HDI-Class}_{\text{M\&L}}$:Top5Downloads_weighed						0.132*
						(0.049)
N	79,704	74,358	74,358	72,414	72,414	72,414
AIC	−443591.5571	−409605.3857	−409719.3936	−397193.2874	−397490.8729	−397847.8903
F test	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001

Note: Estimates and standard errors (in parentheses) reported ×10,000.
*$p < .05$; **$p < .01$; ***$p < .001$.

(IV) introduces the GDP development and the change in the Education Development Index for each country. Countries with a growing GDP and education system are represented significantly more often. The interaction coefficient for GDP change and Sci-Hub downloads is positive and significant at the 5% level. Model (V) shows that downloading publications from the top five journals of the respective subfield has a significantly positive effect on the increase of country representation. In model (VI), this effect loses significance as the introduced interaction terms of "$\text{HDI-Class}_{\text{High}}$" and "$\text{HDI-Class}_{\text{M\&L}}$" with "Top5Downloads_weighed" reveal a positive effect of the

availability of top journals in developing countries that we hypothesized in H2. Again, we see a difference between high and middle/low developed countries because the coefficients indicate a significantly higher impact of downloading from the top five journals of a subfield on the countries' representation in a given journal for highly developed countries (Wald test for difference of the coefficients of the interactions is significant at $p < .001$). A small but significant negative effect is found for the interaction of "Top5Downloads_weighed" and the "Expected Citations." This implies that downloading papers from top journals in a field helps authors to publish in less-cited journals in the same field. As the reputation of a journal (expressed in "Expected Citations") increases, the effect of downloading papers from top journals on the country's representation decreases.

To account for heteroskedasticity in the models, heteroskedasticity-robust White standard errors are estimated as a robustness check and reported in Table A4. Compared to the clustered standard error models, the journal-level control variables lose statistical significance. The hypothesis testing regressors remain at a similar magnitude and show the same sign as in the regular OLS models and the clustered standard error models. The interaction of "Top5Downloads_weighed" and the "Expected Citations" loses its significance in this robustness check. To control for regional differences of Sci-Hub usage and as an alternative to the usage of the HDI classes, we add regional dummy variables to Models (III) and (IV) (see Table A5). The regions are assigned according to the World Bank's classification scheme. Further, a set of models including a "China"—dummy is estimated. All models in this robustness check support our hypotheses. We can further infer that, while all regions were able to improve their representations compared to North America, it is mainly China, East Asia, and the Pacific, as well as Latin America and the Caribbean, where a significantly positive impact of downloading top journals on an improved country representation is discernable. This is largely in line with the distinction according to HDI classes, because most of the highly developed countries can be found in these regions.

A Shapiro–Wilk normality test reveals that the error terms of the regression models are significantly different from a normal distribution. Therefore, robustness checks of the variables included in Model (VI) are estimated using quantile regression for a set of distribution percentiles of the dependent variable (see Table A6). Model (VI) was chosen because it is the most complex model estimated above. As a result, the negative effect of "Top5-Downloads_weighed" prevails only in the lower quantiles. The positive and highly significant effect of the interaction between "HDI-$\text{Class}_{\text{High}}$" and "HDI-$\text{Class}_{\text{M\&L}}$" and "Top5Downloads_weighed" is driven by the lower quartiles in this model. The effects of "HDI-$\text{Class}_{\text{High}}$" and "HDI-$\text{Class}_{\text{M\&L}}$" vary strongly from positive to negative with increasing percentiles, rendering the effect on the conditional mean in Model (VI) rather ambiguous. As a caveat, a look at R^2 statistics (which are all well below 0.1) reveals a low percentage of variance explained by our analyses, indicating that important determinants of change in country representation are probably not revealed through our investigation.

6 | DISCUSSION AND CONCLUSION

Our investigation set out to shed light on a possible connection between free access to academic literature and the representation of authors from developing countries in scientific journals. The different distribution models currently pursued by scientific publishers are tainted by contradictory effects on scientists and institutions with few resources (paywalled articles are exclusionary in terms of access, article processing charges (APC) based gold OA is exclusionary in terms of publishing). Using Sci-Hub, the biggest shadow library for scholarly literature, as a data source allows us to construct a counterfactual: a mode of distribution of scientific knowledge where financial resources are neither needed for access nor disclosure of state-of-the-art research. We matched data from WoS with accesses via Sci-Hub to establish whether increased availability of access to cutting-edge research results in a country leads to an increase in the representation of authors from this country in a specific journal. We deliberately chose mathematics journals as our sample field to avoid endowment effects. We find that increased downloads of a specific journal's articles are positively correlated to a higher representation of the downloading country's authors in this journal. This especially benefits countries that are classified by the United Nations Development Programme as having a high development status, as we find a positive effect on country representation for the access of top journals in a specific field when downloaded from these countries.

At first glance, these results are encouraging. Contributions of researchers from less-developed countries are now better recognized than before Sci-Hub was established. Furthermore, researchers from countries that make use of shadow libraries are better able to publish in international scholarly journals. While basic socio-economic indicators have been included in the statistical model, it does not control for all factors possibly impacting the representation of a country's researchers in scholarly journals, such as individual and nuanced changes in research and higher education

policies in the countries under investigation. Still, the effects measured lend support to the fundamental promise of the OA movement that "truly free" access to and availability of research results will lead to better recognition of contributions by researchers, which have been overlooked previously.

This study is an early attempt to utilize the Sci-Hub access protocols beyond descriptive accounts of who is downloading (Bohannon, 2016) and how much is available (Himmelstein et al., 2018). It faces several serious shortcomings and restrictions, aside from the fact that other shadow libraries and copyright circumvention mechanisms, such as Library Genesis or #canihazpdf, are not considered here. First, we are not able to directly relate downloaded papers to publications in which they are cited for several reasons. This would allow us to assess the ability of downloaders to enter higher-quality journals compared to the status quo before the advent of Sci-Hub (the positive coefficients for our controls hint in this direction). Second, the users of Sci-Hub (and supporters of OA) might differ fundamentally in their publication behavior. We focus exclusively on journals that are usually covered by licensing models (non-OA) and cannot rule out the possibility that downloaders publish mainly in OA journals themselves. Finally, although others have pointed out that Sci-Hub is used extensively in the academic realm (Greshake, 2016), the possibility exists that it is actually employed mainly by practitioners or the industry. Therefore, our analyses cannot give a full account of possible benefits (or damages) of free availability, as they do not take the user-/demand-side effects beyond the academic realm into account. Other economic factors, such as endowment effects and brain drain toward developed countries, are still major hindrances to the small positive effect of the ability to access highly influential journals while researching in a country that is not very highly developed according to the United Nations Development Programme, especially in disciplines that are more investment-intensive than mathematics.

ACKNOWLEDGMENT

Open access funding enabled and organized by Projekt DEAL.

ORCID

Kilian Buehling https://orcid.org/0000-0002-5244-7547
Dorothea Strecker https://orcid.org/0000-0002-9754-3807

ENDNOTES

[1] The increase of coauthorship in all science and engineering disciplines is led by medicine, while mathematics papers are authored by the fewest number of authors (Wuchty et al., 2007).

[2] For example, the Berlin declaration on open access to knowledge in the sciences and humanities (Max Planck Society, 2003).

[3] Information about specific policies of funders and research institutions that have mandated OA publication can be found in the Registry of Open Access Repository Mandates and Policies: http://roarmap.eprints.org/.

[4] This tradition is partly rooted in the political constraints put on readers by the Soviet regime, and a decline in public funding of the publishing industry and libraries in the post-Soviet era. In the 1990s, digital technologies and the emergence of the internet led to the build-up of collections of digitized documents, allowing users to more effectively circumvent copyright restrictions. These small collections were shared, then consolidated, and later specialized and reorganized by various shadow libraries. Whereas the contents of earlier, small shadow libraries were sourced by many contributors, most adding only a few documents to the collections, the majority of contributions to the Library Genesis database appear to be ingests of larger digital collections (Bodó, 2018; Cabanac, 2016).

[5] Information about recent major subscription deal cancellations and renegotiations can be found on https://sparcopen.org/our-work/big-deal-cancellation-tracking/.

[6] Admittedly, this effect is only robust for high-income countries in the analysis.

[7] Available at: Dryad Digital Repository https://doi.org/10.5061/dryad.q447c.

[8] The underlying assumption of our approach is that reception/access to a certain journal is a pre-condition to publish in this journal.

[9] Structuring text data in this manner implies a bag-of-words assumption, which is appropriate when considering the keywords used in a journal as an indicator of its topical focus regardless of the order in which they appear.

[10] Cosine similarity is a commonly used measure in natural language processing because of its implicit vector length normalization (Tan et al., 2016).

[11] This classification follows the United Nations Development Programme (UNDP)'s Human Development Report (UNDP, 2020, p. 336). For a graphical overview of the classifications, see Figure A1. Countries having a *middle* or *low* HDI are grouped together, because downloading in countries with a low HDI is very rare.

[12] For some countries there are no observations for either ΔGDP or ΔEducation Index. The list of missing countries can be found in Table A1.

REFERENCES

Adcock, J., & Fottrell, E. (2008). The north-south information highway: Case studies of publication access among health researchers in resource-poor countries. *Global Health Action*, *1*(1), 1865. https://doi.org/10.3402/gha.v1i0.1865

Agrawal, A. (2006). Engaging the inventor: Exploring licensing strategies for university inventions and the role of latent knowledge. *Strategic Management Journal*, *27*(1), 63–79. https://doi.org/10.1002/smj.508

Aronson, B. (2004). Improving online access to medical information for low-income countries. *New England Journal of Medicine*, *350*(10), 966–968. https://doi.org/10.1056/nejmp048009

Aronson, B., & Long, M. (2003). HINARI: Health InterNetwork Access to Research Initiative. *Serials*, *16*(1), 7–12. https://doi.org/10.1629/167

Arrow, K. J. (1962). Economic welfare and the allocation of resources to inventive activity. In R. R. Nelson (Ed.), *The rate and direction of technical change*. NBER. https://doi.org/10.1515/9781400879762-024

Arunachalam, S. (2003). Information for research in developing countries—Information technology, a friend or foe? *International Information & Library Review*, *35*(2–4), 133–147. https://doi.org/10.1080/10572317.2003.10762596

Beauchesne, O. H. (2011). Stream of scientific collaborations between world cities. In K. Börner & M. J. Stamper (Eds.), *7th iteration 2011: Science maps as visual interfaces to digital libraries. Places & spaces: Mapping science*. Courtesy of Science-Metrix, Inc..

Bendezú-Quispe, G., Nieto-Gutiérrez, W., Pacheco-Mendoza, J., & Taype-Rondan, A. (2016). Sci-Hub and medical practice: An ethical dilemma in Peru. *The Lancet Global Health*, *4*(9), e608. https://doi.org/10.1016/S2214-109X(16)30188-7

Björk, B. C. (2017). Gold, green, and black open access. *Learned Publishing*, *30*(2), 173–175. https://doi.org/10.1002/leap.1096

Blicharska, M., Smithers, R. J., Kuchler, M., Agrawal, G. K., Gutiérrez, J. M., Hassanali, A., Huq, S., Koller, S. H., Marjit, S., Mshinda, H. M., Masjuki, H. H., Solomons, N. W., Staden, J. V., & Mikusiński, G. (2017). Steps to overcome the north–south divide in research relevant to climate change policy and practice. *Nature Climate Change*, *7*(1), 21–27. https://doi.org/10.1038/nclimate3163

Bodó, B. (2016). Pirates in the library—An inquiry into the guerilla open access movement. In *8th annual workshop of the international society for the history and theory of intellectual property*. CREATe Centre, University of Glasgow. https://doi.org/10.2139/ssrn.2816925

Bodó, B. (2018). The genesis of library genesis. In J. Karaganis (Ed.), *Shadow libraries: Access to knowledge in global higher education* (pp. 25–52). MIT Press.

Bohannon, J. (2016). Who's downloading pirated papers? Everyone. *Science*, *352*, 508–512. https://doi.org/10.1126/science.aaf5664

Bohannon, J., & Elbakyan, A. (2016). *Data from: Who's downloading pirated papers? Everyone*. Dryad Digital Repository. https://doi.org/10.5061/dryad.q447c

Boudry, C., Alvarez-Muñoz, P., Arencibia-Jorge, R., Ayena, D., Brouwer, N. J., Chaudhuri, Z., Chawner, B., Epee, E., Erraïs, K., Fotouhi, A., Gharaibeh, A. M., Hassanein, D. H., Herwig-Carl, M. C., Howard, K., Kaimbo Wa Kaimbo, D., Laughrea, P.-A., Lopez, F. A., Machin-Mastromatteo, J. D., Malerbi, F. K., … Mouriaux, F. (2019). Worldwide inequality in access to full text scientific articles: The example of ophthalmology. *PeerJ*, *7*, e7850. https://doi.org/10.7717/peerj.7850

Brown, D. J. (2015). *Access to scientific research*. De Gruyter Saur. https://doi.org/10.1515/9783110369991

Cabanac, G. (2016). Bibliogifts in LibGen? A study of a text-sharing platform driven by biblioleaks and crowdsourcing. *Journal of the Association for Information Science and Technology*, *67*(4), 874–884. https://doi.org/10.1002/asi.23445

Coccia, M., & Wang, L. (2016). Evolution and convergence of the patterns of international scientific collaboration. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(8), 2057–2061. https://doi.org/10.1073/pnas.1510820113

Cohen, W. M., & Levinthal, D. A. (1990). Absorptive capacity: A new perspective on learning and innovation. *Administrative Science Quarterly*, *35*(1), 128–152. https://doi.org/10.2307/2393553

Correa, J. C., Laverde-Rojas, H., Tejada, J., & Marmolejo-Ramos, F. (2022). The Sci-Hub effect on papers' citations. *Scientometrics*. *127*(1), 99–126. https://doi.org/10.1007/s11192-020-03806-w

Crawford, W. (2019). *Gold open access 2013–2018: Articles in journals (GOA4)*. Cites & Insights Books.

Cummings, S., & Hoebink, P. (2017). Representation of academics from developing countries as authors and editorial board members in scientific journals: Does this matter to the field of development studies? *European Journal of Development Research*, *29*(2), 369–383. https://doi.org/10.1057/s41287-016-0002-2

Dahdouh-Guebas, F., Ahimbisibwe, J., Van Moll, R., & Koedam, N. (2003). Neo-colonial science by the most industrialised upon the least developed countries in peer-reviewed publishing. *Scientometrics*, *56*(3), 329–343. https://doi.org/10.18356/b00b9c34-en

Davidson Frame, J., & Carpenter, M. P. (1979). International research collaboration. *Social Studies of Science*, *9*(4), 481–497. https://doi.org/10.1177/030631277900900405

Demeter, M. (2020). *Academic knowledge production and the global south: Questioning inequality and under-representation*. Springer. https://doi.org/10.1007/978-3-030-52701-3

Elbakyan, A. (2016, February 24). Why Sci-Hub is the true solution for open access: Reply to criticism. *Engineuring*. Retrieved from https://engineuring.wordpress.com/2016/02/24/why-sci-hub-is-the-true-solution-for-open-access-reply-to-criticism/

Elbakyan, A. (2018). *Sci-Hub download log of 2017*. Zenodo. https://doi.org/10.5281/zenodo.1158301

Elbakyan, A. (2019, March 30). Sci-Hub and Alexandra basic information. *Engineuring*. Retrieved from https://engineuring.wordpress.com/2019/03/31/sci-hub-and-alexandra-basic-information/

Elbakyan, A., & Bozkurt, A. (2021). A critical conversation with Alexandra Elbakyan: Is she the pirate queen, Robin Hood, a scholarly activist, or a butterfly flapping its wings? *Asian Journal of Distance Education*, *16*(1), 111–118.

Feinerer, I. (2013). *Introduction to the tm package text mining in R*. Retrieved from http://cran.r-project.org/web/packages/tm/vignettes/tm.Pdf

Forero-Pineda, C., & Jaramillo-Salazar, H. (2002). The access of researchers from developing countries to international science and technology. *International Social Science Journal*, *54*(171), 129–140. https://doi.org/10.1111/1468-2451.00364

Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., Milojević, S., Petersen, A. M., Radicchi, F., Sinatra, R., Uzzi, B., Vespignani, A., Waltman, L., Wang, D., & Barabási, A.-L. (2018). Science of science. *Science*, *359*(6379), eaao0185. https://doi.org/10.1126/science.aao0185

Gardner, C. C., & Gardner, G. J. (2017). Fast and furious (at publishers): The motivations behind crowdsourced research sharing. *College & Research Libraries*, *78*(2), 131. https://doi.org/10.5860/crl.78.2.131

Gordon, M. D. (1979). Deficiencies of scientific information access and output in less developed countries. *Journal of the American Society for Information Science*, *30*(6), 340–342. https://doi.org/10.1002/asi.4630300607

Greshake, B. (2016). Correlating the Sci-Hub data with World Bank indicators and identifying academic use. *Winnower*, *3*, e146485.57797. https://doi.org/10.15200/winn.146485.57797

Greshake, B. (2017). Looking into Pandora's box: The content of sci-hub and its usage. *F1000Research*, *6*, 541. https://doi.org/10.12688/f1000research.11366.1

Grossmann, A., & Brembs, B. (2019). Assessing the size of the affordability problem in scholarly publishing. *PeerJ Preprints.*, *7*, e27809v1. https://doi.org/10.7287/peerj.preprints.27809v1

Gui, Q., Liu, C., & Du, D. (2019). Globalization of science and international scientific collaboration: A network perspective. *Geoforum*, *105*, 1–12. https://doi.org/10.1016/j.geoforum.2019.06.017

Hahnel, M. (2017). *List of DOIs of papers collected by SciHub*. Figshare. https://doi.org/10.6084/m9.figshare.4765477.v1

Hampl, M., Finke, P., & Voigt, M. (2021). *Share of open access journal articles published by Berlin authors from 2019: Data*. Deposit Once. https://doi.org/10.14279/depositonce-11775

Himmelstein, D. S., Romero, A. R., Levernier, J. G., Munro, T. A., McLaughlin, S. R., Tzovaras, B. G., & Greene, C. S. (2018). Sci-Hub provides access to nearly all scholarly literature. *Elife*, *7*, e32822. https://doi.org/10.7554/elife.32822.022

Hoey, J., & Todkill, A. M. (2002). JPN and the electronic revolution. *Journal of Psychiatry and Neuroscience*, *27*(3), 158–160.

Hooydonk, G. V. (1997). Fractional counting of multiauthored publications: Consequences for the impact of authors. *Journal of the American Society for Information Science*, *48*(10), 944–945. https://doi.org/10.1002/(SICI)1097-4571(199710)48:10<944::AID-ASI8>3.0.CO;2-1

Houghton, J. W. (2001). Crisis and transition: The economics of scholarly communication. *Learned Publishing*, *14*(3), 167–176. https://doi.org/10.1087/095315101750240412

Huang, M.-H., Dong, H.-R., & Chen, D.-Z. (2012). Globalization of collaborative creativity through cross-border patent activities. *Journal of Informetrics*, *6*(2), 226–236. https://doi.org/10.1016/j.joi.2011.10.003

Jaffe, K., Caicedo, M., Manzanares, M., Gil, M., Rios, A., Florez, A., Montoreano, C., & Davila, V. (2013). Productivity in physical and chemical science predicts the future economic growth of developing countries better than other popular indices. *PLoS One*, *8*(6), e66239. https://doi.org/10.1371/journal.pone.0066239

James, J. E. (2020). Pirate open access as electronic civil disobedience: Is it ethical to breach the paywalls of monetized academic publishing? *Journal of the Association for Information Science and Technology*, *71*(12), 1500–1504. https://doi.org/10.1002/asi.24351

Karaganis, J. (2018). Introduction: Access from above, access from below. In J. Karaganis (Ed.), *Shadow libraries: Access to knowledge in global higher education* (pp. 1–24). MIT Press.

Karlsson, S., Srebotnjak, T., & Gonzales, P. (2007). Understanding the north–south knowledge divide and its implications for policy: A quantitative analysis of the generation of scientific knowledge in the environmental sciences. *Environmental Science & Policy*, *10*(7), 668–684. https://doi.org/10.1016/j.envsci.2007.04.001

Kutlača, D., Živković, L., Štrbac, D., Babić, D., & Semenčenko, D. (2014). Scientific research publication productivity in the areas of mathematics and physics in South Eastern Europe. *Yugoslav Journal of Operations Research*, *24*(3), 415–427. https://doi.org/10.2298/YJOR131112005K

Laverde-Rojas, H., & Correa, J. C. (2019). Can scientific productivity impact the economic complexity of countries? *Scientometrics*, *120*(1), 267–282. https://doi.org/10.1007/s11192-019-03118-8

Lawson, S. (2017). Access, ethics and piracy. *Insights: The UKSG Journal*, *30*(1), 25–30. https://doi.org/10.1629/uksg.333

Machin-Mastromatteo, J. D., Uribe-Tirado, A., & Romero-Ortiz, M. E. (2016). Piracy of scientific papers in Latin America: An analysis of Sci-Hub usage data. *Information Development*, *32*(5), 1806–1814. https://doi.org/10.1177/0266666916671080

Moksness, L., & Olsen, S. O. (2020). Perceived quality and self-identity in scholarly publishing. *Journal of the Association for Information Science and Technology*, *71*(3), 338–348. https://doi.org/10.1002/asi.24235

Max Planck Society (2003). Berlin declaration on open access to knowledge in the sciences and humanities. Retrieved from https://openaccess.mpg.de/Berliner-Erklaerung.

Papin-Ramcharan, J., & Dawe, R. A. (2006). The other side of the coin for open access publishing—A developing country view. *Libri*, *56*(1), 16–27. https://doi.org/10.1515/LIBR.2006.16

Persson, O., Glänzel, W., & Danell, R. (2004). Inflationary bibliometric values: The role of scientific collaboration and the need for relative indicators in evaluative studies. *Scientometrics*, *60*(3), 421–432. https://doi.org/10.1023/B:SCIE.0000034384.35498.7d

Polanyi, M. (1958). *Personal knowledge*. Routledge and Kegan Paul.

Pourret, O., Hedding, D. W., Ibarra, D. E., Irawan, D. E., Liu, H., & Tennant, J. P. (2021). International disparities in open access practices in the Earth Sciences. *European Science Editing*, *47*, e63663. https://doi.org/10.3897/ese.2021.e63663

Rafols, I., Ciarli, T., & Chavarro, D. (2019). Under-reporting research relevant to local needs in the South. *The Transformation of Research in the South. Policies and outcomes*. Éditions des Archives Contemporaines (pp. 105–110). Éditions des Archives Contemporaines. Retrieved from https://eac.ac/articles/2080

Sagemüller, F., Meißner, L., & Mußhoff, O. (2021). Where can the crow make friends? Sci-Hub's activities in the library of development studies and its implications for the field. *Development and Change*, *52*(3), 670–683. https://doi.org/10.1111/dech.12638

Saleem, F., Hasaali, M. A., & ul Haq, N. (2017). Sci-Hub & ethical issues. *Research in Social and Administrative Pharmacy*, *13*(1), 253. https://doi.org/10.1016/j.sapharm.2016.09.001

Schimmer, R., Geschuhn, K. K., & Vogler, A. (2015). *Disrupting the subscription journals' business model for the necessary large-scale transformation to open access*. MPG.PuRe. https://doi.org/10.17617/1

Sengupta, P. (2021). Open access publication: Academic colonialism or knowledge philanthropy? *Geoforum*, *118*, 203–206. https://doi.org/10.1016/j.geoforum.2020.04.001

Solarin, S. A., & Yen, Y. Y. (2016). A global analysis of the impact of research output on economic growth. *Scientometrics*, *108*(2), 855–874. https://doi.org/10.1007/s11192-016-2002-6

Stiglitz, J. E. (1999). Knowledge as a global public good. *Global public goods: International Cooperation in the 21st Century* (pp. 308–325). Oxford University Press.

Strielkowski, W. (2017). Will the rise of Sci-Hub pave the road for the subscription-based access to publishing databases? *Information Development*, *33*(5), 540–542. https://doi.org/10.1177/0266666917728674

Suber, P. (2012). *Open access*. MIT Press.

Tan, P. N., Steinbach, M., & Kumar, V. (2016). *Introduction to data mining*. Pearson Education India.

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. *F1000Research*, *5*, 632. https://doi.org/10.12688/f1000research.8460.3

Tenopir, C., King, D. W., Edwards, S., & Wu, L. (2009). Electronic journals and changes in scholarly article seeking and reading patterns. *Aslib Proceedings*, *61*(1), 5–32. https://doi.org/10.1108/00012530910932267

Till, B. M., Rudolfson, N., Saluja, S., Gnanaraj, J., Samad, L., Ljungman, D., & Shrime, M. (2019). Who is pirating medical literature? A bibliometric review of 28 million Sci-Hub downloads. *Lancet Global Health*, *7*(1), e30–e31. https://doi.org/10.1016/S2214-109X(18)30388-7

Travis, J. (2016). In survey, most give thumbs-up to pirated papers. *Science News*, https://doi.org/10.1126/science.aaf5704

United Nations Development Programme. (2020). *The next frontier: Human development and the Anthropocene*. Author. Retrieved from http://hdr.undp.org/en/2020-report

Van Noorden, R. (2013a). Open access: The true cost of science publishing. *Nature*, *495*(7442), 426–429. https://doi.org/10.1038/495426a

Van Noorden, R. (2013b). Half of 2011 papers now free to read. *Nature*, *500*(7463), 386–387. https://doi.org/10.1038/500386a

Wagner, C. S., Whetsell, T. A., & Mukherjee, S. (2019). International research collaboration: Novelty, conventionality, and atypicality in knowledge recombination. *Research Policy*, *48*(5), 1260–1270. https://doi.org/10.1016/j.respol.2019.01.002

Ware, M., & Mabe, M. (2015). *The STM report: An overview of scientific and scholarly journal publishing*. The Hague: International Association of Scientific, Technical and Medical Publishers.

Wouters, P. (2019). Globalization and the rise of rankings. In *Handbook on science and public policy*. Edward Elgar. https://doi.org/10.4337/9781784715946.00035

Wuchty, S., Jones, B. F., & Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. *Science*, *316*(5827), 1036–1039. https://doi.org/10.1126/science.1136099

How to cite this article: Buehling, K., Geissler, M., & Strecker, D. (2022). Free access to scientific literature and its influence on the publishing activity in developing countries: The effect of Sci-Hub in the field of mathematics. *Journal of the Association for Information Science and Technology*, 1–20. https://doi.org/10.1002/asi.24636

APPENDIX A

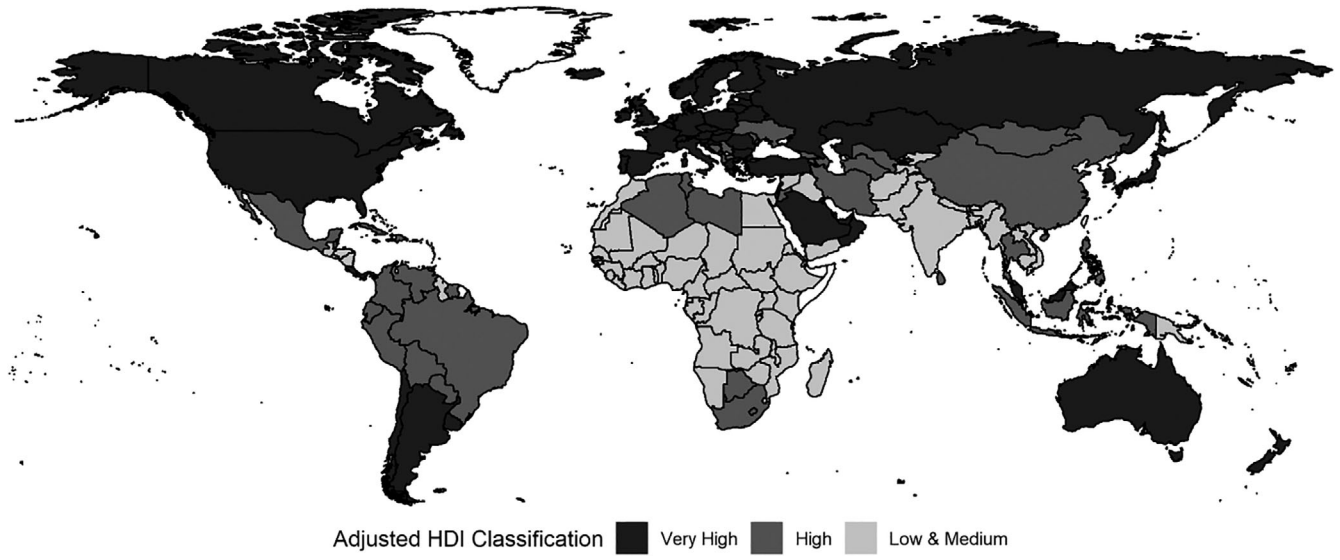

FIGURE A1 Adjusted HDI development classifications

TABLE A1 Countries missing in GDP and education data

Missing GDP data	Missing HDI data
New Caledonia	New Caledonia
Martinique	Martinique
Guadeloupe	Guadeloupe
Korea (Democratic People's Republic of)	Korea (Democratic People's Republic of)
French Polynesia	French Polynesia
Sint Maarten (Dutch part)	Sint Maarten (Dutch part)
Taiwan	Taiwan
Réunion	Réunion
Curaçao	Curaçao
Syrian Arab Republic	Eswatini
	Puerto Rico
	Macao
	Moldova, Republic of
	Hong Kong

TABLE A2 Correlation table of regression covariates

	Downloads	Share Not Available	Expected Citations$_{2010}$	Gini Country$_{2010}$	Gini Keyword$_{2010}$	HDI-Class$_{VeryHigh}$	HDI-Class$_{High}$	HDI-Class$_{M\&L}$	ΔGDP	ΔEducation Index	Top5Downloads_weighed
Downloads	1	−0.048	0.056	0.023	0.016	−0.024	0.043	−0.016	0.017	0.034	0.362
Share Not Available		1	−0.061	−0.158	−0.087	−0.000	−0.000	−0.000	0.000	−0.000	−0.005
Expected Citations$_{2010}$			1	0.003	0.093	0.000	−0.000	0.000	−0.000	0.000	0.003
Gini Country$_{2010}$				1	0.046	0.000	−0.000	0.000	0.000	0.000	−0.019
Gini Keyword$_{2010}$					1	0.000	−0.000	0.000	0.000	0.000	0.024
HDI-Class$_{VeryHigh}$						1	−0.516	−0.550	−0.245	−0.243	−0.059
HDI-Class$_{High}$							1	−0.432	0.030	−0.081	0.084
HDI-Class$_{M\&L}$								1	0.229	0.335	−0.020
ΔGDP									1	0.109	0.036
ΔEducation Index										1	0.072
Top5Downloads_weighed											1

TABLE A3 OLS regression

	(I)	(II)	(III)	(IV)	(V)	(VI)
(Intercept)	−0.094	0.017	1.885	4.767	12.975	4.949
	(1.100)	(22.985)	(22.968)	(23.563)	(23.518)	(23.468)
Downloads	0.193***		0.185***		0.087***	
	(0.017)		(0.017)		(0.019)	
Share Not Available	0.650	−0.459	0.457	−0.172	0.191	−0.378
	(1.908)	(2.062)	(2.062)	(2.115)	(2.112)	(2.104)
Expected Citations_{2010}	−0.099	0.034	−0.066	0.013	0.230	0.258
	(0.164)	(0.175)	(0.175)	(0.181)	(0.182)	(0.182)
Gini Country_{2010}		−4.001	−4.745	−4.445	−1.338	−2.300
		(4.617)	(4.614)	(4.734)	(4.728)	(4.716)
Gini Keyword_{2010}		−4.778	−6.266	−5.794	−19.284	−12.071
		(22.904)	(22.887)	(23.484)	(23.444)	(23.391)
$\text{HDI-Class}_{\text{High}}$		19.128***	18.562***			9.667***
		(1.387)	(1.387)			(1.474)
$\text{HDI-Class}_{\text{M\&L}}$		9.626***	9.608***			2.573
		(1.345)	(1.344)			(1.538)
ΔGDP				0.177***	0.200***	0.180***
				(0.021)	(0.021)	(0.022)
ΔEducation Index				0.477***	0.318***	0.294**
				(0.092)	(0.092)	(0.097)
ΔGDP:Downloads				0.007***		
				(0.000)		
$\text{Expected Citations}_{2010}$:Downloads				−0.007***		
				(0.002)		
Top5Downloads_weighed					0.114***	−0.081***
					(0.006)	(0.017)
$\text{Expected Citations}_{2010}$: Top5Downloads_weighed					−0.009***	−0.007***
					(0.001)	(0.001)
$\text{HDI-Class}_{\text{High}}$:Top5Downloads_weighed						0.225***
						(0.016)
$\text{HDI-Class}_{\text{M\&L}}$:Top5Downloads_weighed						0.132***
						(0.007)
N	79,704	74,358	74,358	72,414	72,414	72,414
AIC	−443591.5571	−409605.3857	−409719.3936	−397193.2874	−397490.8729	−397847.8903
F-test	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001

Note: Estimates and standard errors (in parentheses) reported ×10,000.
*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

TABLE A4 OLS regression with heteroscedasticity robust White standard errors

	(I)	(II)	(III)	(IV)	(V)	(VI)
(Intercept)	−0.094	0.017	1.885	4.767	12.975	4.949
	(1.108)	(27.943)	(27.942)	(28.637)	(28.602)	(28.655)
Downloads	0.193***		0.185***		0.087	
	(0.051)		(0.050)		(0.047)	
Share Not Available	0.650	−0.459	0.457	−0.172	0.191	−0.378
	(1.944)	(2.097)	(2.094)	(2.155)	(2.136)	(2.148)
Expected Citations$_{2010}$	−0.099	0.034	−0.066	0.013	0.230	0.258
	(0.140)	(0.148)	(0.149)	(0.157)	(0.155)	(0.156)
Gini Country$_{2010}$		−4.001	−4.745	−4.445	−1.338	−2.300
		(3.571)	(3.566)	(3.663)	(3.768)	(3.787)
Gini Keyword$_{2010}$		−4.778	−6.266	−5.794	−19.284	−12.071
		(27.833)	(27.833)	(28.583)	(28.562)	(28.540)
HDI-Class$_{High}$		19.128***	18.562***			9.667***
		(1.555)	(1.536)			(1.420)
HDI-Class$_{M\&L}$		9.626***	9.608***			2.573*
		(1.275)	(1.276)			(1.209)
ΔGDP				0.177***	0.200***	0.180***
				(0.023)	(0.023)	(0.024)
ΔEducation Index				0.477***	0.318***	0.294***
				(0.063)	(0.065)	(0.053)
ΔGDP:Downloads				0.007***		
				(0.002)		
Expected Citations$_{2010}$:Downloads				−0.007		
				(0.008)		
Top5Downloads_weighed					0.114***	−0.081
					(0.022)	(0.053)
Expected Citations$_{2010}$: Top5Downloads_weighed					−0.009	−0.007
					(0.005)	(0.005)
HDI-Class$_{High}$:Top5Downloads_weighed						0.225***
						(0.052)
HDI-Class$_{M\&L}$:Top5Downloads_weighed						0.132**
						(0.051)
N	79,704	74,358	74,358	72,414	72,414	72,414
AIC	−443591.557	−409605.386	−409719.394	−397193.287	−397490.873	−397847.89
F test	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001

Note: Estimates and standard errors (in parentheses) reported ×10,000.
*$p < .05$; **$p < .01$; ***$p < .001$.

TABLE A5 OLS regression with regional dummy variables

	(III)	(III_{China})	(VI)	(VI_{China})
(Intercept)	−119.790***	−119.906***	−106.653***	−113.569***
	(22.099)	(21.762)	(24.023)	(23.834)
Downloads	0.197***	0.047**		
	(0.017)	(0.017)		
Share Not Available	0.253	−0.446	−0.368	−0.375
	(1.938)	(1.909)	(2.089)	(2.072)
Expected $Citations_{2010}$	−0.097	−0.021	0.218	0.232
	(0.165)	(0.162)	(0.181)	(0.179)
Gini $Country_{2010}$	−5.034	−4.468	−1.515	−2.242
	(4.337)	(4.271)	(4.681)	(4.644)
Gini $Keyword_{2010}$	−7.864	−6.740	−20.081	−12.493
	(21.515)	(21.187)	(23.218)	(23.035)
East Asia and Pacific	145.291***	128.906***	127.795***	125.581***
	(5.003)	(4.938)	(6.354)	(6.313)
Europe and Central Asia	126.570***	125.265***	120.150***	120.547***
	(4.891)	(4.816)	(6.140)	(6.091)
Latin America and Caribbean	134.147***	132.696***	124.600***	126.765***
	(4.949)	(4.874)	(6.248)	(6.199)
Middle East and North Africa	136.040***	135.542***	129.106***	129.044***
	(5.036)	(4.959)	(6.282)	(6.232)
South Asia	137.282***	136.664***	126.296***	129.898***
	(5.431)	(5.349)	(6.717)	(6.665)
Sub-Saharan Africa	132.817***	131.245***	124.708***	126.080***
	(4.928)	(4.853)	(6.197)	(6.148)
China-Dummy		345.028***		298.474***
		(6.941)		(10.701)
ΔGDP			0.070**	−0.016
			(0.024)	(0.024)
ΔEducation Index			0.249*	0.199*
			(0.099)	(0.098)
Top5Downloads_weighed			−0.001	0.005
			(0.029)	(0.029)
Expected $Citations_{2010}$:Top5Downloads_weighed			−0.007***	−0.008***
			(0.001)	(0.001)
East Asia and Pacific:Top5Downloads_weighed			0.264***	0.065*
			(0.029)	(0.031)
Europe and Central Asia:Top5Downloads_weighed			0.058	0.049
			(0.034)	(0.033)
Latin America and Caribbean: Top5Downloads_weighed			0.292***	0.259***
			(0.074)	(0.073)
Middle East and North Africa: Top5Downloads_weighed			0.058*	0.055
			(0.029)	(0.029)

TABLE A5 (Continued)

	(III)	(III_{China})	(VI)	(VI_{China})
South Asia:Top5Downloads_weighed			0.060*	0.056
			(0.030)	(0.030)
Sub-Saharan Africa:Top5Downloads_weighed			−0.12	−0.051
			(0.244)	(0.242)
China-Dummy:Top5Downloads_weighed				0.036*
				(0.017)
N	79,218	79,218	72,414	72,414
AIC	−441274.6825	−443705.9707	−398949.0798	−400101.496
F test	<0.001	<0.001	<0.001	<0.001

Note: Estimates and standard errors (in parentheses) reported ×10,000. OLS regressions with regional dummy variables according to the World Bank's regional classification. Countries classified as "North America" serve as category of reference.
*$p < .05$; **$p < .01$; ***$p < .001$.

TABLE A6 Quantile regression estimates for Model VI variables

	$\tau = 0.05$	$\tau = 0.1$	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$	$\tau = 0.9$	$\tau = 0.95$
(Intercept)	−198.554***	−82.211***	0.000	0.000	0.000	80.927***	197.077***
	(2.145)	(0.939)	(0.019)	(0.000)	(0.151)	(1.271)	(7.349)
Share Not Available	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	(0.014)	(0.037)	(0.001)	(0.000)	(0.010)	(0.042)	(0.240)
Expected Citations_{2010}	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	(0.006)	(0.009)	(0.001)	(0.000)	(0.004)	(0.012)	(0.024)
Gini Country_{2010}	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	(0.053)	(0.073)	(0.016)	(0.000)	(0.060)	(0.093)	(0.512)
Gini Keyword_{2010}	0.000	0.000	0.000	0.000	0.000	0.000	−15.613*
	(0.779)	(0.556)	(0.014)	(0.000)	(0.136)	(1.180)	(7.568)
$\text{HDI-Class}_{\text{High}}$	198.554***	82.211***	0.000	0.000	0.000	−80.927***	−172.776***
	(1.989)	(0.733)	(0.006)	(0.000)	(0.016)	(0.429)	(1.841)
$\text{HDI-Class}_{\text{M\&L}}$	198.554***	82.211***	0.000	0.000	0.000	−80.927***	−181.464***
	(1.989)	(0.732)	(0.004)	(0.000)	(0.013)	(0.413)	(1.057)
ΔGDP	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	(0.001)	(0.001)	(0.000)	(0.000)	(0.001)	(0.001)	(0.004)
ΔEducation Index	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	(0.002)	(0.003)	(0.000)	(0.000)	(0.001)	(0.003)	(0.014)
Top5Downloads_weighed	−4.212***	−2.883***	−0.832***	0.000	0.720***	1.894***	2.387***
	(0.079)	(0.083)	(0.014)	(0.000)	(0.030)	(0.073)	(0.126)
$\text{Estimated Citations}_{2010}$:Downloads	0.023**	0.010	−0.001	0.000	−0.009***	−0.022***	−0.041***
	(0.007)	(0.006)	(0.002)	(0.000)	(0.002)	(0.006)	(0.006)
$\text{HDI-Class}_{\text{High}}$: Top5Downloads_weighed	3.534***	2.613***	0.829***	0.000	−0.425***	−0.877***	−0.727***
	(0.085)	(0.087)	(0.015)	(0.000)	(0.034)	(0.086)	(0.128)
$\text{HDI-Class}_{\text{M\&L}}$: Top5Downloads_weighed	3.702***	2.661***	0.832***	0.000	−0.608***	−1.450***	−1.489***
	(0.072)	(0.077)	(0.014)	(0.000)	(0.027)	(0.065)	(0.121)

Note: Quantile regressions of Model VI for the 5th, 10th, 25th, 50th, 75th, 90th, and 95th percentile, respectively. Estimates and Bootlegged standard errors (in parentheses) reported $\times 10^4$.

*$p < 0.1$; **$p < .05$; ***$p < .01$.