Abstract

The website Sci-Hub enables users to download PDF versions of scholarly articles, including many articles that are paywalled at their journal's site. Sci-Hub has grown rapidly since its creation in 2011, but the extent of its coverage was unclear. Here we report that, as of March 2017, Sci-Hub's database contains 68.9% of the 81.6 million scholarly articles registered with Crossref and 85.1% of articles published in toll access journals. We find that coverage varies by discipline and publisher, and that Sci-Hub preferentially covers popular, paywalled content. For toll access articles, we find that Sci-Hub provides greater coverage than the University of Pennsylvania, a major research university in the United States. Green open access to toll access articles via licit services, on the other hand, remains quite limited. Our interactive browser at https://greenelab.github.io/scihub allows users to explore these findings in more detail. For the first time, nearly all scholarly literature is available gratis to anyone with an Internet connection, suggesting the toll access business model may become unsustainable.

Introduction

Recent estimates suggest paywalls on the web limit access to three-quarters of scholarly literature [1–3]. The open access movement strives to remedy this situation [4]. After decades of effort by the open access community [5], nearly 50% of newly published articles are available without paywalls [1,6,7].

Despite these gains, access to scholarly literature remains a pressing global issue. Foremost, widespread subscription access remains restricted to institutions, such as universities or medical centers. Smaller institutions or those in the developing world often have poor access to scholarly literature [8–10]. As a result, only a tiny percentage of the world's population has been able to access much of the scholarly literature, despite the fact that the underlying research was often publicly or philanthropically funded. Compounding the problem is that publications have historically been the primary, if not sole, output of scholarship. Although copyright does not apply to ideas, journals leverage the copyright covering an article's prose, figures, and typesetting to effectively paywall its knowledge.

Since each article is unique, libraries cannot substitute one journal subscription for another without depriving their users of potentially crucial access. As a result, the price of journal subscriptions has grown at a faster rate than inflation for several decades [11], leading to an ever-present "serials crisis" that has pushed library budgets to their brink while diverting funds from other services [12]. Meanwhile, publishing has trended towards oligopoly [13], with nondisclosure clauses obfuscating price information among subscribers [14] while publishers profit immensely [15–17]. Price increases have persisted over the last decade [18–20]. For example, EBSCO estimates that per-journal subscription costs increased by 25% from 2013–2017, with annual subscription to a journal for research libraries now averaging $1,396 [21].

In this study, we use the term "toll access" (also known as "closed access") to refer to paywalled literature [22]. On the other hand, we refer to literature that is free to read as "open access". Furthermore, we discuss two variants of open access: "libre" and "gratis" [22,23]. Libre open access refers to literature that is openly licensed to allow reuse. Gratis open access refers to literature that is accessible free of charge, although permission barriers may remain (usually due to copyright) [24].

The website Sci-Hub, now in its sixth year of existence, provides gratis access to scholarly literature, despite the continued presence of paywalls. Sci-Hub brands itself as "the first pirate website in the world to provide mass and public access to tens of millions of research papers." The website, started in 2011, is run by Alexandra Elbakyan, a graduate student and native of Kazakhstan who now resides in Russia [25,26]. Elbakyan describes herself as motivated to provide universal access to knowledge [27–29].

Sci-Hub does not restrict itself to only openly licensed content. Instead, it retrieves and distributes scholarly literature without regard to copyright. **Readers should note that, in many jurisdictions, use of Sci-Hub may constitute copyright infringement. Users of Sci-Hub do so at their own risk. This study is not an endorsement of using Sci-Hub, and its authors and publishers accept no responsibility on behalf of readers. There is a possibility that Sci-Hub users — especially those not using privacy-enhancing services such as Tor — could have their usage history unmasked and face legal or reputational consequences.**

Sci-Hub is currently served at domains including https://sci-hub.hk, https://sci-hub.la, https://sci-hub.mn, https://sci-hub.name, https://sci-hub.tv, and https://sci-hub.tw, as well as at scihub22266oqcxt.onion — a Tor Hidden Service [30]. Elbakyan described the project's technical scope in July 2017 [31]: "Sci-Hub technically is by itself a repository, or a library if you like, and not a search engine for some other repository. But of course, the most important part in Sci-Hub is not a repository, but the script that can download papers closed behind paywalls."

One method Sci-Hub uses to bypass paywalls is by obtaining leaked authentication credentials for educational institutions [31]. These credentials enable Sci-Hub to use institutional networks as proxies and gain subscription journal access. While the open access movement has progressed slowly [32], Sci-Hub represents a seismic shift in access to scholarly literature. Since its inception, Sci-Hub has experienced sustained growth, with spikes in interest and awareness driven by legal proceedings, service outages, news coverage, and social media (Figure 1 and 1—figure supplement 1). Here we investigate the extent to which Sci-Hub provides access to scholarly literature. If Sci-Hub's coverage is sufficiently broad, then a radical shift may be underway in how individuals access scholarly literature.

In Figure 1, The ⓛⓔⓣⓣⓔⓡⓢ refer to the following events:

- Ⓐ Created by Alexandra Elbakyan, the Sci-Hub website goes live on September 5, 2011.
- Ⓑ Several LibGen domains go down when their registration expires, allegedly due to a longtime site administrator passing away from cancer [36].
- Ⓒ Elsevier files a civil suit against Sci-Hub and LibGen — at the respective domains sci-hub.org and libgen.org — in the U.S. District Court for the Southern District of New York [37,38]. The complaint seeks a "prayer for relief" that includes domain name seizure, damages, and "an order disgorging Defendants' profits".
- Ⓓ Elsevier is granted a preliminary injunction to suspend domain names and restrain the site operators from distributing Elsevier's copyrighted works [39,40]. Shortly after, Sci-Hub and

LibGen resurface at alternative domains outside of U.S. court jurisdiction, including on the dark web [26,41].

- ⒺThe article “Meet the Robin Hood of Science” by Simon Oxenham spurs a wave of attention and news coverage on Sci-Hub and Alexandra Elbakyan [42], culminating in *The New York Times* asking “Should all research papers be free?” [43].
- ⒻThe article “Who’s downloading pirated papers? Everyone” by John Bohannon shows Sci-Hub is used worldwide, including in developed countries [44,45]. These findings spark debate among scholars, with a large contingent of scientists supporting Sci-Hub’s mission [46,47].
- ⒼAlexandra Elbakyan is named one of “*Nature*’s 10”, which featured “ten people who mattered” in 2016 [48]. Written by Richard Van Noorden, the story profiles Alexandra and includes an estimate that Sci-Hub serves “3% of all downloads from science publishers worldwide.”
- ⒽThe court finds that Alexandra Elbakyan, Sci-Hub, and LibGen are “liable for willful copyright infringement” in a default judgment, since none of the defendants answered Elsevier’s complaint [49–51]. The court issues a permanent injunction and orders the defendants to pay Elsevier $15 million, or $150,000 for each of 100 copyrighted works. The statutory damages, which the defendants do not intend to pay, now bear interest.
- ⒾThe American Chemical Society files suit against Sci-Hub in the U.S. District Court for the Eastern District of Virginia. Their “prayer for relief” requests that Internet search engines and Internet service providers “cease facilitating access” to Sci-Hub [52,53].
- ⒿThe version 1 preprint of this study is published [54], generating headlines such as *Science*’s “subscription journals are doomed” [55] and *Inside Higher Ed*’s “Inevitably Open” [56].
- ⓀSci-Hub blocks access to Russian IP addresses due to disputes with the Russian Scientific establishment and the naming of a newly discovered parasitoid wasp species, *Idiogramma elbakyanae*, after Alexandra Elbakyan [57]. Four days later, Sci-Hub restores access after receiving “many letters of support from Russian researchers” [59].
- ⓁThe court rules on the American Chemical Society suit, ordering Sci-Hub to pay $4.8 million in damages and that “any person or entity in active concert or participation” with Sci-Hub “including any Internet search engines, web hosting and Internet service providers, domain name registrars, and domain name registries, cease facilitating access” [60,61]. Within five weeks, the domains sci-hub.io, sci-hub.ac, sci-hub.cc, and sci-hub.bz were suspended by their respective domain name registries [62], leaving only the Tor hidden service and several newly-registered/revealed domains in operation.

Past research sheds some light on Sci-Hub’s reach. From the Spring of 2013 until the end of 2014, Sci-Hub relied on the Library Genesis (LibGen) scimag repository to store articles [31]. Whenever a user requested an article, Sci-Hub would check LibGen for a copy. If the article was not in LibGen, Sci-Hub would fetch the article for the user and then upload it to LibGen. Cabanac compared the number of articles in the LibGen scimag database at the start of 2014 to the total number of Crossref DOIs, estimating that LibGen contained 36% of all published scholarly articles [34]. Coverage was higher for several prominent publishers: 77% for Elsevier, 73% for Wiley, and 53% for Springer (prior to its merger with Macmillan / Nature [63]).

Later, Bohannon analyzed six months of Sci-Hub's server access logs, starting in September 2015 [44]. He found a global pattern of usage. Based on these logs, Gardner, McLaughlin, and Asher estimated the ratio of publisher downloads to Sci-Hub downloads within the U.S. for several publishers [64]. They estimated this ratio at 20:1 for the Royal Society of Chemistry and 48:1 for Elsevier. They also noted that 25% of Sci-Hub downloads in the U.S. were for articles related to clinical medicine. Greshake also analyzed the logs to identify per capita Sci-Hub usage [65]. Portugal, Iran, Tunisia, and Greece had the highest usage, suggesting Sci-Hub is preferentially used in countries with poor institutional access to scholarly literature. In a subsequent study, Greshake found especially high Sci-Hub usage in chemistry, with 12 of the top 20 requested journals specializing in chemistry [66,67].

Since 2015, Sci-Hub has operated its own repository, distinct from LibGen. On March 19, 2017, Sci-Hub released the list of DOIs for articles in its database. Greshake retrieved metadata for 77% of Sci-Hub DOIs [66,67]. He found that 95% of articles in Sci-Hub were published after 1950. Sci-Hub requests were even more skewed towards recent articles, with only 5% targeting articles published before 1983. Greshake's study did not incorporate a catalog of all scholarly literature. This study analyzes Sci-Hub's catalog in the context of all scholarly literature and thus assesses coverage. In other words, what percentage of articles in a given domain does Sci-Hub have in its repository?

Results

To define the extent of the scholarly literature, we relied on DOIs from the Crossref database, as downloaded on March 21, 2017. We define the "scholarly literature" as 81,609,016 texts identified by their DOIs. We refer to these texts as "articles", although Sci-Hub encompasses a range of text types, including book chapters, conference papers, and journal front matter. To assess the articles available from Sci-Hub, we relied on a list of DOIs released by Sci-Hub on March 19, 2017. All DOIs were lowercased to be congruent across datasets (see Methods). Sci-Hub's offerings included 56,246,220 articles from the corpus of scholarly literature, equating to 68.9% of all articles.

Coverage by article type

Each article in Crossref's database is assigned a type. Figure 2 shows coverage by article type. The scholarly literature consisted primarily of journal articles, for which Sci-Hub had 77.8% coverage. Sci-Hub's coverage was also strong for the 5 million proceedings articles at 79.7%. Overall coverage suffered from the 10 million book chapters, where coverage was poor (14.2%). The remaining Crossref types were uncommon, and hence contributed little to overall coverage.

Coverage by journal

We defined a comprehensive set of scholarly publishing venues, referred to as "journals", based on the Scopus database. In reality, these include conferences with proceedings as well as book series. For inclusion in this analysis, each required an ISSN and at least one article as part of the Crossref-derived catalog of scholarly literature. Accordingly, our catalog consisted of 23,037 journals encompassing 56,755,671 articles. Of these journals, 4,598 (20.0%) were inactive (i.e. no longer publishing articles), and 2,933 were open access (12.7%). Only 70 journals were inactive and also open access.

We calculated Sci-Hub's coverage for each of the 23,037 journals (examples in Table 1). A complete journal coverage table is available in our Sci-Hub Stats Browser. The Browser also provides views for each journal and publisher with detailed coverage and access-log information.

In general, a journal's coverage was either nearly complete or near zero (Figure 3). As a result, relatively few journals had coverage between 5–75%. At the extremes, 2,574 journals had zero coverage in Sci-Hub, whereas 2,095 journals had perfect coverage. Of zero-coverage journals, 22.2% were inactive, and 27.9% were open access. Of perfect-coverage journals, 81.6% were inactive, and 2.0% were open access. Hence, inactive, toll access journals make up the bulk of perfect-coverage journals.

Next, we explored article coverage according to journal attributes (Figure 4). Sci-Hub covered 83.1% of the 56,755,671 articles that were attributable to a journal. Articles from inactive journals had slightly lower coverage than active journals (77.3% versus 84.1%). Strikingly, coverage was substantially higher for articles from toll rather than open access journals (85.1% versus 48.3%). Coverage did vary by subject area, with the highest coverage in chemistry at 93.0% and the lowest coverage in computer science at 76.3%. Accordingly, no discipline had coverage below 75%. See Figure 4—figure supplement 1 for coverage according to a journal's country of publication.

We also evaluated whether journal coverage varied by journal impact. We assessed journal impact using the 2015 CiteScore, which measures the average number of citations that articles published in 2012–2014 received during 2015. Highly cited journals tended to have higher coverage in Sci-Hub (Figure 9A). The 1,734 least cited journals (lowest decile) had 40.9% coverage on average, whereas the 1,733 most cited journals (top decile) averaged 90.0% coverage.

Coverage by publisher

Next, we evaluated coverage by publisher (Figure 5, full table online). The largest publisher was Elsevier, with 13,115,639 articles from 3,410 journals. Sci-Hub covered 96.9% of Elsevier articles. For the eight publishers with more than one million articles, the following coverage was observed: 96.9% of Elsevier, 89.7% of Springer Nature, 94.7% of Wiley-Blackwell, 92.6% of Taylor & Francis, 79.4% of Wolters Kluwer, 88.3% of Oxford University Press, 90.9% of SAGE, and 98.8% of American Chemical Society articles. In total, 3,832 publishers were represented in the journal catalog. The coverage distribution among publishers resembled the journal coverage distribution, with most publishers occupying the extremities (Figure 3). Sci-Hub had zero coverage for 1,249 publishers, and complete coverage for 341 publishers.

Coverage by year

Next, we investigated coverage based on the year an article was published (Figure 6). For most years since 1850, annual coverage is between 60–80%. However, there is a dropoff in coverage, starting in 2010, for recently published articles. For example, 2016 coverage was 56.0% and 2017 coverage (for part of the year) was 45.3%. One factor is that it can take some time for Sci-Hub to retrieve articles following their publication, as many articles are not downloaded until requested by a user. Another possible factor is that some publishers are now deploying more aggressive measures to deter unauthorized article downloads [68,69], making recent articles less accessible.

In addition, the prevalence of open access has been increasing, while Sci-Hub preferentially covers articles in toll access journals. Figure 6—figure supplement 1 tracks yearly coverage separately for articles in toll and open access journals. Toll access coverage exceeded 80% every year since 1950 except for 2016 and 2017. For both toll and open articles, the recent dropoff in coverage appears to begin in 2014 (Figure 6—figure supplement 1) compared to 2010 when calculated across all articles (Figure 6). We speculate this discrepancy results from the proliferation of obscure, low-quality journals over the last decade [70], as these journals generally issue DOIs but are not indexed in Scopus, and therefore would be included in Figure 6 but not in Figure 6—figure supplement 1. In addition to having limited readership demand, these journals are generally open access, and thus less targeted by Sci-Hub.

Sci-Hub's coverage of 2016 articles in open access journals was just 32.7% compared to 78.8% for articles in toll access journals (Figure 6—figure supplement 1). Upon further investigation, we discovered that in June 2015, Sci-Hub ceased archiving articles in *PeerJ*, *eLife*, and PLOS journals, although they continued archiving articles in other open access journals such as *Scientific Reports*, *Nature Communications*, and BMC-series journals. Sci-Hub currently redirects requests for these delisted journals to the publisher's site, unless it already possesses the article, in which case it serves the PDF. These findings suggest Sci-Hub prioritizes circumventing access barriers rather than creating a single repository containing every scholarly article.

Coverage by category of access status

In the previous analyses, open access status was determined at the journal level according to Scopus. This category of access is frequently referred to as "gold" open access, meaning that all articles from the journal are available gratis. However, articles in toll access journals may also be available without charge. Adopting the terminology from the recent "State of OA" study [1], articles in toll access journals may be available gratis from the publisher under a license that permits use (termed "hybrid") or with all rights reserved (termed "bronze"). Alternatively, "green" articles are paywalled on the publisher's site, but available gratis from an open access repository (e.g. a pre- or post-print server, excluding Sci-Hub and academic social networks).

The State of OA study determined the access status of 290,120 articles using the oaDOI utility (see Methods). Figure 7 shows Sci-Hub's coverage for each category of access status. In line with our findings on the entire Crossref article catalog where Sci-Hub covered 49.1% of articles in open access journals, Sci-Hub's coverage of gold articles in the State of OA dataset was 49.2%. Coverage of the 165,340 closed articles was 90.4%.

Sci-Hub's coverage was higher for closed and green articles than for hybrid or bronze articles. Furthermore, Sci-Hub's coverage of closed articles was similar to its coverage of green articles (Figure 7). These findings suggest a historical pattern where users resort to Sci-Hub after encountering a paywall but before checking oaDOI or a search engine for green access. As such, Sci-Hub receives requests for green articles, triggering it to retrieve green articles at a similar rate to closed articles. However, hybrid and bronze articles, which are available gratis from their publisher, are requested and thus retrieved at a lower rate.

Coverage of Penn Libraries

As a benchmark, we decided to compare Sci-Hub's coverage to the access provided by a major research library. Since we were unaware of any studies that comprehensively profiled library access to scholarly articles, we collaborated with Penn Libraries to assess the extent of access available at the University of Pennsylvania (Penn). Penn is a private research university located in Philadelphia and founded by the open science pioneer Benjamin Franklin in 1749. It is one of the world's wealthiest universities, with an endowment of over $10 billion. According to the Higher Education Research and Development Survey, R&D expenditures at Penn totaled $1.29 billion in 2016, placing it third among U.S. colleges and universities. In 2017, Penn Libraries estimates that it spent $13.13 million on electronic resources, which includes subscriptions to journals and ebooks. During this year, its users accessed 7.3 million articles and 860 thousand ebook chapters, averaging a per-download cost of $1.61.

Penn Libraries uses the Alma library resource management system from Ex Libris. Alma includes an OpenURL resolver, which the Penn Libraries use to provide a service called PennText for looking up scholarly articles. PennText indicates whether an article's fulltext is available online, taking into account Penn's digital subscriptions. Using API calls to PennText's OpenURL resolver, we retrieved Penn's access status for the 290,120 articles analyzed by the State of OA study (see the greenelab/library-access repository). We randomly selected 500 of these articles to evaluate manually and assessed whether their fulltexts were available from within Penn's network as well as from outside of any institutional network. We defined access as fulltext availability at the location redirected to by an article's DOI, without providing any payment, credentials, or login information. This definition is analogous to the union of oaDOI's gold, hybrid, and bronze categories.

Using these manual access calls, we found PennText correctly classified access 88.2% [85.2%–90.8%] of the time (bracketed ranges refer to 95% confidence intervals calculated using Jeffreys interval for binomial proportions [71]). PennText claimed to have access to 422 of the 500 articles [81.0%–87.4%]. When PennText asserted access, it was correct 94.8% [92.4%–96.6%] of the time. However, when PennText claimed no access, it was only correct for 41 of 78 articles [41.6%–63.4%]. This error rate arose because PennText was not only unaware of Penn's access to 23 open articles, but also unaware of Penn's subscription access to 14 articles. Despite these issues, PennText's estimate of Penn's access at 84.4% did not differ significantly from the manually evaluated estimate of 87.4% [84.3%–90.1%]. Nonetheless, we proceed by showing comparisons for both the 500 articles with manual access calls as well as the 290,120 articles with PennText calls.

Coverage combining access methods

In practice, readers of the scholarly literature likely use a variety of methods for access. Figure 8 compares several of these methods, as well as their combinations. Users without institutional access may simply attempt to view an article on its publisher's site. Based on our manual evaluation of 500 articles, we found 34.8% [30.7%–39.1%] of articles were accessible this way. The remaining 326 articles that were not accessible from their publisher's site are considered toll access. oaDOI — a utility that redirects paywalled DOIs to gratis, licit versions, when possible [1] — was able to access 15.3% [11.7%–19.5%] of these toll access articles, indicating that green open access is still limited in its reach. This remained true on the full set of 208,786 toll access articles from the State of OA dataset, where oaDOI only provided access to 12.4% [12.3%–12.6%]. Although oaDOI's overall access rate was 37.0% [36.8%–37.2%], this access consisted largely of gold, hybrid, and bronze articles, whereby gratis access is provided by the publisher.

Sci-Hub and Penn had similar coverage on all articles: 85.2% [81.9%–88.1%] versus 87.4% [84.3%–90.1%] on the manual article set and 84.8% [84.7%–84.9%] versus 84.4% [84.3%–84.5%] on the larger but automated set. However, when considering only toll access articles, Sci-Hub's coverage exceeds Penn's: 94.2% [91.2%–96.3%] versus 80.7% [76.1%–84.7%] on the manual set and 90.7% [90.5%–90.8%] versus 83.5% [83.4%–83.7%] on the automated set. This reflects Sci-Hub's focus on paywalled articles. In addition, Sci-Hub's coverage is a lower bound for its access rate, since it can retrieve articles on demand, so in practice Sci-Hub's access to toll access articles could exceed Penn's by a higher margin. Remarkably, Sci-Hub provided greater access to paywalled articles than a leading research university spending millions of dollars per year on subscriptions. However, since Sci-Hub is able to retrieve articles through many university networks, it is perhaps unsurprising that its coverage would exceed that of any single university.

Combining access methods can also be synergistic. Specifically when including open access articles, combining Sci-Hub's repository with oaDOI's or Penn's access increased coverage from around 85% to 95%. The benefits of oaDOI were reduced when only considering toll access articles, where oaDOI only improved Sci-Hub's or Penn's coverage by approximately 1%. On toll access articles, Penn's access appeared to complement Sci-Hub's. Together, Sci-Hub's repository and Penn's access covered approximately 96% of toll access articles [95.0%–98.6% (manual set), 95.9%–96.1% (automated set)]. Our findings suggest that users with institutional subscriptions comparable to those at Penn as well as knowledge of oaDOI and Sci-Hub are able to access over 97% of all articles [96.7%–99.1% (manual set), 97.3%–97.5% (automated set)], online and without payment.

Coverage of recently cited articles

The coverage metrics presented thus far give equal weight to each article. However, we know that article readership and by extension Sci-Hub requests are not uniformly distributed across all articles. Instead, most articles receive little readership, with a few articles receiving great readership. Therefore, we used recent citations to estimate Sci-Hub's coverage of articles weighted by user needs.

We identified 7,312,607 outgoing citations from articles published since 2015. 6,657,410 of the recent citations (91.0%) referenced an article that was in Sci-Hub. However, if only considering the 6,264,257 citations to articles in toll access journals, Sci-Hub covered 96.2% of recent citations. On the other hand, for the 866,115 citations to articles in open access journals, Sci-Hub covered only 62.3%.

Sci-Hub access logs

Sci-Hub released article access records from its server logs, covering 165 days from September 2015 through February 2016 [33,44,45]. After processing, the logs contained 26,984,851 access events. Hence, Sci-Hub provided access to an average of 164,000 valid requests per day in late 2015–early 2016.

In the first version of this study [54], we mistakenly treated the log events as requests rather than downloads. Fortunately, Sci-Hub reviewed the preprint in a series of tweets, and pointed out the error, stating “in Sci-Hub access logs released previous year, all requests are resolved requests, i.e. user successfully downloaded PDF with that DOI … unresolved requests are not saved”. Interestingly however, 198,600 access events from the logs pointed to DOIs that were not in Sci-Hub’s subsequent DOI catalog. 99.1% of these events — corresponding to DOIs logged as accessed despite later being absent from Sci-Hub — were for book chapters. Upon further investigation, we identified several DOIs in this category that Sci-Hub redirected to LibGen book records as of September 2017. The LibGen landing pages were for the entire books, which contained the queried chapters, and were part of LibGen’s book (not scimag) collection. The explanation that Sci-Hub outsources some book access to LibGen (and logged such requests as accessed) is corroborated by Elbakyan’s statement that [31]: “Currently, the Sci-Hub does not store books, for books users are redirected to LibGen, but not for research papers. In future, I also want to expand the Sci-Hub repository and add books too.” Nonetheless, Sci-Hub’s catalog contains 72.4% of the 510,760 distinct book chapters that were accessed according to the logs. Therefore, on a chapter-by-chapter basis, Sci-Hub does already possess many of the requested scholarly books available from LibGen.

We computed journal-level metrics based on average article downloads. The “visitors” metric assesses the average number of IP addresses that accessed each article published by a journal during the 20 months preceding September 2015 (the start date of the Sci-Hub logs). In aggregate, articles from toll access journals averaged 1.30 visitors, whereas articles from open access journals averaged 0.25 visitors. Figure 9B shows that articles from highly cited journals were visited much more frequently on average. Articles in the least cited toll access journals averaged almost zero visitors, compared to approximately 15 visitors for the most cited journals. In addition, Figure 9B shows that articles in toll access journals received many times more visitors than those in open access journals, even after accounting for journal impact. One limitation of using this analysis to judge Sci-Hub’s usage patterns is that we do not know to what extent certain categories of articles were resolved (and thus logged) at different rates.

Discussion

Sci-Hub’s repository contained 69% of all scholarly articles with DOIs. Coverage for the 54.5 million articles attributed to toll access journals — which many users would not otherwise be able to access — was 85.1%. Since Sci-Hub can retrieve, in real time, requested articles that are not in its database, our coverage figures are a lower bound. Furthermore, Sci-Hub preferentially covered popular, paywalled articles. We find that 91.0% of citations since 2015 were present in Sci-Hub’s repository, which increased to 96.2% when excluding citations to articles in open access journals. Journals with very low (including zero) coverage tended to be obscure, less cited venues, while average coverage of the most cited journals exceeded 90%.

We find strong evidence that Sci-Hub is primarily used to circumvent paywalls. In particular, users accessed articles from toll access journals much more frequently than open access journals. Additionally, within toll access journals, Sci-Hub provided higher coverage of articles in the closed and green categories (paywalled by the publisher) as opposed to the hybrid and bronze categories (available gratis from the publisher). Accordingly, many users likely only resort to Sci-Hub when access through a commercial database is cumbersome or costly. Finally, we observed evidence that Sci-Hub's primary operational focus is circumventing paywalls rather than compiling all literature, as archiving was deactivated in 2015 for several journals that exemplify openness. Attesting to its success in this mission, Sci-Hub's database already contains more toll access articles than are immediately accessible via the University of Pennsylvania, a leading research university.

Judging from donations, many users appear to value Sci-Hub's service. In the past, Sci-Hub accepted donations through centralized and regulated payment processors such as PayPal, Yandex, WebMoney, and QiQi [38,72]. Now however, Sci-Hub only advertises donation via Bitcoin, presumably to avoid banking blockades or government seizure of funds. Since the ledger of bitcoin transactions is public, we can evaluate the donation activity to known Sci-Hub addresses (`1K4t2vSBSS2xFjZ6PofYnbgZewjeqbG1TM`, `14ghuGKDAPdEcUQN4zuzGwBUrhQgACwAyA`, `1EVkHpdQ8VJQRpQ15hSRoohCztTvDMEepm`). We find that, prior to 2018, these addresses have received 1,232 donations, totaling □94.494 (Figure 10). Using the U.S. dollar value at the time of transaction confirmation, Sci-Hub has received an equivalent of $69,224 in bitcoins. □85.467 has been withdrawn from the Sci-Hub addresses via 174 transactions. Since the price of bitcoins has risen, the combined U.S. dollar value at time of withdrawal was $421,272. At the conclusion of 2017, the Sci-Hub accounts had an outstanding balance of □9.027, valued at roughly $120,000. In response to this study's preprint [54], Sci-Hub tweeted: "the information on donations … is not very accurate, but I cannot correct it: that is confidential." Therefore, presumably, Sci-Hub has received considerable donations via alternative payment systems or to unrevealed Bitcoin addresses, which our audit did not capture. Since we do not know the identity of the depositors, another possibility would be that Sci-Hub transfered bitcoins from other addresses it controlled to the identified donation addresses.

The largest, most prominent academic publishers are thoroughly covered by Sci-Hub, and these publishers have taken note. Elsevier (whose 13.5 million works are 96.9% covered by Sci-Hub) and the American Chemical Society (whose 1.4 million works are 98.8% covered) both filed suit against Sci-Hub, despite the limited enforcement options of United States courts. The widespread gratis access that Sci-Hub provides to previously paywalled articles calls into question the sustainability of the subscription publishing model [55,73]. Avoiding biblioleaks and retaining exclusive possession of digital media may prove an insurmountable challenge for publishers [74]. As distributed and censorship-resistant file storage protocols mature [75,76], successors to Sci-Hub may emerge that no longer rely on a centralized service. Indeed, Alexandra Elbakyan is only one individual in the larger "guerilla access" movement [77–79], which will persist regardless of Sci-Hub's fate. As such, Sci-Hub's corpus of gratis scholarly literature may be extremely difficult to suppress.

Surveys from 2016 suggest awareness and usage of Sci-Hub was not yet commonplace [47,80]. However, adoption appears to be growing. According to Elbakyan, the number of Sci-Hub downloads increased from 42 million in 2015 to 75 million in 2016, equating to a 79% gain [48]. Comparing the search interest peaks following Ⓓand Ⓛin Figure 1, which both correspond to domain outages and hence existing users searching how to access Sci-Hub, we estimate annual growth of 88%. As per Figure 1—figure supplement 1, Sci-Hub averaged 185,243 downloads per day in January–February 2016, whereas in 2017 daily downloads averaged 458,589.

Accordingly, the ratio of Sci-Hub to Penn Libraries downloads in 2017 was 20:1. In addition, adoption of Sci-Hub or similar sites could accelerate due to new technical burdens on authorized access (the flip side of anti-piracy measures) [81,82], crackdowns on article sharing via academic social networks [83,84], or large-scale subscription cancellations by libraries [85].

Historically, libraries have often canceled individual journal subscriptions or switched from bundled to à-la-carte selections [12,86,87]. More recently, library consortia have threatened wholesale cancellation of specific publishers. In 2010, Research Libraries of the UK threatened to let Elsevier contracts expire [14,88], while the University of California raised the possibility of boycotting Nature Publishing Group. But these disputes were ultimately resolved before major cancellations transpired. But in 2017, researchers began losing access to entire publishers. Universities in the Netherlands canceled all Oxford University Press subscriptions in May 2017 [89]. University of Montreal reduced its subscriptions to Taylor & Francis periodicals by 93%, axing 2,231 journals [90]. Negotiations with Elsevier reached impasses in Germany, Peru, and Taiwan. As a result, hundreds of universities have cancelled all Elsevier subscriptions [91,92]. These developments echo the predictions of Elsevier's attorneys in 2015 [93]: "Defendants' actions also threaten imminent irreparable harm to Elsevier because it appears that the Library Genesis Project repository may be approaching (or will eventually approach) a level of 'completeness' where it can serve as a functionally equivalent, although patently illegal, replacement for ScienceDirect."

In the worst case for toll access publishers, growing Sci-Hub usage will become both the cause and the effect of dwindling subscriptions. Librarians rely on usage metrics and user feedback to evaluate subscriptions [12]. Sci-Hub could decrease the use of library subscriptions as many users find it more convenient than authorized access [47]. Furthermore, librarians may receive fewer complaints after canceling subscriptions, as users become more aware of alternatives. Green open access also provides an access route outside of institutional subscription. The posting of preprints and postprints has been growing rapidly [1,94], with new search tools to help locate them [95]. The trend of increasing green availability is poised to continue as funders mandate postprints [96] and preprints help researchers sidestep the slow pace of scholarly publishing [97]. In essence, scholarly publishers may have already lost the access battle. Publishers will be forced to adapt quickly to open access publishing models. In the words of Alexandra Elbakyan [98]: "The effect of long-term operation of Sci-Hub will be that publishers change their publishing models to support Open Access, because closed access will make no sense anymore."

Sci-Hub is poised to fundamentally disrupt scholarly publishing. The transition to gratis availability of scholarly articles is currently underway, and such a model may be inevitable in the long term [99–101]. However, we urge the community to take this opportunity to fully liberate scholarly articles, as well as explore more constructive business models for publishing [102–104]. Only libre access, enabled by open licensing, allows building applications on top of scholarly literature without fear of legal consequences [24]. For example, fulltext mining of scholarly literature is an area of great potential [105], but is currently impractical due to the lack of a large-scale preprocessed corpus of articles. The barriers here are legal, not technological [106,107]. In closing, were all articles libre, there would be no such thing as a "pirate website" for accessing scholarly literature.

Methods

This project was performed entirely in the open, via the GitHub repository greenelab/scihub. Several authors of this study became involved after we mentioned their usernames in GitHub discussions. This project's fully transparent and online model enabled us to assemble an international team of individuals with complementary expertise and knowledge.

We managed our computational environment using Conda, allowing us to specify and install dependencies for both Python and R. We performed our analyses using a series of Jupyter notebooks. In general, data integration and manipulation were performed in Python 3, relying heavily on Pandas, while plotting was performed with ggplot2 in R. Tabular data were saved in TSV (tab-separated values) format, and large datasets were compressed using XZ. We used Git Large File Storage (Git LFS) to track large files, enabling us to make nearly all of the datasets generated and consumed by the analyses available to the public. The Sci-Hub Stats Browser is a single-page application built using React and hosted via GitHub Pages. Frontend visualizations use Vega-Lite [108]. Certain datasets for the browser are hosted in the greenelab/scihub-browser-data repository.

The manuscript source for this study is located at greenelab/scihub-manuscript. We used the Manubot to automatically generate the manuscript from Markdown files. This system — originally developed for the Deep Review to enable collaborative writing on GitHub [109] — uses continuous analysis to fetch reference metadata and rebuild the manuscript upon changes [110].

Digital Object Identifiers

We used DOIs (Digital Object Identifiers) to uniquely identify articles. The Sci-Hub and LibGen scimag repositories also uniquely identify articles by their DOIs, making DOIs the natural primary identifier for our analyses. The DOI initiative began in 1997, and the first DOIs were registered in 2000 [111,112]. Note that DOIs can be registered retroactively. For example, Antony van Leewenhoeck's discovery of protists and bacteria — published in 1677 by *Philosophical Transactions of the Royal Society of London* [113] — has a DOI (`10.1098/rstl.1677.0003`), retroactively assigned in 2006.

Not all scholarly articles have DOIs. By evaluating the presence of DOIs in other databases of scholarly literature (such as PubMed, Web of Science, and Scopus), researchers estimate around 90% of newly published articles in the sciences have DOIs [114,115]. The prevalence of DOIs varies by discipline and country of publication, with DOI assignment in newly published Arts & Humanities articles around 60% [114]. Indeed, DOI registration is almost entirely lacking for publishers from many Eastern European countries [115]. In addition, the prevalence of DOI assignment is likely lower for older articles [115]. The incomplete and non-random assignment of DOIs to scholarly articles is a limitation of this study. However, DOIs are presumably the least imperfect and most widespread identifier for scholarly articles.

An often overlooked aspect of the DOI system is that DOIs are case-insensitive within the ASCII character range [111,116]. In other words, `10.7717/peerj.705` refers to the same article as `10.7717/PeerJ.705`. Accordingly, DOIs make a poor standard identifier unless they are consistently cased. While the DOI handbook states that "all DOI names are converted to upper case upon registration” [111], we lowercased DOIs in accordance with Crossref’s behavior. Given the risk of unmatched DOIs, we lowercased DOIs for each input resource at the earliest opportunity in our processing pipeline. Consistent casing considerably influenced our findings as different resources used different casings of the same DOI.

Crossref-derived catalog of scholarly articles

To catalog all scholarly articles, we relied on the Crossref database. Crossref is a DOI Registration Agency (an entity capable of assigning DOIs) for scholarly publishing [117]. There are presently 10 Registration Agencies. We estimate that Crossref has registered 67% of all DOIs in existence. While several Registration Agencies assign DOIs to scholarly publications, Crossref is the preeminent registrar. In March 2015, of the 1,464,818 valid DOI links on the English version of Wikipedia, 99.9% were registered with Crossref [118]. This percentage was slightly lower for other languages: 99.8% on Chinese Wikipedia and 98.0% on Japanese Wikipedia. Hence, the overwhelming majority of DOI-referenced scholarly articles are registered with Crossref. Since Crossref has the most comprehensive and featureful programmatic access, there was a strong incentive to focus solely on Crossref-registered DOIs. Given Crossref’s preeminence, the omission of other Registration Agencies is unlikely to substantially influence our findings.

We queried the `works` endpoint of the Crossref API to retrieve the metadata for all DOIs, storing the responses in a MongoDB database. The queries began on March 21, 2017 and took 12 days to complete. In total, we retrieved metadata for 87,542,370 DOIs, corresponding to all Crossref works as of March 21, 2017. The source code for this step is available on GitHub at greenelab/crossref. Due to its large file size (7.4 GB), the MongoDB database export of DOI metadata is not available on GitHub, and is instead hosted via figshare [119]. We created TSV files with the minimal information needed for this study: First, a DOI table with columns for work type and date issued. Date issued refers to the earliest known publication date, i.e. the date of print or online publication, whichever occurred first. Second, a mapping of DOI to ISSN for associating articles with their journal of publication.

We selected a subset of Crossref work types to include in our Sci-Hub coverage analyses that corresponded to scholarly articles (i.e. publications). Since we could not locate definitions for the Crossref types, we used our best judgment and evaluated sample works of a given type in the case of uncertainty. We included the following types: `book-chapter`, `book-part`, `book-section`, `journal-article`, `proceedings-article`, `reference-entry`, `report`, and `standard`. Types such as `book`, `journal`, `journal-issue`, and `report-series` were excluded, as they are generally containers for individual articles rather than scholarly articles themselves. After filtering by type, 81,609,016 DOIs remained (77,201,782 of which had their year of publication available). For the purposes of this study, these DOIs represent the entirety of the scholarly literature.

Scopus-derived catalog of journals

Prior to June 2017, the Crossref API had an issue that prevented exhaustively downloading journal metadata. Therefore, we instead relied on the Scopus database to catalog scholarly journals. Scopus uses “title” to refer to all of the following: peer-reviewed journals, trade journals, book series, and conference proceedings. For this study, we refer to all of these types as journals. From the October 2017 data release of Scopus titles, we extracted metadata for 72,502 titles including their names, ISSNs, subject areas, publishers, open access status, and active status. The publisher information was poorly standardized — e.g. both “ICE Publishing” and “ICE Publishing Ltd.” were present — so name variants were combined using OpenRefine. This version of Scopus determined open access status by whether a journal was registered in DOAJ or ROAD as of April 2017. Note that Scopus does not index every scholarly journal [120], which is one reason why 30.5% of articles (24,853,345 DOIs) were not attributable to a journal.

We tidied the Scopus Journal Metrics, which evaluate journals based on the number of citations their articles receive. Specifically, we extracted a 2015 CiteScore for 22,256 titles, 17,336 of which were included in our journal catalog. Finally, we queried the Elsevier API to retrieve homepage URLs for 20,992 Scopus titles. See dhimmel/scopus for the source code and data relating to Scopus.

LibGen scimag’s catalog of articles

Library Genesis (LibGen) is a shadow library primarily comprising illicit copies of academic books and articles. Compared to Sci-Hub, the operations of LibGen are more opaque, as the contributors maintain a low profile and do not contact journalists [31]. LibGen hosts several collections, including distinct repositories for scientific books and textbooks, fiction books, and comics [34]. In 2012, LibGen added the “scimag” database for scholarly literature. Since the spring of 2013, Sci-Hub has uploaded articles that it obtains to LibGen scimag [31]. At the end of 2014, Sci-Hub forked LibGen scimag and began managing its own distinct article repository.

We downloaded the LibGen scimag metadata database on April 7, 2017 as a SQL dump. We imported the SQL dump into MySQL, and then exported the scimag table to a TSV file [121]. Each row of this table corresponds to an article in LibGen, identified by its DOI. The `TimeAdded` field apparently indicates when the publication was uploaded to LibGen. After removing records missing `TimeAdded`, 64,195,940 DOIs remained. 56,205,763 (87.6%) of the DOIs were in our Crossref-derived catalog of scholarly literature. The 12.4% of LibGen scimag DOIs missing from our Crossref catalog likely comprise incorrect DOIs, DOIs whose metadata availability postdates our Crossref export, DOIs from other Registration Agencies, and DOIs for excluded publication types.

Next, we explored the cumulative size of LibGen scimag over time according to the `TimeAdded` field (Figure 11). However, when we compared our plot to one generated from the LibGen scimag database SQL dump on January 1, 2014 [34,35], we noticed a major discrepancy. The earlier analysis identified a total of 22,829,088 DOIs, whereas we found only 233,707 DOIs as of January 1, 2014. We hypothesize that the discrepancy arose because `TimeAdded` indicates the date modified rather than created. Specifically, when an article in the database is changed, the database record for that DOI is entirely replaced. Hence, the `TimeAdded` value is effectively overwritten upon every update to a record. Unfortunately, many research questions require the date first added. For example, lag-time analyses (the time from study publication to LibGen upload) may be unreliable. Therefore, we do not report on these findings in this manuscript. Instead, we provide Figure 11—figure supplement 1 as an example analysis that would be highly informative were reliable creation dates available. In addition, findings from some previous studies may require additional scrutiny. For example, Cabanac writes [34]: “The growth of LibGen suggests that it has benefited from a few isolated, but massive, additions of scientific articles to its cache. For instance, 71% of the article collection was uploaded in 13 days at a rate of 100,000+ articles a day. It is likely that such massive collections of articles result from biblioleaks [74], but one can only speculate about this because of the undocumented source of each file cached at LibGen." While we agree this is most likely the case, confirmation is needed that the bulk addition of articles does not simply correspond to bulk updates rather than bulk initial uploads.

Sci-Hub's catalog of articles

On March 19, 2017, Sci-Hub tweeted: "If you like the list of all DOI collected on Sci-Hub, here it is: sci-hub.cc/downloads/doi.7z … 62,835,101 DOI in alphabetical order". The tweet included a download link for a file with the 62,835,101 DOIs that Sci-Hub claims to provide access to. Of these DOIs, 56,246,220 were part of the Crossref-derived catalog of scholarly articles, and 99.5% of the DOIs from Sci-Hub's list were in the LibGen scimag repository (after filtering). Hence, the LibGen scimag and Sci-Hub repositories have largely stayed in sync since their split. On Twitter, the Sci-Hub account confirmed this finding, commenting "with a small differences, yes the database is the same". Therefore, the LibGen scimag and Sci-Hub DOI catalogs can essentially be used interchangeably for research purposes.

State of OA Datasets

oaDOI, short for open access DOI, is a service that determines whether a DOI is available gratis somewhere online [122]. oaDOI does not index articles posted to academic social networks or available from illicit repositories such as Sci-Hub [1]. Using the oaDOI infrastructure, the State of OA study investigated the availability of articles from three collections [1]. Each collection consists of a random sample of approximately 100,000 articles from a larger corpus. We describe the collections below and report the number of articles after intersection with our DOI catalog:

- **Web of Science**: 103,491 articles published between 2009–2015 and classified as citable items in Web of Science.
- **Unpaywall**: 87,322 articles visited by Unpaywall users from June 5–11, 2017.
- **Crossref**: 99,952 articles with Crossref type of `journal-article`.

Unpaywall is a web-browser extension that notifies its user if an article is available via oaDOI [123]. Since the Unpaywall collection is based on articles that users visited, it's a better reflection of the actual access needs of contemporary scholars. Unfortunately, since the number of visits per article is not preserved by this dataset, fulfillment rate estimates are biased against highly-visited articles and become scale-variant (affected by the popularity of Unpaywall).

The State of OA study ascertained the accessibility status of each DOI in each collection using oaDOI [1,124]. Articles for which oaDOI did not identify a full-text were considered "closed". Otherwise, articles were assigned a color/status of bronze, green, hybrid, or gold. oaDOI classifies articles not available from their publisher's site as either green or closed. The version of oaDOI used in the State of OA study identified green articles by searching PubMed Central and BASE. Readers should note that this implementation likely undercounts green articles, especially if considering articles available from academic social networks as green.

Recent citation catalog

OpenCitations is an public domain resource containing scholarly citation data [125]. OpenCitations extracts its information from the Open Access Subset of PubMed Central. In the greenelab/opencitations repository, we processed the July 25, 2017 OpenCitations data release [126,127], creating a DOI–cites–DOI catalog of bibliographic references. For quality control, we removed DOIs that were not part of the Crossref-derived catalog of articles. Furthermore, we removed outgoing citations from articles published before 2015. Incoming citations to articles predating 2015 were not removed. The resulting catalog consisted of 7,312,607 citations from 200,206 recent articles to 3,857,822 referenced articles.

Sci-Hub access logs

The 2016 study titled "Who's downloading pirated papers? Everyone" analyzed a dataset of Sci-Hub access logs [44,45]. Alexandra Elbakyan worked with journalist John Bohannon to produce a dataset of Sci-Hub's resolved requests from September 1, 2015 through February 29, 2016 [33]. In November 2015, Sci-Hub's domain name was suspended as the result of legal action by Elsevier [26,41]. According to Bohannon, this resulted in "an 18-day gap in the data starting November 4, 2015 when the domain sci-hub.org went down and the server logs were improperly configured." We show this downtime in Figure 1.

We filtered the access events by excluding DOIs not included in our literature catalog and omitting records that occurred before an article's publication date. This filter preserved 26,984,851 access events for 10,293,836 distinct DOIs (97.5% of the 10,552,418 distinct prefiltered DOIs). We summarized the access events for each article using the following metrics:

1. downloads: total number of times the article was accessed
2. visitors: number of IP addresses that accessed the article
3. countries: number of countries (geolocation by IP address) from which the article was accessed
4. days: number of days on which the article was accessed
5. months: number of months in which the article was accessed

Next, we calculated journal-level access metrics based on articles published from January 1, 2014 until the start of the Sci-Hub access log records on September 1, 2015. For each journal, we calculated the average values for the five access log metrics described above. Interestingly, the journal *Medicine - Programa de Formación Médica Continuada Acreditado* received the most visitors per article, averaging 33.4 visitors for each of its 326 articles.

Note that these analyses do not include Sci-Hub's access logs for 2017 [129], which were released on January 18, 2018. Unfortunately, at that time we had already adopted a freeze on major new analyses. Nonetheless, we did a quick analysis to assess growth in Sci-Hub downloads over time that combined the 2015–2016 and 2017 access log data (Figure 1—figure supplement 1).

Tables

Table 1: Coverage for the ten journals with the most articles. The total number of articles published by each journal is noted in the Crossref column. The table provides the number (Sci-Hub column) and percentage (Coverage column) of these articles that are in Sci-Hub's repository.

Journal	Sci-Hub	Crossref	Coverage
The Lancet	457,650	458,580	99.8%
Nature	385,619	399,273	96.6%
British Medical Journal (Clinical Research Edition)	17,141	392,277	4.4%
Lecture Notes in Computer Science	103,675	356,323	29.1%
Science	230,649	251,083	91.9%
Journal of the American Medical Association	191,950	248,369	77.3%
Journal of the American Chemical Society	189,142	189,567	99.8%
Scientific American	22,600	186,473	12.1%
New England Journal of Medicine	180,321	180,467	99.9%
PLOS ONE	4,731	177,260	2.7%

Figures

Figure 1: The history of Sci-Hub. Weekly interest from Google Trends is plotted over time for the search terms "Sci-Hub" and "LibGen". The light green period indicates when Sci-Hub used LibGen as its database for storing articles [31]. Light blue indicates the collection period of the Sci-Hub access logs that we analyze throughout this study [33]. Based on these logs and newly released logs for 2017, Figure 1—figure supplement 1 shows the number of articles downloaded from Sci-Hub over time, providing an alternative look into Sci-Hub's growth. The first pink dotted line represents the collection date of the LibGen scimag metadata used in Cabanac's study [34,35]. The second pink dotted line shows the date of Sci-Hub's tweeted DOI catalog used in this study.

Figure 2: Coverage by article type. Coverage is plotted for the Crossref work types included by this study. We refer to all of these types as "articles".

Figure 3: Distributions of journal & publisher coverages. The histograms show the distribution of Sci-Hub's coverage for all 23,037 journals (top) and 3,832 publishers (bottom). Each bin spans 2.5 percentage points. For example, the top-left bar indicates Sci-Hub's coverage is between 0.0%–2.5% for 3,892 journals.

Figure 4: Coverage by journal attributes. Each bar represents Sci-Hub's coverage of articles in journals with the specified attributes, according to Scopus. Active refers to whether a journal still publishes articles. Open refers to whether a journal is open access. Subject area refers to a journal's discipline. Note that some journals are assigned to multiple subject areas. As an example, we identified 588 neuroscience journals, which contained 1.8 million articles. Sci-Hub possessed 87.7% of these articles.

Figure 5: Coverage by publisher. Article coverage is shown for all Scopus publishers with at least 200,000 articles.

Figure 6: Coverage of articles by year published. Sci-Hub's article coverage is shown for each year since 1850.

Figure 7: Sci-Hub's coverage by oaDOI access status. Using oaDOI calls from the State of OA study, we show Sci-Hub's coverage on each access status. Gray indicates articles that were not accessible via oaDOI (referred to as closed). Here, all three State of OA collections were combined, yielding 290,120 articles. Figure 7—figure supplement 1 shows coverage separately for the three State of OA collections.

Figure 8: Coverage of several access methods and their combinations. This figure compares datasets of article coverage corresponding to various access methods. These article sets refer to manually evaluated access via the publisher's site from outside of an institutional network (labeled None) or from inside Penn's network (labeled Penn); access according to Penn's library system (labeled PennText); access via the oaDOI utility (labeled oaDOI); and inclusion in Sci-Hub's database (labeled Sci-Hub). Each diagram shows the coverage of three access methods and their possible combinations. Within a diagram, each section notes the percent coverage achieved by the corresponding combination of access methods. **Contrary to traditional Venn diagrams**, each section does not indicate disjoint sets of articles. Instead, each section shows coverage on the same set of articles, whose total number is reported in the diagram's title. The top two diagrams show coverage on a small set of manually evaluated articles (confidence intervals provided in the main text). The bottom two diagrams show coverage on a larger set of automatically evaluated articles. The two lefthand diagrams show coverage on all articles, whereas the two righthand diagrams show coverage on toll access articles only. Specifically, the top-right diagram assesses coverage on articles that were inaccessible from outside of an institutional network. Similarly, the bottom-right diagram assesses coverage of articles that were classified as closed or green by oaDOI, and thus excludes gold, hybrid, and bronze articles (those available gratis from their publisher).

Figure 9: Relation to journal impact. A) Average coverage for journals divided into 2015 CiteScore deciles. The CiteScore range defining each decile is shown by the x-axis labels.

The ticks represent 99% confidence intervals of the mean. This is the only analysis where "Sci-Hub Coverage" refers to journal-level rather than article-level averages. **B)** The association between 2015 CiteScore and average visitors per article is plotted for open and toll access journals. Curves show the 95% confidence band from a Generalized Additive Model.

Figure 10: Number of bitcoin donations per month. The number of bitcoin donations to Sci-Hub is shown for each month from June 2015 to December 2017. Since February 2016, Sci-Hub has received over 25 donations per month. Each donation corresponds to an incoming transaction to a known Sci-Hub address. See Figure 10—figure supplement 1 for the amount donated each month, valued in BTC or USD.

Figure 11: Number of articles in LibGen scimag over time. The figure shows the number of articles in LibGen scimag, according to its `TimeAdded` field, for two database dumps. The number of articles added per day for the January 1, 2014 LibGen database dump was provided by Cabanac and corresponds to Figure 1 of [34]. Notice the major discrepancy whereby articles from the April 7, 2017 database dump were added at later dates. Accordingly, we hypothesize that the `TimeAdded` field is replaced upon modification, making it impossible to assess date of first upload.

Figure Supplements

Figure 1—figure supplement 1: Downloads per day on Sci-Hub for months with access logs. The number of articles downloaded from Sci-Hub is shown over time. Sci-Hub access logs were combined from two releases: [33] covering 27,819,963 downloads from September 2015 to February 2016 and [129] covering 150,875,862 downloads from 2017. The plot shows the average number of downloads per day for months with data. There were 54 days within the collection periods without any logged access events, due presumably to service outages or server misconfiguration. Hence, we ignored days without logs when computing monthly averages. Point color indicates the proportion of days with logs for a given month. For example, November 2015 and October 2017, which were missing logs for 17 and 23 days respectively, are thus lighter. The December 2017 dropoff in downloads likely reflects the effect of domain suspensions that occurred in late November [62]. Unlike the Sci-Hub log analyses elsewhere in this study, this plot does not filter for valid articles (i.e. DOIs in our Crossref-derived catalog of scholarly literature).

Figure 4—figure supplement 1: Coverage by country of publication. Scopus assigns each journal a country of publication. Sci-Hub's coverage is shown for countries with at least 100,000 articles.

Figure 6—figure supplement 1: Coverage of articles by year published and journal access status. Sci-Hub's coverage is shown separately for articles in open versus toll access journals, for each year since 1950.

Figure 7—figure supplement 1: Coverage by oaDOI access status on each State of OA collection. Coverage by oaDOI access status is shown for Sci-Hub, PennText, and the union of Sci-Hub and PennText. Each panel refers to a different State of OA collection, with Combined referring to the union of the Crossref, Unpaywall, and Web of Science collections. The Sci-Hub section of the Combined panel is the same as Figure 7. Impressively, Sci-Hub's coverage of the closed articles in the Web of Science collection was 97.8%. This remarkable coverage likely reflects that these articles were published from 2009–2015 and classified as citable items by Web of Science, which is selective when indexing journals [120]. Note that PennText does not have complete coverage of bronze, hybrid, and gold articles, which should be the case were all metadata systems perfect. These anomalies likely result from errors in both PennText (whose accuracy we estimated at 88.2%) and oaDOI (whose accuracy the State of OA study estimated at 90.4%, i.e. Table 1 of [1] reports 5 false positives and 43 false negatives on oaDOI calls for 500 articles).

Figure 10—figure supplement 1: Bitcoin donations to Sci-Hub per month. For months since June 2015, total bitcoin donations (deposits to known Sci-Hub addresses) were assessed. Donations in USD refers to the United States dollar value at time of transaction confirmation.

Figure 11—figure supplement 1: Lag-time from publication to LibGen upload. For each year of publication from 2010–2017, we plot the relationship between lag-time and LibGen scimag's coverage. For example, this plot shows that 75% of articles published in 2011 were uploaded to LibGen within 60 months. This analysis only considers articles for which a month of publication can reliably be extracted, which excludes all articles allegedly published on January 1. This plot portrays lag-times as decreasing over time, with maximum coverage declining. For example, coverage for 2016 articles exceeded 50% within 6 months, but appears to have reached an asymptote around 60%. Alternatively, coverage for 2014 took 15 months to exceed 50%, but has since reached 75%. However, this signal could result from post-dated LibGen upload timestamps. Therefore, we caution against drawing any conclusions from the `TimeAdded` field in LibGen scimag until its accuracy can be established more reliably.

Acknowledgements

We'd like to thank the individuals, not listed as authors, who provided comments on GitHub issues or pull requests. Specifically, Ross Mounce, Richard Smith-Unna, Guillaume Cabanac, and Stuart Taylor provided valuable input while the study was underway. In addition, we're grateful to GitHub for offering gratis Large File Storage as part of their education program.

References

1. **The State of OA: A large-scale analysis of the prevalence and impact of Open Access articles**
Heather Piwowar, Jason Priem, Vincent Larivière, Juan Pablo Alperin, Lisa Matthias, Bree Norlander, Ashley Farley, Jevin West, Stefanie Haustein
PeerJ (2017) https://doi.org/10.7287/peerj.preprints.3119v1

2. **The Number of Scholarly Documents on the Public Web**
Madian Khabsa, C. Lee Giles
PLoS ONE (2014) https://doi.org/10.1371/journal.pone.0093949

3. **Open access levels: a quantitative exploration using Web of Science and oaDOI data**
Jeroen Bosman, Bianca Kramer
PeerJ (2018) https://doi.org/10.7287/peerj.preprints.3520v1

4. **The academic, economic and societal impacts of Open Access: an evidence-based review**
Jonathan P. Tennant, François Waldner, Damien C. Jacques, Paola Masuzzo, Lauren B. Collister, Chris. H. J. Hartgerink
F1000Research (2016) https://doi.org/10.12688/f1000research.8460.3

5. **A Brief History of Open Access**
Paul Royster
Library Conference Presentations and Speeches (2016)
http://digitalcommons.unl.edu/library_talks/123

6. **Proportion of Open Access Papers Published in Peer-Reviewed Journals at the European and World Levels—1996–2013**
Éric Archambault, Didier Amyot, Philippe Deschamps, Aurore Nicol, Françoise Provencher, Lise Rebout, Guillaume Roberge
Copyright, Fair Use, Scholarly Communication, etc. (2014)
http://digitalcommons.unl.edu/scholcom/8

7. **Half of 2011 papers now free to read**
Richard Van Noorden
Nature (2013) https://doi.org/10.1038/500386a

8. **Beyond Open: Expanding Access to Scholarly Content**
Alice Meadows
The Journal of Electronic Publishing (2015) https://doi.org/10.3998/3336451.0018.301

9. **Transforming Access to Research Literature for Developing Countries**
Barbara Kirsop, Leslie Chan
Serials Review (2005) https://doi.org/10.1080/00987913.2005.10764998

10. **Sci-Hub and medical practice: an ethical dilemma in Peru**
Guido Bendezú-Quispe, Wendy Nieto-Gutiérrez, Josmel Pacheco-Mendoza, Alvaro Taype-Rondan
The Lancet Global Health (2016) https://doi.org/10.1016/s2214-109x(16)30188-7

11. **Graph 4: Expenditure Trends in ARL Libraries, 1986-2015**
Association of Research Libraries
ARL Statistics 2014–2015 (2017) http://www.arl.org/storage/documents/expenditure-trends.pdf

12. **The Serials Crisis Revisited**
Dana L. Roth
The Serials Librarian (1990) https://doi.org/10.1300/j123v18n01_09

13. **The Oligopoly of Academic Publishers in the Digital Era**
Vincent Larivière, Stefanie Haustein, Philippe Mongeon
PLOS ONE (2015) https://doi.org/10.1371/journal.pone.0127502

14. **Evaluating big deal journal bundles**
Theodore C. Bergstrom, Paul N. Courant, R. Preston McAfee, Michael A. Williams
Proceedings of the National Academy of Sciences (2014) https://doi.org/10.1073/pnas.1403006111

15. **Freedom for scholarship in the internet age**
Heather Grace Morrison
(2012) http://summit.sfu.ca/item/12537

16. **Is the staggeringly profitable business of scientific publishing bad for science?**
Stephen Buranyi, Stephen Buranyi
the Guardian (2017) https://www.theguardian.com/science/2017/jun/27/profitable-business-scientific-publishing-bad-for-science

17. **Open access: The true cost of science publishing**
Richard Van Noorden
Nature (2013) https://doi.org/10.1038/495426a

18. **New World, Same Model | Periodicals Price Survey 2017**
Stephen Bosch, Kittie Henderson
Library Journal (2017) http://lj.libraryjournal.com/2017/04/publishing/new-world-same-model-periodicals-price-survey-2017/

19. **Journal subscription costs - FOIs to UK universities**
Stuart Lawson, Ben Meghreblian, Michelle Brook
Figshare (2015) https://doi.org/10.6084/m9.figshare.1186832.v23

20. **Journal subscription expenditure in the UK 2015-16**
Stuart Lawson
Figshare (2017) https://doi.org/10.6084/m9.figshare.4542433.v6

21. **Five Year Journal Price Increase History (2013-2017)**
EBSCO
(2017) https://www.ebscohost.com/promoMaterials/Five_Year_Journal_Price_Increase_History_EBSCO_2013-2017.pdf

22. **Open Access**
Peter Suber
MIT Press (2017) https://mitpress.mit.edu/books/open-access

23. **Gratis and libre open access**
Peter Suber
Peter Suber, Gratis and libre open access, SPARC Open Access Newsletter, August 2, 2008. (2008) https://dash.harvard.edu/handle/1/4322580

24. **The licensing of bioRxiv preprints**
Daniel Himmelstein
Satoshi Village (2016) http://blog.dhimmel.com/biorxiv-licenses/

25. **The frustrated science student behind Sci-Hub**
John Bohannon
Science (2016) https://doi.org/10.1126/science.aaf5675

26. **Pirate research-paper sites play hide-and-seek with publishers**
Quirin Schiermeier
Nature (2015) https://doi.org/10.1038/nature.2015.18876

27. **Sci-Hub is a goal, changing the system is a method**
Alexandra Elbakyan
Engineuring (2016) https://engineuring.wordpress.com/2016/03/11/sci-hub-is-a-goal-changing-the-system-is-a-method/

28. **Letter – Document #50 of Elsevier Inc. v. Sci-Hub – Case 1:15-cv-04282-RWS**
Alexandra Elbakyan
Southern District Court of New York (2015) https://www.courtlistener.com/docket/4355308/50/elsevier-inc-v-sci-hub/

29. **Alexandra Elbakyan – Science Should be Open to all Not Behind Paywalls**
Elena Milova
Life Extension Advocacy Foundation (2017) http://www.leafscience.org/alexandra-elbakyan/

30. **Tor: The Second-Generation Onion Router**
Roger Dingledine, Nick Mathewson, Paul Syverson
(2004) http://www.dtic.mil/docs/citations/ADA465464

31. **Some facts on Sci-Hub that Wikipedia gets wrong**
Alexandra Elbakyan
Engineuring (2017) https://engineuring.wordpress.com/2017/07/02/some-facts-on-sci-hub-that-wikipedia-gets-wrong/

32. **Scholarly journal publishing in transition- from restricted to open access**
Bo-Christer Björk
Electronic Markets (2017) https://doi.org/10.1007/s12525-017-0249-2

33. **Sci-Hub download data**
Alexandra Elbakyan, John Bohannon
Dryad Digital Repository (2016) https://doi.org/10.5061/dryad.q447c/1

34. **Bibliogifts in LibGen? A study of a text-sharing platform driven by biblioleaks and crowdsourcing**
Guillaume Cabanac
Journal of the Association for Information Science and Technology (2015) https://doi.org/10.1002/asi.23445

35. **Scimag catalogue of LibGen as of January 1st, 2014**
Guillaume Cabanac
Figshare (2017) https://doi.org/10.6084/m9.figshare.4906367.v1

36. **Scholar Subreddit: Libgen down**
BitterCoffeeMan, BigHogBalls
Reddit (2016) https://redd.it/2raea8

37. **Elsevier Cracks Down on Pirated Scientific Articles**
Ernesto Van der Sar
TorrentFreak (2015) https://torrentfreak.com/elsevier-cracks-down-on-pirated-scientific-articles-150609/

38. **Complaint – Document #1 of Elsevier Inc. v. Sci-Hub – Case 1:15-cv-04282-RWS**
Joseph V. DeMarco, David M. Hirschberg, Urvashi Sen
Southern District Court of New York (2015) https://www.courtlistener.com/docket/4355308/1/elsevier-inc-v-sci-hub/

39. **Court Orders Shutdown of Libgen, Bookfi and Sci-Hub**
Ernesto Van der Sar
TorrentFreak (2015) https://torrentfreak.com/court-orders-shutdown-of-libgen-bookfi-and-sci-hub-151102/

40. **Memorandum & Opinion – Document #53 of Elsevier Inc. v. Sci-Hub – Case 1:15-cv-04282-RWS**
Robert W. Sweet
Southern District Court of New York (2015) https://www.courtlistener.com/docket/4355308/53/elsevier-inc-v-sci-hub/

41. **Sci-Hub, BookFi and LibGen Resurface After Being Shut Down**
Ernesto Van der Sar
TorrentFreak (2015) https://torrentfreak.com/sci-hub-and-libgen-resurface-after-being-shut-down-151121/

42. **Meet the Robin Hood of Science**
Simon Oxenham
Big Think (2016) http://bigthink.com/neurobonkers/a-pirate-bay-for-science

43. **Should All Research Papers Be Free?**
Kate Murphy

New York Times (2016) https://www.nytimes.com/2016/03/13/opinion/sunday/should-all-research-papers-be-free.html

44. **Who’s downloading pirated papers? Everyone**
J. Bohannon
Science (2016) https://doi.org/10.1126/science.352.6285.508

45. **Who’s downloading pirated papers? Everyone**
John Bohannon
Science (2016) https://doi.org/10.1126/science.aaf5664

46. **Paper piracy sparks online debate**
Chris Woolston
Nature (2016) https://doi.org/10.1038/nature.2016.19841

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

48. **Nature’s 10***Nature* (2016) https://doi.org/10.1038/540507a

49. **US court grants Elsevier millions in damages from Sci-Hub**
Quirin Schiermeier
Nature (2017) https://doi.org/10.1038/nature.2017.22196

50. **Sci-Hub Ordered to Pay $15 Million in Piracy Damages**
Ernesto Van der Sar
TorrentFreak (2017) https://torrentfreak.com/sci-hub-ordered-to-pay-15-million-in-piracy-damages-170623/

51. **Judgement – Document #87 of Elsevier Inc. v. Sci-Hub – Case 1:15-cv-04282-RWS**
Robert W. Sweet
Southern District Court of New York (2017)
https://www.documentcloud.org/documents/3878258-2017-06-21-Elsevier-Sci-Hub-Final-Judgement.html

52. **New Lawsuit Demands ISP Blockades Against “Pirate” Site Sci-Hub**
Ernesto Van der Sar
TorrentFreak (2017) https://torrentfreak.com/new-lawsuit-demands-isp-blockades-against-pirate-site-sci-hub-170629/

53. **Complaint – Document #1 of American Chemical Society v. Sci-Hub – Case 1:17-cv-00726-LMB-JFA**
Barnes Attison L. III, Weslow David E., Gardner Matthew J.
Eastern District Court of Virginia (2017)
https://www.courtlistener.com/docket/6146630/1/american-chemical-society-v-does-1-99/

54. **Sci-Hub provides access to nearly all scholarly literature**
Daniel S Himmelstein, Ariel R Romero, Stephen R McLaughlin, Bastian Greshake Tzovaras,

Casey S Greene
PeerJ (2017) https://doi.org/10.7287/peerj.preprints.3100v1

55. **Sci-Hub's cache of pirated papers is so big, subscription journals are doomed, data analyst suggests**
Lindsay McKenzie
Science (2017) https://doi.org/10.1126/science.aan7164

56. **Library Babel Fish Blog: Inevitably Open**
Barbara Fister
Inside Higher Ed (2017) https://www.insidehighered.com/blogs/library-babel-fish/inevitably-open

57. **The World's Largest Free Scientific Resource Is Now Blocked in Russia**
Reid Standish
Foreign Policy (2017) http://foreignpolicy.com/2017/09/06/the-worlds-largest-free-scientific-resource-is-now-blocked-in-russia/

58. **Ichneumonidae (Hymenoptera) associated with xyelid sawflies (Hymenoptera, Xyelidae) in Mexico**
Andrey I. Khalaim, Enrique Ruíz-Cancino
Journal of Hymenoptera Research (2017) https://doi.org/10.3897/jhr.58.12919

59. **Вернуть Sci-Hub**
Алла Астахова
Status Praesens (2017) http://alla-astakhova.ru/sci-hub/

60. **Pirate paper website Sci-Hub dealt another blow by US courts**
Quirin Schiermeier
Nature (2017) https://doi.org/10.1038/nature.2017.22971

61. **Order – Document #37 of American Chemical Society v. Sci-Hub – Case 1:17-cv-00726-LMB-JFA**
Brinkema Leonie M.
Eastern District Court of Virginia (2017)
https://regmedia.co.uk/2017/11/07/sci_hub_block_order_short.pdf

62. **Sci-Hub domains inactive following court order**
Andrew Silver
The Register (2017)
http://www.theregister.co.uk/2017/11/23/scihubs_become_inactive_following_court_order/

63. **Nature owner merges with publishing giant**
Richard Van Noorden
Nature (2015) https://doi.org/10.1038/nature.2015.16731

64. **Shadow Libraries and You: Sci-Hub Usage and the Future of ILL**
Gabriel J. Gardner, Stephen R. McLaughlin, Andrew D. Asher
(2017) https://hdl.handle.net/10760/30981

65. **Correlating the Sci-Hub data with World Bank Indicators and Identifying Academic Use**
Bastian Greshake
The Winnower https://doi.org/10.15200/winn.146485.57797

66. **Looking into Pandora's Box: The Content of Sci-Hub and its Usage**
Bastian Greshake
F1000Research (2017) https://doi.org/10.12688/f1000research.11366.1

67. **Data And Scripts For Looking Into Pandora'S Box: The Content Of Sci-Hub And Its Usage**
Bastian Greshake
Zenodo (2017) https://doi.org/10.5281/zenodo.472493

68. **Online Access To ACS Publications Is Restored After Some Customers Were Unintentionally Blocked | Chemical & Engineering News**
Sophie L. Rovner
(2014) http://cen.acs.org/articles/92/web/2014/04/Online-Access-ACS-Publications-Restored.html

69. **Publisher under fire for fake article webpages**
Rachel Becker
Nature (2016) https://doi.org/10.1038/535011f

70. **"Predatory" open access: a longitudinal study of article volumes and market characteristics**
Cenyu Shen, Bo-Christer Björk
BMC Medicine (2015) https://doi.org/10.1186/s12916-015-0469-2

71. **Logit-Based Interval Estimation for Binomial Data Using the Jeffreys Prior**
Donald B. Rubin, Nathaniel Schenker
Sociological Methodology (1987) https://doi.org/10.2307/271031

72. **Declaration in Support of Motion – Document #8 Attachment #23 of Elsevier Inc. v. Sci-Hub – Case 1:15-cv-04282-RWS**
Anthony Woltermann
Southern District Court of New York (2015)
https://www.courtlistener.com/docket/4355308/8/23/elsevier-inc-v-sci-hub/

73. **Access, ethics and piracy**
Stuart Lawson
Insights the UKSG journal (2017) https://doi.org/10.1629/uksg.333

74. **Is Biblioleaks Inevitable?**
Adam G Dunn, Enrico Coiera, Kenneth D Mandl
Journal of Medical Internet Research (2014) https://doi.org/10.2196/jmir.3331

75. **IPFS - Content Addressed, Versioned, P2P File System**
Juan Benet
arXiv (2014) https://arxiv.org/abs/1407.3561v1

76. **Decentralized Storage: The Backbone of the Third Web**
ConsenSys
ConsenSys Media (2016) https://media.consensys.net/decentralized-storage-the-backbone-of-the-third-web-d4bc54e79700

77. **Pirates in the Library – An Inquiry into the Guerilla Open Access Movement**
Balázs Bodó
8th Annual Workshop of the International Society for the History and Theory of Intellectual Property, CREATe (2016) https://doi.org/10.2139/ssrn.2816925

78. **Libraries in the Post-Scarcity Era**
Balázs Bodó
Copyrighting Creativity: Creative values, Cultural Heritage Institutions and Systems of Intellectual Property (2015) https://doi.org/10.2139/ssrn.2616636

79. **The Rise of Pirate Libraries**
Sarah Laskow
Atlas Obscura (2016) http://www.atlasobscura.com/articles/the-rise-of-illegal-pirate-libraries

80. **Use, knowledge, and perception of the scientific contribution of Sci-Hub in medical students: Study in six countries in Latin America**
Christian R. Mejia, Mario J. Valladares-Garrido, Armando Miñan-Tapia, Felipe T. Serrano, Liz E. Tobler-Gómez, William Pereda-Castro, Cynthia R. Mendoza-Flores, Maria Y. Mundaca-Manay, Danai Valladares-Garrido
PLOS ONE (2017) https://doi.org/10.1371/journal.pone.0185673

81. **Two-step Authentication: Finally Coming to a University Near You**
Phil Davis
The Scholarly Kitchen (2016) https://scholarlykitchen.sspnet.org/2016/06/21/two-step-authentication-finally-coming-to-a-university-near-you/

82. **Sci-Hub and the Four Horsemen of the Internet**
Joseph J. Esposito
The Scholarly Kitchen (2016) https://scholarlykitchen.sspnet.org/2016/03/02/sci-hub-and-the-four-horsemen-of-the-internet/

83. **Publishers go after networking site for illicit sharing of journal papers**
Dalmeet Chawla
Science (2017) https://doi.org/10.1126/science.aaq0132

84. **Publishers take ResearchGate to court, alleging massive copyright infringement**
Dalmeet Chawla
Science (2017) https://doi.org/10.1126/science.aaq1560

85. **Sci-Hub Moves to the Center of the Ecosystem**
Joseph J. Esposito
The Scholarly Kitchen (2017) https://scholarlykitchen.sspnet.org/2017/09/05/sci-hub-moves-center-ecosystem/

86. **Factors in Science Journal Cancellation Projects: The Roles of Faculty Consultations and Data**
Peter Fernandez Jeanine Williamson
Issues in Science and Technology Librarianship (2014) https://doi.org/10.5062/f4g73bp3

87. **Walking away from the American Chemical Society**
Jenica Rogers
Attempting Elegance (2012) http://www.attemptingelegance.com/?p=1765

88. **Reassessing the value proposition: first steps towards a fair(er) price for scholarly journals**
David C Prosser
Serials: The Journal for the Serials Community (2011) https://doi.org/10.1629/2460

89. **Dutch lose access to OUP journals in subscription standoff**
Holly Else
Times Higher Education (2017) https://www.timeshighereducation.com/news/dutch-lose-access-oup-journals-subscription-standoff

90. **UdeM Libraries cancel Big Deal subscription to 2,231 periodical titles published by Taylor & Francis Group**
Stéphanie Gagnon
Communiqués, Bibliothèques, Université de Montréal (2017) http://www.bib.umontreal.ca/communiques/20170504-DC-annulation-taylor-francis-va.htm

91. **Scientists in Germany, Peru and Taiwan to lose access to Elsevier journals**
Quirin Schiermeier, Emiliano Rodríguez Mega
Nature (2016) https://doi.org/10.1038/nature.2016.21223

92. **Germany vs Elsevier: universities win temporary journal access after refusing to pay fees**
Quirin Schiermeier
Nature (2018) https://doi.org/10.1038/d41586-018-00093-7

93. **Memorandum of Law in Support of Motion – Document #6 of Elsevier Inc. v. Sci-Hub – Case 1:15-cv-04282-RWS**
Joseph V. DeMarco, David M. Hirschberg, Urvashi Sen
Southern District Court of New York (2015) https://www.courtlistener.com/docket/4355308/6/elsevier-inc-v-sci-hub/

94. **Are preprints the future of biology? A survival guide for scientists**
Jocelyn Kaiser
Science (2017) https://doi.org/10.1126/science.aaq0747

95. **Need a paper? Get a plug-in**
Dalmeet Singh Chawla
Nature (2017) https://doi.org/10.1038/d41586-017-05922-9

96. **Funders punish open-access dodgers**
Richard Van Noorden
Nature (2014) https://doi.org/10.1038/508161a

97. **Does it take too long to publish research?**
Kendall Powell
Nature (2016) https://doi.org/10.1038/530148a

98. **Why Sci-Hub is the true solution for Open Access: reply to criticism**
Alexandra Elbakyan
Engineuring (2016) https://engineuring.wordpress.com/2016/02/24/why-sci-hub-is-the-true-solution-for-open-access-reply-to-criticism/

99. **The Inevitability of Open Access**
David W. Lewis
College & Research Libraries (2012) https://doi.org/10.5860/crl-299

100. **Is free inevitable in scholarly communication?: The economics of open access**
Caroline Sutton
College & Research Libraries News (2011) https://doi.org/10.5860/crln.72.11.8671

101. **Open access to research is inevitable, says Nature editor-in-chief**
Alok Jha
the Guardian (2012) https://www.theguardian.com/science/2012/jun/08/open-access-research-inevitable-nature-editor

102. **The Transition to Open Access: The State of the Market, Offsetting Deals, and a Demonstrated Model for Fair Open Access with the Open Library of Humanities**
Eve Martin Paul, de Vries Saskia C.J., Rooryck Johan
Stand Alone (2017) https://doi.org/10.3233/978-1-61499-769-6-118

103. **A bold open-access push in Germany could change the future of academic publishing**
Gretchen Vogel
Science (2017) https://doi.org/10.1126/science.aap7562

104. **We can shift academic culture through publishing choices**
Corina J Logan
F1000Research (2017) https://doi.org/10.12688/f1000research.11415.2

105. **Text mining of 15 million full-text scientific articles**
David Westergaard, Hans-Henrik Stærfeldt, Christian Tønsberg, Lars Juhl Jensen, Søren Brunak
Cold Spring Harbor Laboratory (2017) https://doi.org/10.1101/162099

106. **The Social, Political and Legal Aspects of Text and Data Mining (TDM)**
Michelle Brook, Peter Murray-Rust, Charles Oppenheim
D-Lib Magazine (2014) https://doi.org/10.1045/november14-brook

107. **Trouble at the text mine**
Richard Van Noorden
Nature (2012) https://doi.org/10.1038/483134a

108. **Vega-Lite: A Grammar of Interactive Graphics**
Arvind Satyanarayan, Dominik Moritz, Kanit Wongsuphasawat, Jeffrey Heer
IEEE Transactions on Visualization and Computer Graphics (2017)
https://doi.org/10.1109/tvcg.2016.2599030

109. **Opportunities And Obstacles For Deep Learning In Biology And Medicine**
Travers Ching, Daniel S. Himmelstein, Brett K. Beaulieu-Jones, Alexandr A. Kalinin, Brian T. Do, Gregory P. Way, Enrico Ferrero, Paul-Michael Agapow, Wei Xie, Gail L. Rosen, … Casey S. Greene
Cold Spring Harbor Laboratory (2017) https://doi.org/10.1101/142760

110. **Reproducibility of computational workflows is automated using continuous analysis**
Brett K Beaulieu-Jones, Casey S Greene
Nature Biotechnology (2017) https://doi.org/10.1038/nbt.3780

111. **DOI® Handbook***International DOI Foundation* (2017) https://doi.org/10.1000/182

112. **Digital Object Identifiers and Their Use in Libraries**
Jue Wang
Serials Review (2007) https://doi.org/10.1016/j.serrev.2007.05.006

113. **Observations, Communicated to the Publisher by Mr. Antony van Leewenhoeck, in a Dutch Letter of the 9th of Octob. 1676. Here English'd: concerning Little Animals by Him Observed in Rain-Well-Sea. and Snow Water; as Also in Water Wherein Pepper Had Lain Infused**
A. van Leewenhoeck
Philosophical Transactions of the Royal Society of London (1677)
https://doi.org/10.1098/rstl.1677.0003

114. **Availability of digital object identifiers (DOIs) in Web of Science and Scopus**
Juan Gorraiz, David Melero-Fuentes, Christian Gumpenberger, Juan-Carlos Valderrama-Zurián
Journal of Informetrics (2016) https://doi.org/10.1016/j.joi.2015.11.008

115. **Availability of digital object identifiers in publications archived by PubMed**
Christophe Boudry, Ghislaine Chartron
Scientometrics (2017) https://doi.org/10.1007/s11192-016-2225-6

116. **BS ISO 26324:2012. Information and documentation. Digital object identifier system***BSI British Standards* (2012) https://doi.org/10.3403/30177056

117. **CrossRef developments and initiatives: an update on services for the scholarly publishing community from CrossRef**
Rachael Lammey
Science Editing (2014) https://doi.org/10.6087/kcse.2014.1.13

118. **DOI Links on Wikipedia**
Jiro Kikkawa, Masao Takaku, Fuyuki Yoshikane
Digital Libraries: Knowledge, Information, and Data in an Open Access Society (2016)
https://doi.org/10.1007/978-3-319-49304-6_40

119. **Metadata for all DOIs in Crossref: JSON MongoDB exports of all works from the Crossref API**
Daniel Himmelstein, Kurt Wheeler, Casey Greene
Figshare (2017) https://doi.org/10.6084/m9.figshare.4816720.v1

120. **The journal coverage of Web of Science and Scopus: a comparative analysis**
Philippe Mongeon, Adèle Paul-Hus
Scientometrics (2015) https://doi.org/10.1007/s11192-015-1765-5

121. **A user-friendly extract of the LibGen scimag metadata SQL dump on 2017-04-07**
Daniel Himmelstein, Stephen McLaughlin
Figshare (2017) https://doi.org/10.6084/m9.figshare.5231245.v1

122. **Introducing oaDOI: resolve a DOI straight to OA**
Heather Piwowar
Impactstory blog (2016) http://blog.impactstory.org/introducting-oadoi/

123. **Unpaywall finds free versions of paywalled papers**
Dalmeet Singh Chawla
Nature (2017) https://doi.org/10.1038/nature.2017.21765

124. **Data From: The State Of Oa: A Large-Scale Analysis Of The Prevalence And Impact Of Open Access Articles**
Heather Piwowar, Jason Priem, Vincent Larivière, Juan Pablo Alperin, Lisa Matthias, Bree Norlander, Ashley Farley, Jevin West, Stefanie Haustein
Zenodo (2017) https://doi.org/10.5281/zenodo.837902

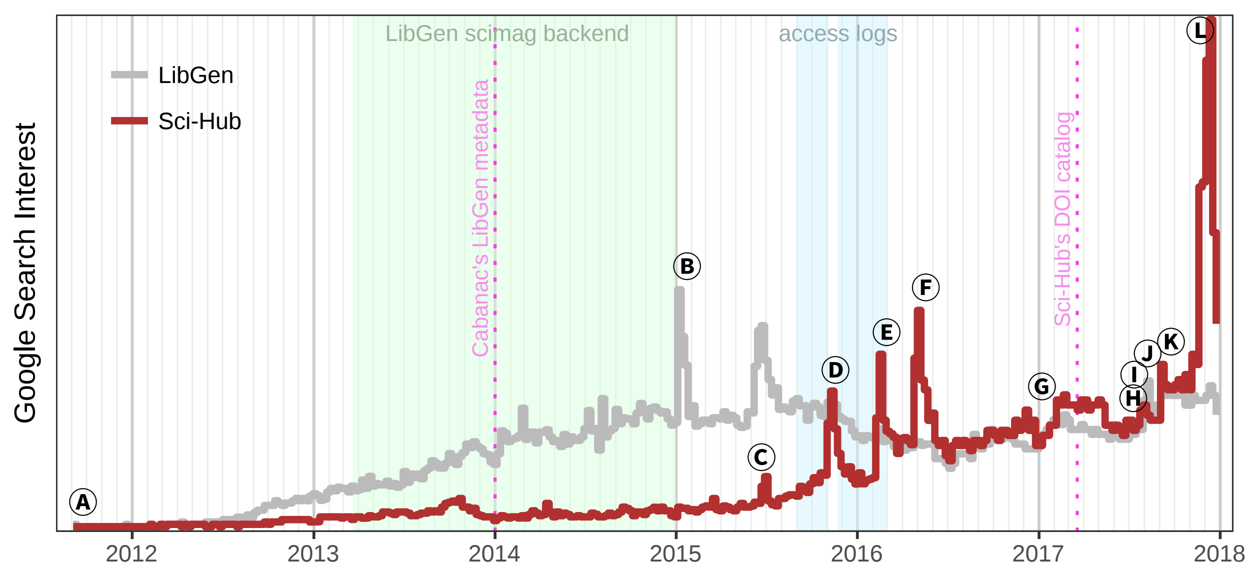
125. **Setting our bibliographic references free: towards open citation data**
Silvio Peroni, Alexander Dutton, Tanya Gray, David Shotton
Journal of Documentation (2015) https://doi.org/10.1108/jd-12-2013-0166

126. **Metadata for the OpenCitations Corpus**
Silvio Peroni, David Shotton
Figshare (2016) https://doi.org/10.6084/m9.figshare.3443876.v3

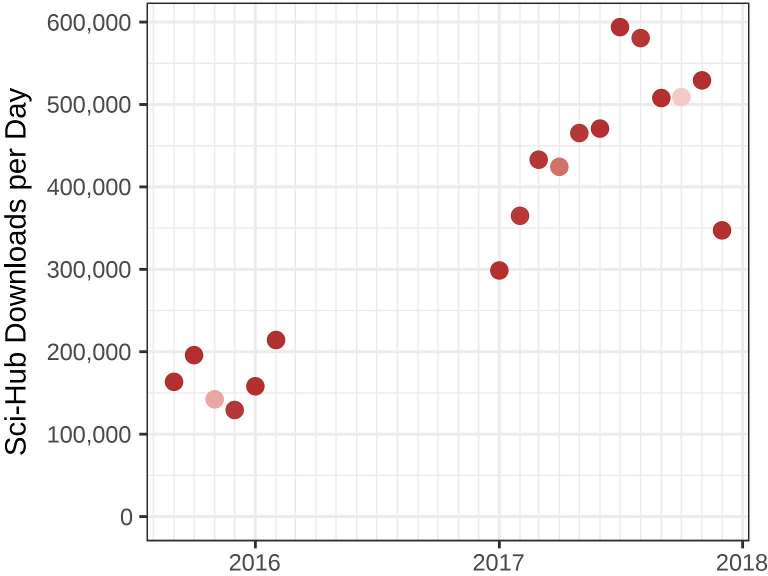
127. **OCC dataset of all the identifiers, made on 2017-07-25**
OpenCitations
Figshare (2017) https://doi.org/10.6084/m9.figshare.5255368.v1

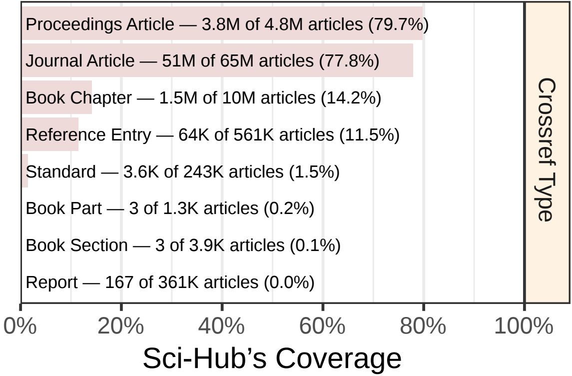
128. **OCC dataset of all the bibliographic resources, made on 2017-07-25**
OpenCitations
Figshare (2017) https://doi.org/10.6084/m9.figshare.5255365.v1

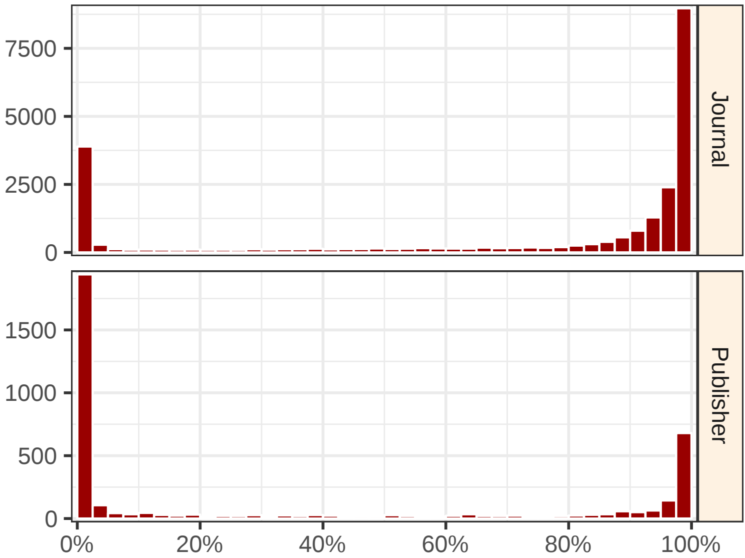
129. **Sci-Hub Download Log Of 2017**
Bastian Greshake Tzovaras
Zenodo (2018) https://doi.org/10.5281/zenodo.1158301



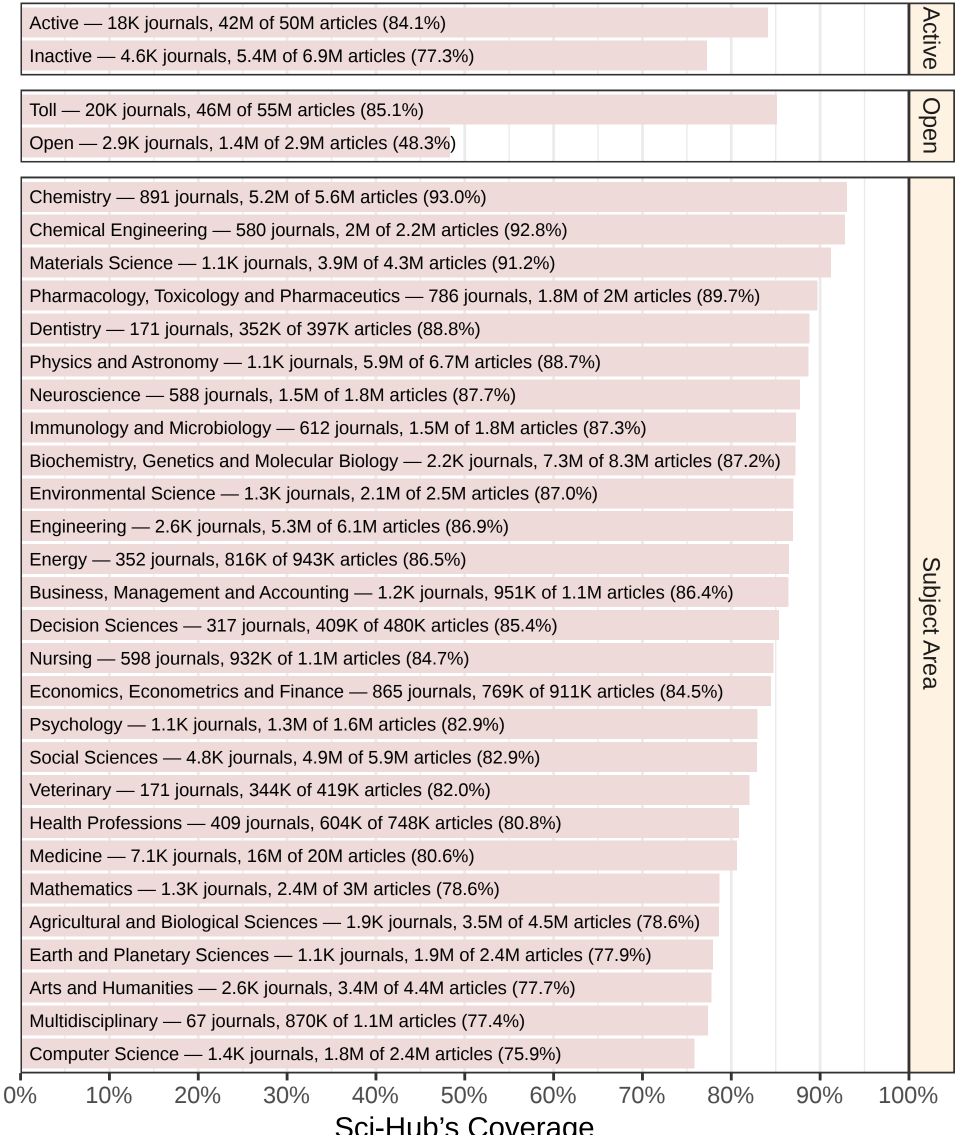
LibGen scimag backend
access logs
LibGen
Sci-Hub
Cabanac's LibGen metadata
Sci-Hub's DOI catalog
Google Search Interest
A
B
C
D
E
F
G
H
I
J
K
L
2012
2013
2014
2015
2016
2017
2018

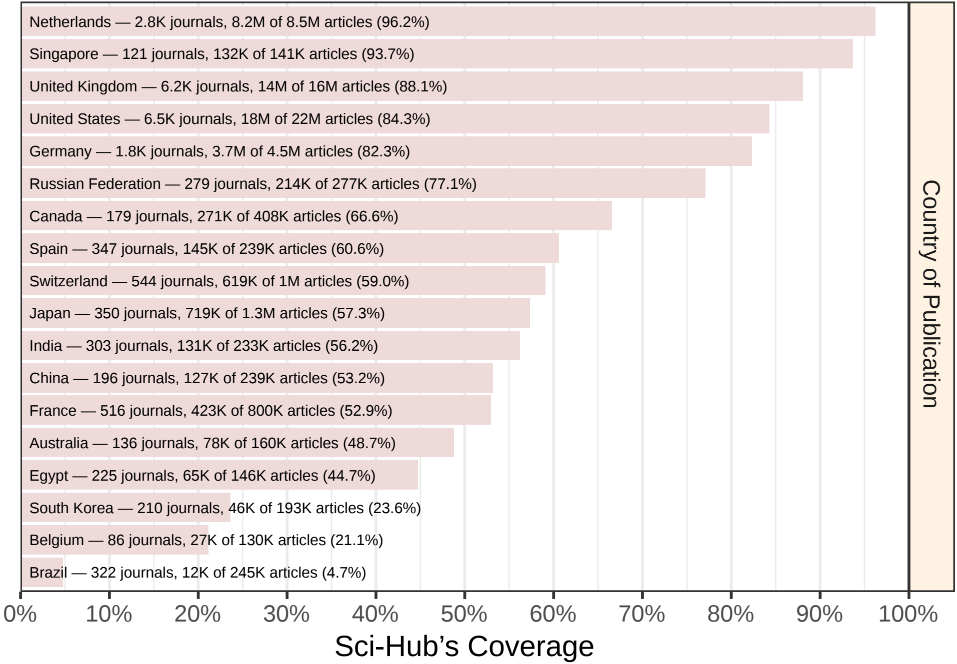

Sci-Hub Downloads per Day
600,000
500,000
400,000
300,000
200,000
100,000
0
2016
2017
2018


Proceedings Article — 3.8M of 4.8M articles (79.7%)
Journal Article — 51M of 65M articles (77.8%)
Book Chapter — 1.5M of 10M articles (14.2%)
Reference Entry — 64K of 561K articles (11.5%)
Standard — 3.6K of 243K articles (1.5%)
Book Part — 3 of 1.3K articles (0.2%)
Book Section — 3 of 3.9K articles (0.1%)
Report — 167 of 361K articles (0.0%)
Crossref Type
0%
20%
40%
60%
80%
100%
Sci-Hub's Coverage

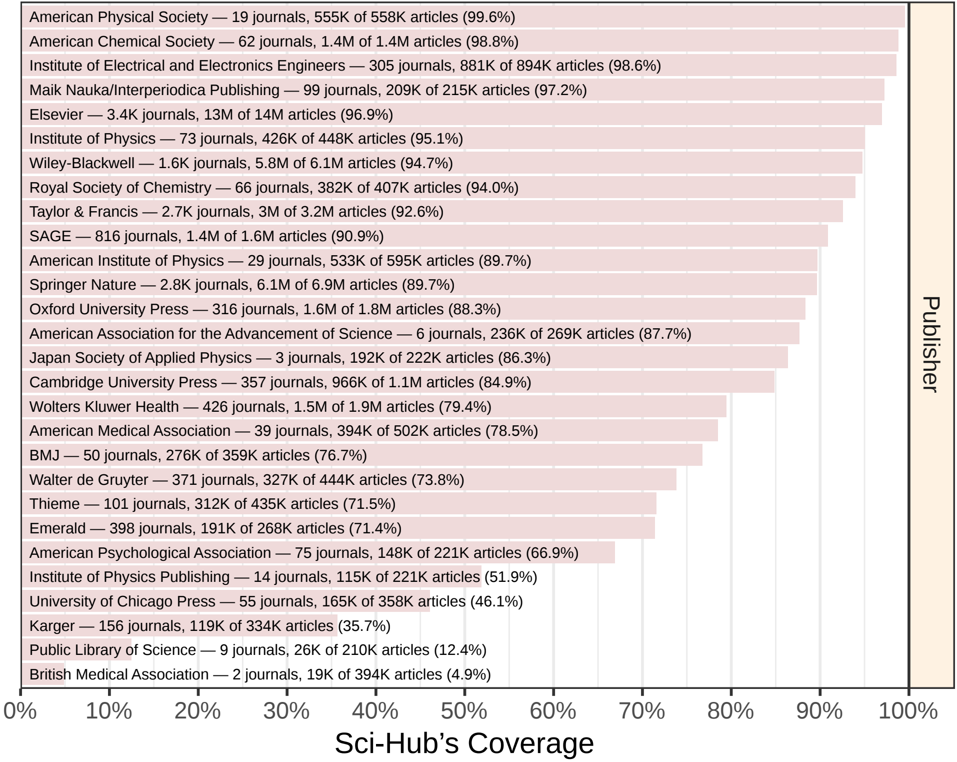


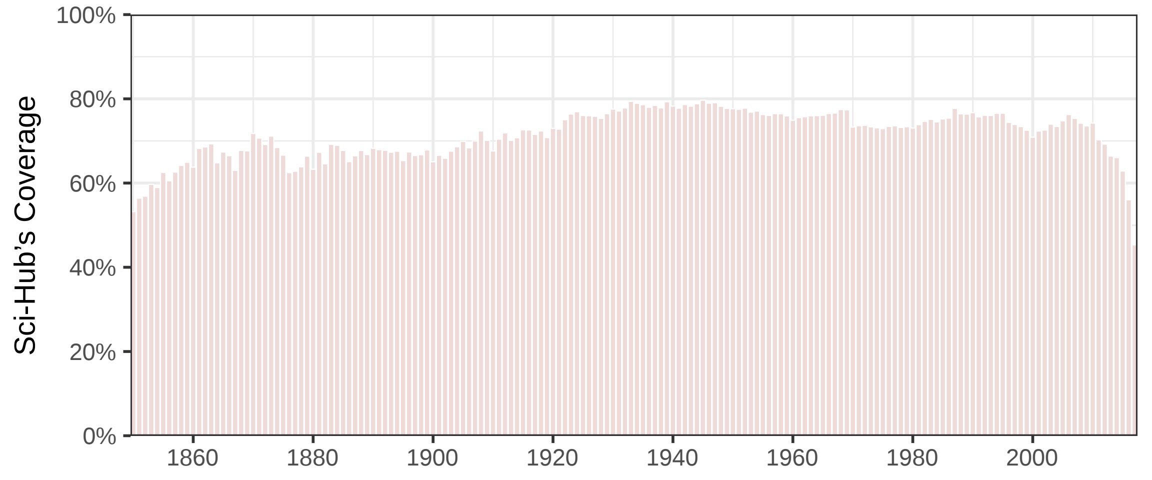
7500
5000
2500
0
Journal
1500
1000
500
0
Publisher
0%
20%
40%
60%
80%
100%


Active — 18K journals, 42M of 50M articles (84.1%)
Inactive — 4.6K journals, 5.4M of 6.9M articles (77.3%)
Active
Toll — 20K journals, 46M of 55M articles (85.1%)
Open — 2.9K journals, 1.4M of 2.9M articles (48.3%)
Open
Chemistry — 891 journals, 5.2M of 5.6M articles (93.0%)
Chemical Engineering — 580 journals, 2M of 2.2M articles (92.8%)
Materials Science — 1.1K journals, 3.9M of 4.3M articles (91.2%)
Pharmacology, Toxicology and Pharmaceutics — 786 journals, 1.8M of 2M articles (89.7%)
Dentistry — 171 journals, 352K of 397K articles (88.8%)
Physics and Astronomy — 1.1K journals, 5.9M of 6.7M articles (88.7%)
Neuroscience — 588 journals, 1.5M of 1.8M articles (87.7%)
Immunology and Microbiology — 612 journals, 1.5M of 1.8M articles (87.3%)
Biochemistry, Genetics and Molecular Biology — 2.2K journals, 7.3M of 8.3M articles (87.2%)
Environmental Science — 1.3K journals, 2.1M of 2.5M articles (87.0%)
Engineering — 2.6K journals, 5.3M of 6.1M articles (86.9%)
Energy — 352 journals, 816K of 943K articles (86.5%)
Business, Management and Accounting — 1.2K journals, 951K of 1.1M articles (86.4%)
Decision Sciences — 317 journals, 409K of 480K articles (85.4%)
Nursing — 598 journals, 932K of 1.1M articles (84.7%)
Economics, Econometrics and Finance — 865 journals, 769K of 911K articles (84.5%)
Psychology — 1.1K journals, 1.3M of 1.6M articles (82.9%)
Social Sciences — 4.8K journals, 4.9M of 5.9M articles (82.9%)
Veterinary — 171 journals, 344K of 419K articles (82.0%)
Health Professions — 409 journals, 604K of 748K articles (80.8%)
Medicine — 7.1K journals, 16M of 20M articles (80.6%)
Mathematics — 1.3K journals, 2.4M of 3M articles (78.6%)
Agricultural and Biological Sciences — 1.9K journals, 3.5M of 4.5M articles (78.6%)
Earth and Planetary Sciences — 1.1K journals, 1.9M of 2.4M articles (77.9%)
Arts and Humanities — 2.6K journals, 3.4M of 4.4M articles (77.7%)
Multidisciplinary — 67 journals, 870K of 1.1M articles (77.4%)
Computer Science — 1.4K journals, 1.8M of 2.4M articles (75.9%)
Subject Area
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
Sci-Hub's Coverage

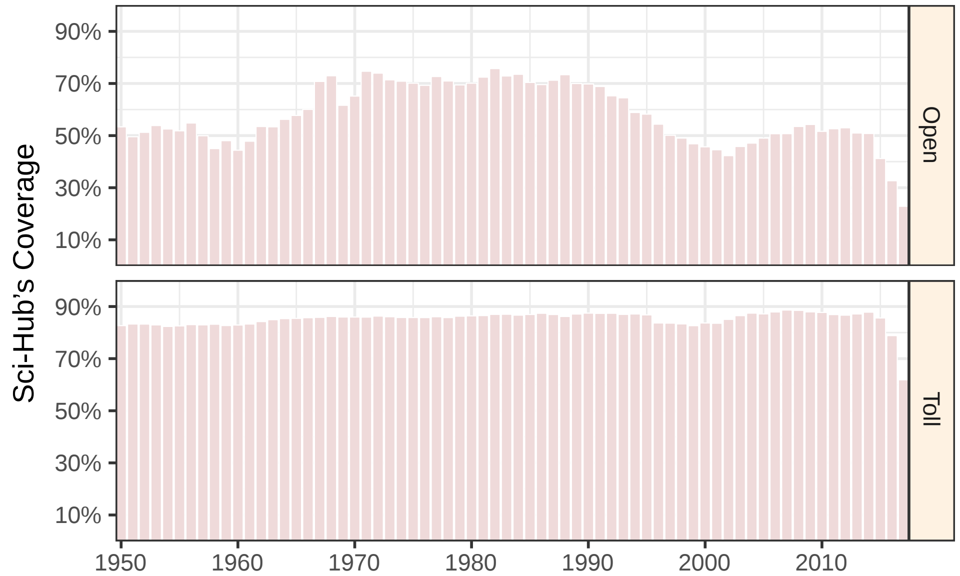


Netherlands — 2.8K journals, 8.2M of 8.5M articles (96.2%)
Singapore — 121 journals, 132K of 141K articles (93.7%)
United Kingdom — 6.2K journals, 14M of 16M articles (88.1%)
United States — 6.5K journals, 18M of 22M articles (84.3%)
Germany — 1.8K journals, 3.7M of 4.5M articles (82.3%)
Russian Federation — 279 journals, 214K of 277K articles (77.1%)
Canada — 179 journals, 271K of 408K articles (66.6%)
Spain — 347 journals, 145K of 239K articles (60.6%)
Switzerland — 544 journals, 619K of 1M articles (59.0%)
Japan — 350 journals, 719K of 1.3M articles (57.3%)
India — 303 journals, 131K of 233K articles (56.2%)
China — 196 journals, 127K of 239K articles (53.2%)
France — 516 journals, 423K of 800K articles (52.9%)
Australia — 136 journals, 78K of 160K articles (48.7%)
Egypt — 225 journals, 65K of 146K articles (44.7%)
South Korea — 210 journals, 46K of 193K articles (23.6%)
Belgium — 86 journals, 27K of 130K articles (21.1%)
Brazil — 322 journals, 12K of 245K articles (4.7%)
Country of Publication
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
Sci-Hub’s Coverage

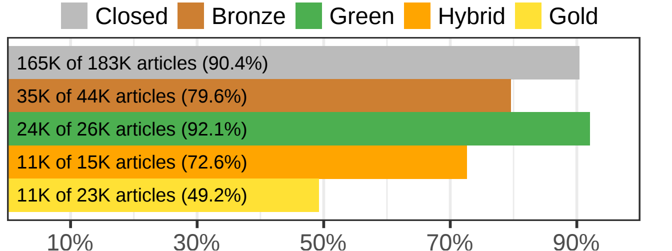

American Physical Society — 19 journals, 555K of 558K articles (99.6%)
American Chemical Society — 62 journals, 1.4M of 1.4M articles (98.8%)
Institute of Electrical and Electronics Engineers — 305 journals, 881K of 894K articles (98.6%)
Maik Nauka/Interperiodica Publishing — 99 journals, 209K of 215K articles (97.2%)
Elsevier — 3.4K journals, 13M of 14M articles (96.9%)
Institute of Physics — 73 journals, 426K of 448K articles (95.1%)
Wiley-Blackwell — 1.6K journals, 5.8M of 6.1M articles (94.7%)
Royal Society of Chemistry — 66 journals, 382K of 407K articles (94.0%)
Taylor & Francis — 2.7K journals, 3M of 3.2M articles (92.6%)
SAGE — 816 journals, 1.4M of 1.6M articles (90.9%)
American Institute of Physics — 29 journals, 533K of 595K articles (89.7%)
Springer Nature — 2.8K journals, 6.1M of 6.9M articles (89.7%)
Oxford University Press — 316 journals, 1.6M of 1.8M articles (88.3%)
American Association for the Advancement of Science — 6 journals, 236K of 269K articles (87.7%)
Japan Society of Applied Physics — 3 journals, 192K of 222K articles (86.3%)
Cambridge University Press — 357 journals, 966K of 1.1M articles (84.9%)
Wolters Kluwer Health — 426 journals, 1.5M of 1.9M articles (79.4%)
American Medical Association — 39 journals, 394K of 502K articles (78.5%)
BMJ — 50 journals, 276K of 359K articles (76.7%)
Walter de Gruyter — 371 journals, 327K of 444K articles (73.8%)
Thieme — 101 journals, 312K of 435K articles (71.5%)
Emerald — 398 journals, 191K of 268K articles (71.4%)
American Psychological Association — 75 journals, 148K of 221K articles (66.9%)
Institute of Physics Publishing — 14 journals, 115K of 221K articles (51.9%)
University of Chicago Press — 55 journals, 165K of 358K articles (46.1%)
Karger — 156 journals, 119K of 334K articles (35.7%)
Public Library of Science — 9 journals, 26K of 210K articles (12.4%)
British Medical Association — 2 journals, 19K of 394K articles (4.9%)
Publisher
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
Sci-Hub’s Coverage

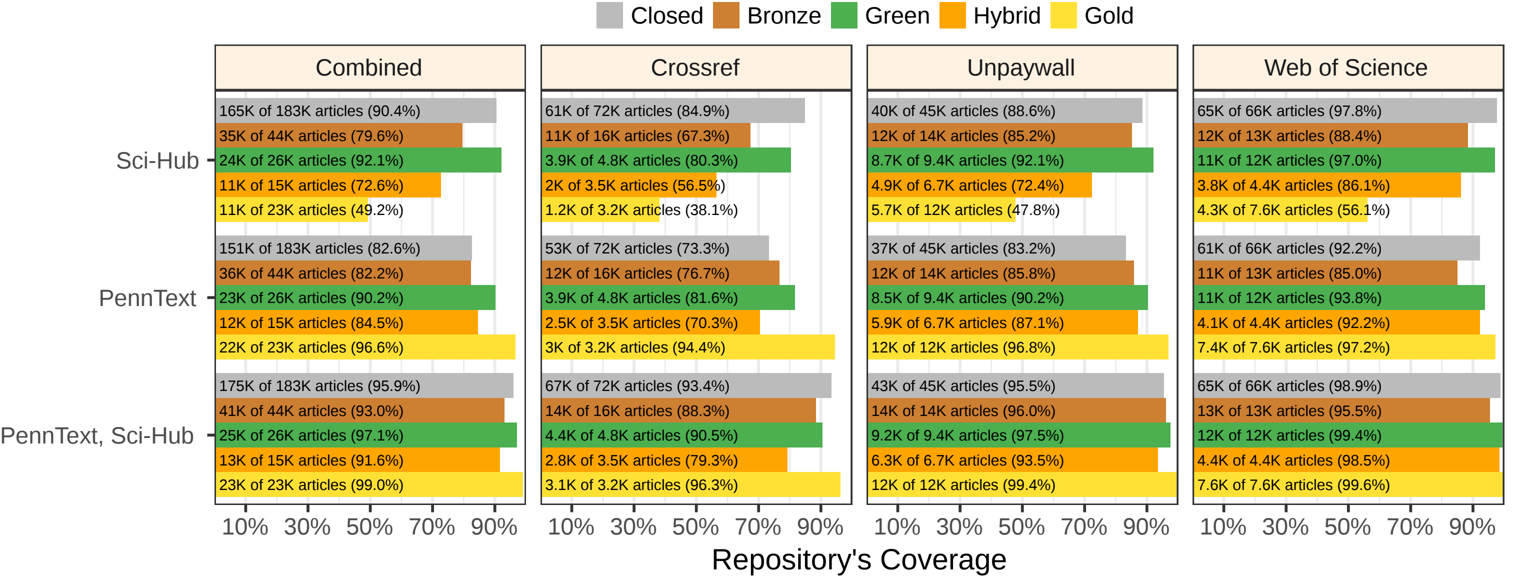


Sci-Hub's Coverage
100%
80%
60%
40%
20%
0%
1860
1880
1900
1920
1940
1960
1980
2000

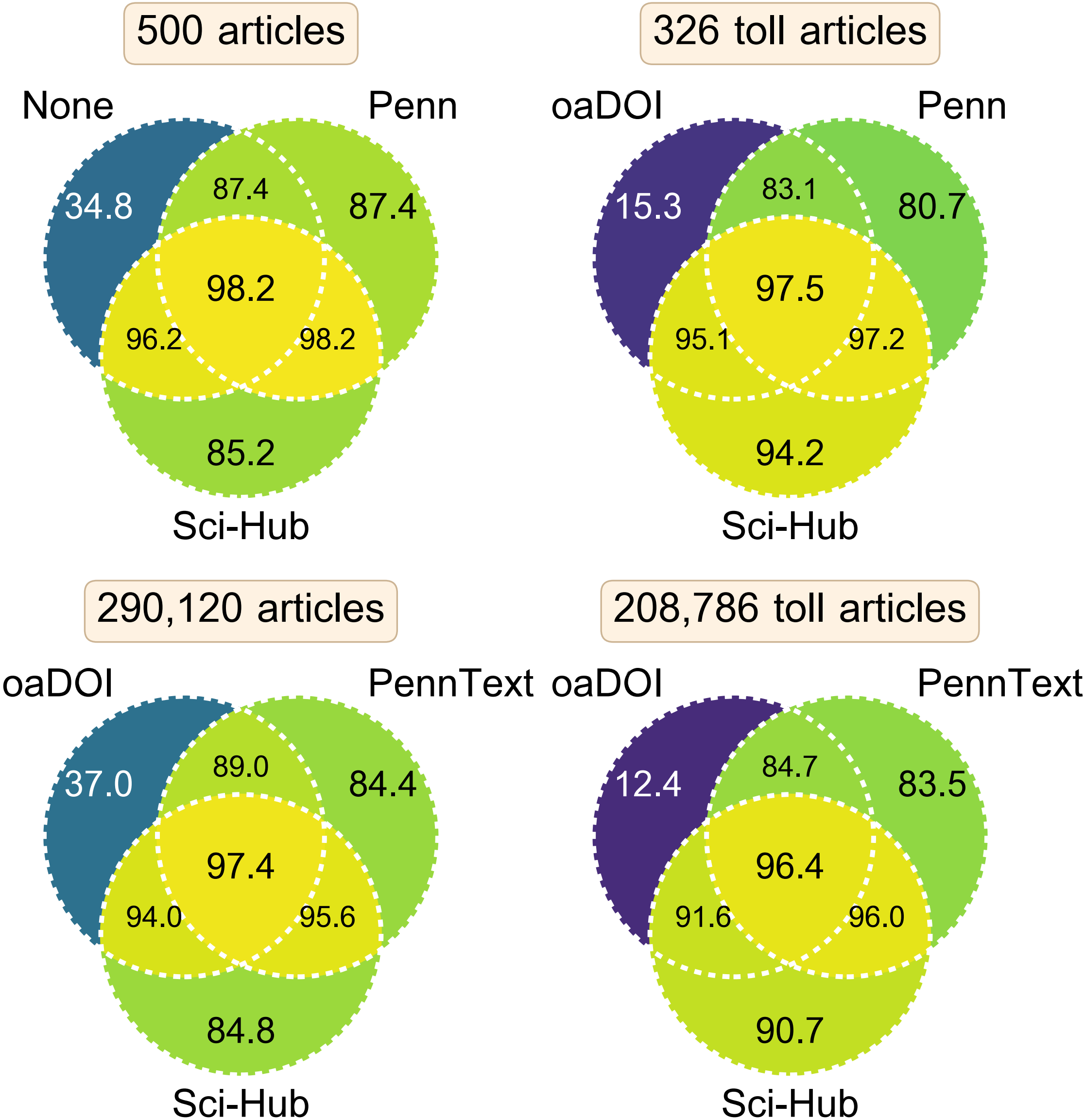


Sci-Hub's Coverage
90%
70%
50%
30%
10%
Open
Toll
1950
1960
1970
1980
1990
2000
2010

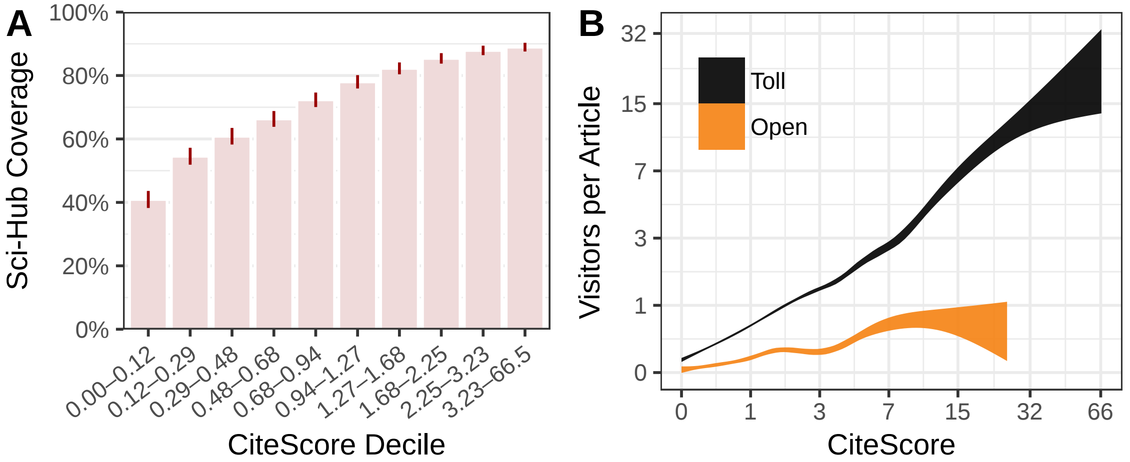

Closed
Bronze
Green
Hybrid
Gold
165K of 183K articles (90.4%)
35K of 44K articles (79.6%)
24K of 26K articles (92.1%)
11K of 15K articles (72.6%)
11K of 23K articles (49.2%)
10%
30%
50%
70%
90%



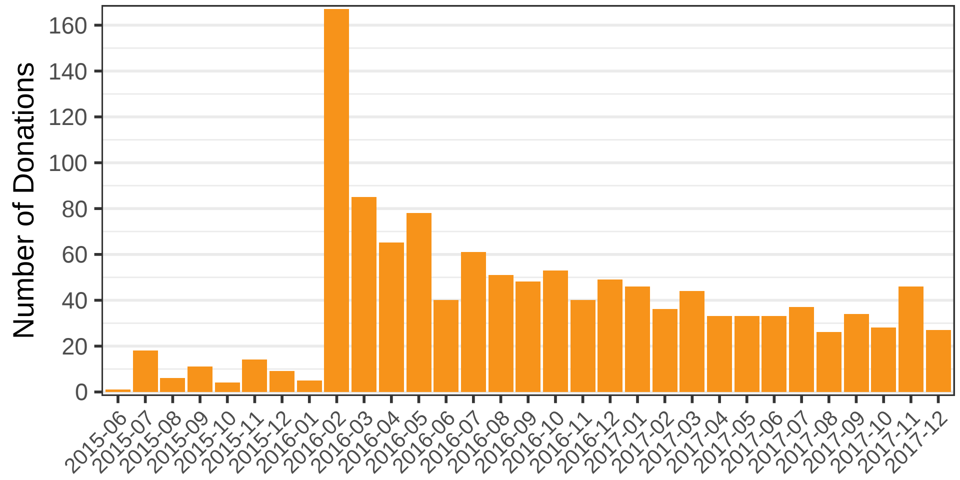
Closed
Bronze
Green
Hybrid
Gold
Combined
Crossref
Unpaywall
Web of Science
Sci-Hub
PennText
PennText, Sci-Hub
165K of 183K articles (90.4%)
35K of 44K articles (79.6%)
24K of 26K articles (92.1%)
11K of 15K articles (72.6%)
11K of 23K articles (49.2%)
151K of 183K articles (82.6%)
36K of 44K articles (82.2%)
23K of 26K articles (90.2%)
12K of 15K articles (84.5%)
22K of 23K articles (96.6%)
175K of 183K articles (95.9%)
41K of 44K articles (93.0%)
25K of 26K articles (97.1%)
13K of 15K articles (91.6%)
23K of 23K articles (99.0%)
61K of 72K articles (84.9%)
11K of 16K articles (67.3%)
3.9K of 4.8K articles (80.3%)
2K of 3.5K articles (56.5%)
1.2K of 3.2K articles (38.1%)
53K of 72K articles (73.3%)
12K of 16K articles (76.7%)
3.9K of 4.8K articles (81.6%)
2.5K of 3.5K articles (70.3%)
3K of 3.2K articles (94.4%)
67K of 72K articles (93.4%)
14K of 16K articles (88.3%)
4.4K of 4.8K articles (90.5%)
2.8K of 3.5K articles (79.3%)
3.1K of 3.2K articles (96.3%)
40K of 45K articles (88.6%)
12K of 14K articles (85.2%)
8.7K of 9.4K articles (92.1%)
4.9K of 6.7K articles (72.4%)
5.7K of 12K articles (47.8%)
37K of 45K articles (83.2%)
12K of 14K articles (85.8%)
8.5K of 9.4K articles (90.2%)
5.9K of 6.7K articles (87.1%)
12K of 12K articles (96.8%)
43K of 45K articles (95.5%)
14K of 14K articles (96.0%)
9.2K of 9.4K articles (97.5%)
6.3K of 6.7K articles (93.5%)
12K of 12K articles (99.4%)
65K of 66K articles (97.8%)
12K of 13K articles (88.4%)
11K of 12K articles (97.0%)
3.8K of 4.4K articles (86.1%)
4.3K of 7.6K articles (56.1%)
61K of 66K articles (92.2%)
11K of 13K articles (85.0%)
11K of 12K articles (93.8%)
4.1K of 4.4K articles (92.2%)
7.4K of 7.6K articles (97.2%)
65K of 66K articles (98.9%)
13K of 13K articles (95.5%)
12K of 12K articles (99.4%)
4.4K of 4.4K articles (98.5%)
7.6K of 7.6K articles (99.6%)
10%
30%
50%
70%
90%
Repository's Coverage



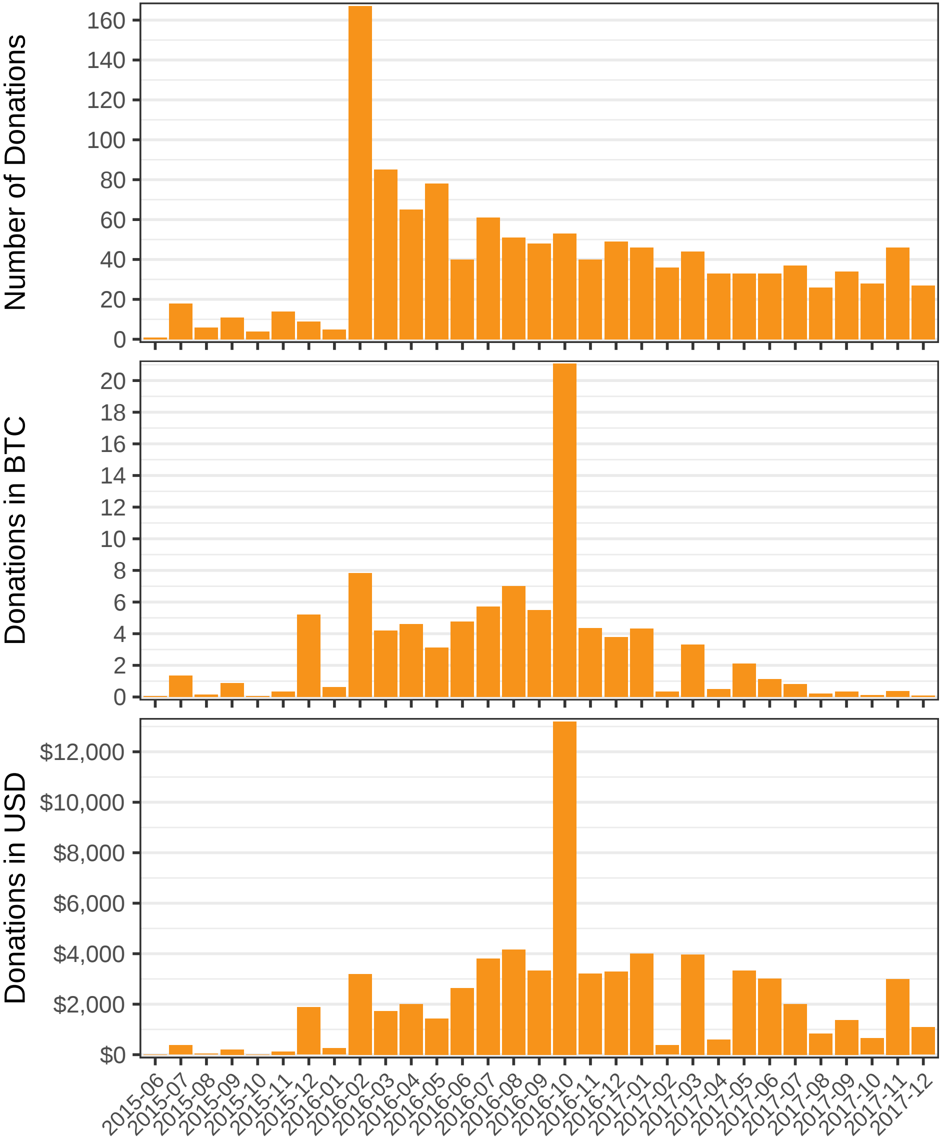
500 articles
None
Penn
34.8
87.4
87.4
98.2
96.2
98.2
85.2
Sci-Hub
326 toll articles
oaDOI
Penn
15.3
83.1
80.7
97.5
95.1
97.2
94.2
Sci-Hub
290,120 articles
oaDOI
PennText
37.0
89.0
84.4
97.4
94.0
95.6
84.8
Sci-Hub
208,786 toll articles
oaDOI
PennText
12.4
84.7
83.5
96.4
91.6
96.0
90.7
Sci-Hub

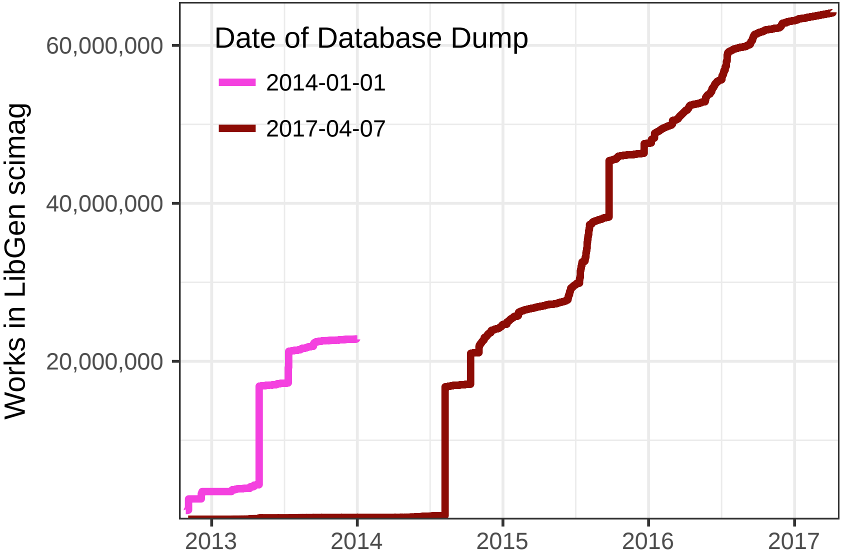


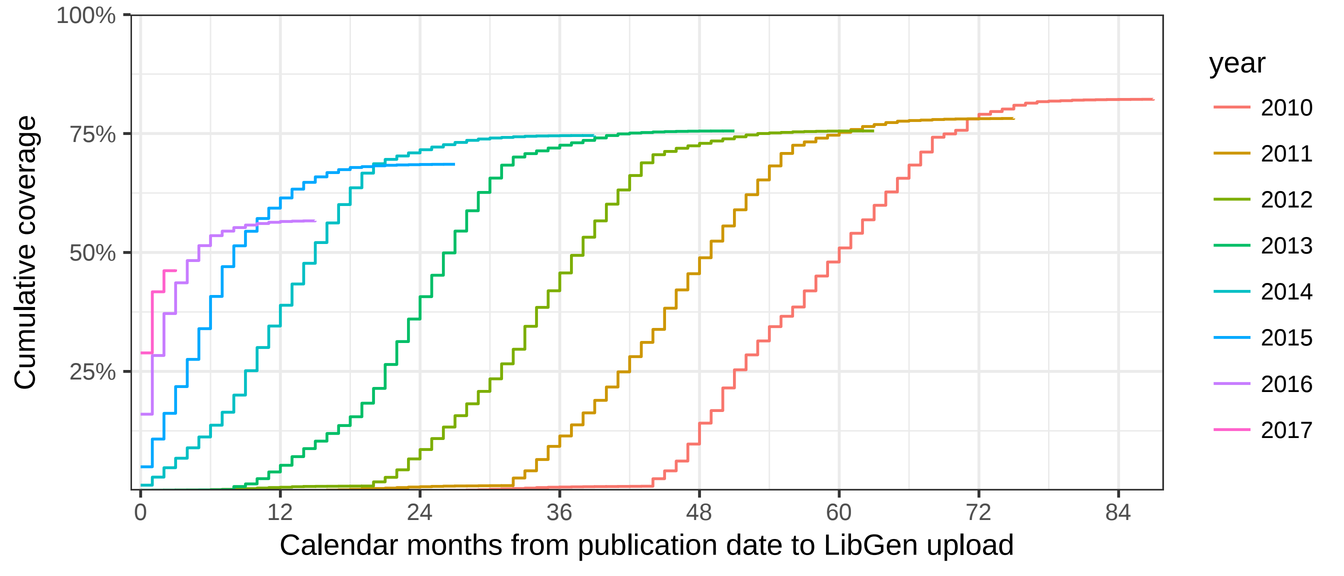
A
100%
80%
60%
40%
20%
0%
Sci-Hub Coverage
0.00–0.12
0.12–0.29
0.29–0.48
0.48–0.68
0.68–0.94
0.94–1.27
1.27–1.68
1.68–2.25
2.25–3.23
3.23–66.5
CiteScore Decile
B
32
15
7
3
1
0
Visitors per Article
Toll
Open
0
1
3
7
15
32
66
CiteScore



Number of Donations
0
20
40
60
80
100
120
140
160
2015-06
2015-07
2015-08
2015-09
2015-10
2015-11
2015-12
2016-01
2016-02
2016-03
2016-04
2016-05
2016-06
2016-07
2016-08
2016-09
2016-10
2016-11
2016-12
2017-01
2017-02
2017-03
2017-04
2017-05
2017-06
2017-07
2017-08
2017-09
2017-10
2017-11
2017-12


Number of Donations
160
140
120
100
80
60
40
20
0
Donations in BTC
20
18
16
14
12
10
8
6
4
2
0
Donations in USD
$12,000
$10,000
$8,000
$6,000
$4,000
$2,000
$0
2015-06
2015-07
2015-08
2015-09
2015-10
2015-11
2015-12
2016-01
2016-02
2016-03
2016-04
2016-05
2016-06
2016-07
2016-08
2016-09
2016-10
2016-11
2016-12
2017-01
2017-02
2017-03
2017-04
2017-05
2017-06
2017-07
2017-08
2017-09
2017-10
2017-11
2017-12


Date of Database Dump
2014-01-01
2017-04-07
Works in LibGen scimag
60,000,000
40,000,000
20,000,000
2013
2014
2015
2016
2017



100%
75%
50%
25%
Cumulative coverage
0
12
24
36
48
60
72
84
Calendar months from publication date to LibGen upload
year
2010
2011
2012
2013
2014
2015
2016
2017