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

# Bibliogifts in LibGen? A study of a text-sharing platform driven by biblioleaks and crowdsourcing

**Guillaume Cabanac**

Received: June 3, 2014 / Revised: August 18, 2014 / Accepted: October 1, 2014

**Abstract** Research papers disseminate the knowledge produced by the scientific community. Access to this literature is crucial for researchers and the general public. Apparently 'bibliogifts' are available online for free from text-sharing platforms. However, little is known about such platforms. What is the size of the underlying digital libraries? What are the topics covered? Where do these documents originally come from? This paper reports a study of the Library Genesis platform (LibGen). The 25 million documents (42 terabytes) it hosts and distributes for free are mostly research papers, textbooks, and books in English. The paper collection stems from isolated but massive paper uploads (71%) in line with a 'biblioleaks' scenario, as well as from daily crowdsourcing (29%) by worldwide users of platforms such as Reddit Scholar and Sci-Hub. By relating the DOIs registered at CrossRef and those cached at LibGen, this study reveals that 36% of all DOI papers are available for free at LibGen. This figure is even higher (68%) for three major publishers: Elsevier, Springer, and Wiley. More research is needed to understand to what extent researchers and the general public have recourse to such text-sharing platforms, and why.

**Keywords** Research · Publication · LibGen · Biblioleaks · Crowdsourcing · #icanhazpdf

## Introduction

Are biblioleaks inevitable? In a stimulating essay, opening like a sci-fi novel Dunn, Coiera, and Mandl (2014) coin the neologism 'biblioleaks' with reference to the information leakage cases that have made the news in the past few years. What if so-called hackers infiltrated the digital libraries of major subscription-based publishers, downloaded scientific articles *en masse*, and released them through anonymous peer-to-peer networks? With growing incentives from people hitting paywalls[1] daily, and virtually no technical barriers high enough to stop high-profile hackers, this essay stresses how feasible this scenario is.

G. Cabanac
Computer Science Department, University of Toulouse, IRIT UMR 5505 CNRS, 118 route de Narbonne, F-31062 Toulouse cedex 9, France
E-mail: guillaume.cabanac@univ-tlse3.fr

[1] According to Pickard and Williams (2014, p. 195) in the context of digital journalism "a paywall acts as a barrier between an internet user and a news organization's online content. To access content behind the paywall, users must pay a fee either on a one-time basis, or as part of a subscription."

Well... Biblioleaks *are* happening right now: various text-sharing platforms distribute tens of millions of documents online for free. However, these are still unknown to most of academia. Only a handful of papers acknowledge their existence in short passages (e.g., Egorov, 2013; Veletsianos, 2013) and no systematic study of the available collections has been undertaken until now.

In this paper, I study a prominent text-sharing platform: the Library Genesis, also known as LibGen (`libgen.org`). As of January 2014, it hosted and distributed 25 million digital documents, 95% of which being for educational purposes (i.e., scientific articles, books, and textbooks) and the other 5% for recreational purposes (i.e., fiction books and comics). This collection arguably covers a significant share of the scientific literature, as 1.3 million journal papers were published in 2006 (Björk, Roos, & Lauri, 2009).

In addition to biblioleaks feeding LibGen (i.e., a few isolated but massive additions of materials), there is evidence of *crowdsourcing* happening continuously. This terms encompasses many activities and more than 40 definitions (Estellés-Arolas & González-Ladrón-de-Guevara, 2012). I use it here to qualify the explicit collaboration of people who build a distributed collection of items that can be shared among users (Doan, Ramakrishnan, & Halevy, 2011, p. 88). I discuss how people crowdsource papers through various channels, such as Reddit and Sci-Hub. They request papers that eventually get cached at LibGen, thus contributing to increasing its collection. The growing popularity of the #icanhazpdf hashtag on Twitter is yet another manifestation of Wark's (2007) gift economy, with people relying on their social relations to access literature. However, such 'bibliogifts' may contravene copyright laws and more research is needed to assess the legal issues related to these materials and text-sharing platforms.

## Data and Method

The LibGen website runs a search engine allowing users to search and download materials directly from its servers. Files are available for individual download[2] or bulk download *via* peer-to-peer torrent files. LibGen releases all its code and data to foster the deployment of several mirrors (e.g., `gen.lib.rus.ec`) and to launch other websites (e.g., `bookzz.org`) according to its online documentation (see Appendix). The data stored in its relational databases associate the metadata of the available files with their URL.

This paper reports a study of the four MySQL databases run by LibGen (Table 1). All the cataloged materials are recorded with various metadata, such as title, authors, DOI,[3] journal or conference title, and volume/issue (for journal papers). Information specific to LibGen is also provided, such as when the file entered the cache and where the full-text version of the cached document is located on LibGen's servers.

In this study, I rely on DOIs assigned to papers to estimate the coverage of the scientific literature by LibGen. On the one hand, papers available from LibGen have their DOIs recorded in the `scimag` database. On the other hand, the CrossRef DOI registration agency represents 4,751 publishers and societies called registrants. This agency runs the CrossRef Depositor service[4] releasing the following data about the registrants and DOI allocations:

[2] See, e.g., `http://libgen.org/scimag/get.php?doi=10.1002%2Fasi.20971` or its mirrored version `http://lib.gen.in/36ba9ab556f46fcfcb52d37756b26891.pdf`.

[3] A Digital Object Identifier (DOI) is a character string (e.g., 10.1098/rstl.1665.0001) used to uniquely identify a digital object (see Davidson & Douglas, 1998). Most papers are published with a DOI today. The DOI of this particular paper is 10.1002/asi.23445.

[4] cf. the CrossRef system reports `http://www.crossref.org/06members/`

- The *journals* published by each registrant with the number of DOIs assigned per journal. These DOIs are used to identify papers, editorials, reviews, and so on. DOIs are grouped according to journal titles. For example, the 65 volumes of *JASIST* are distributed into four records of the Depositor, as the journal has been successively known under four titles (Table 2). Note that the 'journals' listed in the Depositor include bulletins, news, forum, and more. For example, the following are listed for the *Association for Computing Machinery*: *ACM Transactions on X*, *ACM SIGGROUP Bulletin*, *ACM SIGACT News*, *ACM SIGBIO Newsletter*, *ACM SIGIR Forum*, *ACM Inroads*, *ACM SIGPC Notes*, *ACM Queue*, and more.
- The *conference proceedings* published by each registrant with the number of DOIs assigned per proceedings. These DOIs are used to identify conference papers.
- The *books* published by each registrant with the number of DOIs assigned per book. These DOIs are used to identify book parts, such as book chapters.

Each DOI has a prefix that uniquely identifies its registrant. For example, the registrant of the DOI '10.1098/rstl.1665.0001' is 'The Royal Society' (with identifier '10.1098') and 'rstl.1665.0001' is the identifier of (Oldenburg, 1665) in the library of the registrant. Although the DOI System was launched in 1997 (Davidson & Douglas, 1998), publishers assigned DOIs to publications pre-dating the introduction of DOIs retrospectively — even the first paper published in the first journal has a DOI (Oldenburg, 1665). I rely on registrant identifiers to compute the coverage of publishers' digital libraries by LibGen in this paper.

## Browsing the Digital Shelves of LibGen

Figure 1 shows the growth of the collection of educational (`scimag` and `bookwarrior` databases) and recreational materials (`fiction` and `comics` databases) available at LibGen. The largest database (`scimag`) comprises 22 million scientific articles that were progressively collected from October 28, 2012 onwards. The number of papers cached per day is usually

**Table 1** Database dumps of LibGen downloaded from `http://gen.lib.rus.ec/dbdumps` on January 5, 2014. Each database stores the metadata and link to full-text for the cached materials.

| Database | Cached materials | | |
|---|---|---|---|
| | Number | Size (terabytes) | Type of material |
| `scimag` | 22,829,088 | 15 | Scientific articles |
| `bookwarrior` | 1,126,091 | 13 | Scientific books and textbooks |
| `fiction` | 932,307 | 1 | Fiction books |
| `comics` | 472,269 | 13 | Comics |
| Total | 25,359,755 | 42 | — |

**Table 2** History of title changes for the *JASIST* journal.

| Active period | Volumes | Title of the journal |
|---|---|---|
| 1950–1969 | 1–20 | *American Documentation* |
| 1970–2000 | 21–51 | *Journal of the American Society for Information Science* |
| 2001–2013 | 52–64 | *Journal of the American Society for Information Science and Technology* |
| 2014– | 65– | *Journal of the Association for Information Science and Technology* |

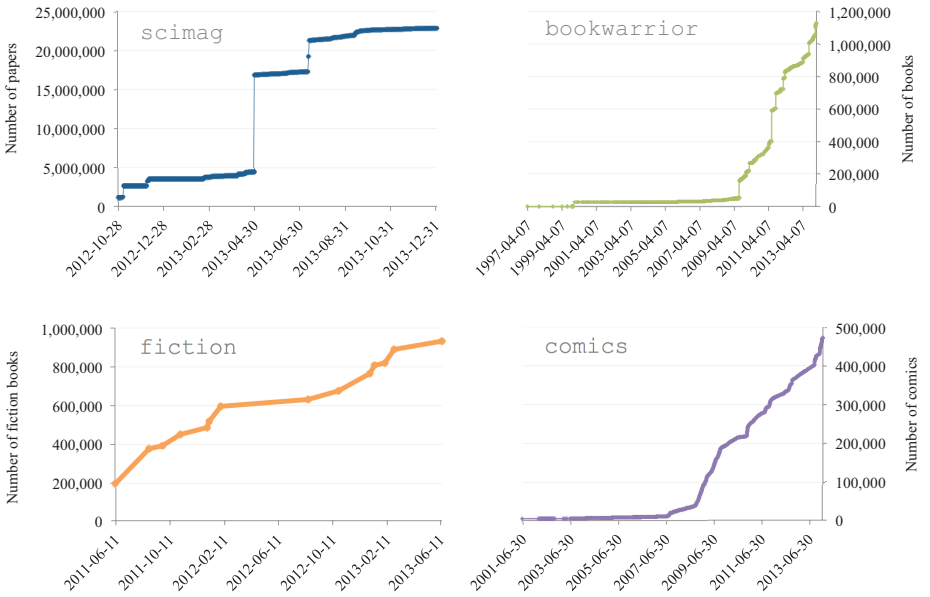


**Fig. 1** Growth of the number of educational and recreational materials available at LibGen as of January 5, 2014.

low ($Mdn = 2{,}720$ and $Avg = 52{,}957$) compared to the four discontinuities clearly visible in the cumulative distribution. April 30, 2013 saw the largest growth with 12,466,342 papers added on that day, representing 55% of the current collection. Overall, 71% of the `scimag` papers stemmed from uploads of more than 100,000 papers a day, which occurred on 13 days in total. These figures suggest that biblioleaks as imagined in the essay by Dunn et al. (2014) have already happened.

The three other databases represent 10% only of the LibGen collection. Scientific books (`bookwarrior` database) were collected from 1997 onwards, with a strong increase in the frequency of additions after 2009. Fiction books (`fiction` database) were added from 2011 onwards, though less frequently than scientific books. Finally, the comics collection (`comics` database) started in 2001 and it has grown on a daily basis since 2009.

The following sections discuss the features of the 25 million materials available from the four databases run by LibGen.

### *Scientific Articles*

LibGen distributes the PDF files of scientific articles published in 27,134 journals by 1,342 publishers (Figure 2). Journals are partially indexed, with only one paper available for some journals and all the papers published in each volume for others.[5] LibGen covers 78% of all journal titles (Table 3). This large coverage suggests that LibGen is not focused on a few publishers or journals. Some publish open-access journals only (e.g., BioMed Central, Hindawi, PLOS), but this is not the case for the majority of publishers shown in Figure 2.

[5] See the individual record of each journal available from the journal list (e.g., `http://libgen.org/scimag/journals.php?letter=A` lists journals with titles starting with 'A').

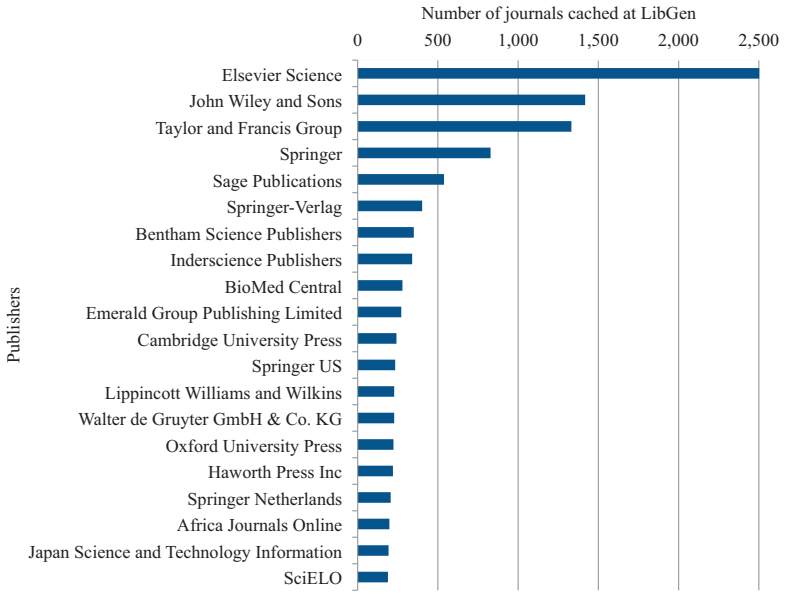


**Fig. 2** Distribution of the number of journals with at least one paper cached at LibGen ($N = 27{,}134$), by publisher — limited to the top 20 publishers among the 1,342 listed in the database.

Table 3 focuses on three of such major publishers, where the number of journals indexed was summed for all 'variants' of a publisher's name (e.g., Springer in Table 3 aggregates the values of Springer, Springer-Verlag, Springer US, Springer Netherlands, and the other lower-ranked variants shown in Figure 2). LibGen covers (at least partially) an average of 59% of all the journals published by Elsevier, Springer, and Wiley.

**Table 3** Number of journals from which at least one paper is available in LibGen and the number of journals published by DOI registrants at CrossRef.

| Publisher | Number of Journals | | Coverage of LibGen |
|---|---|---|---|
| | LibGen | CrossRef | |
| All | 27,134 | 34,670 | 78% |
| Elsevier | 2,503 | 3,925 | 64% |
| Springer | 1,814 | 3,394 | 53% |
| Wiley | 1,418 | 2,401 | 59% |

After assessing the presence of *journals* in LibGen, I wondered about the distribution of the cached *papers* across publishers. The prefix of the DOIs assigned to papers (identifying registrants) was used as a proxy for publishers (database field `scimag.publishers.doicode`). There were 1,064 registrants overall, with a highly skewed distribution of cached papers (Figure 3). After grouping registrants related to a single publisher, Elsevier (42%), Springer (20%), and Wiley (22%) account for 83% of all of the papers cached at LibGen.

The coverage of LibGen in terms of research papers (Table 4) was assessed by dividing the number of cached papers (LibGen) by the number of papers published in each of the

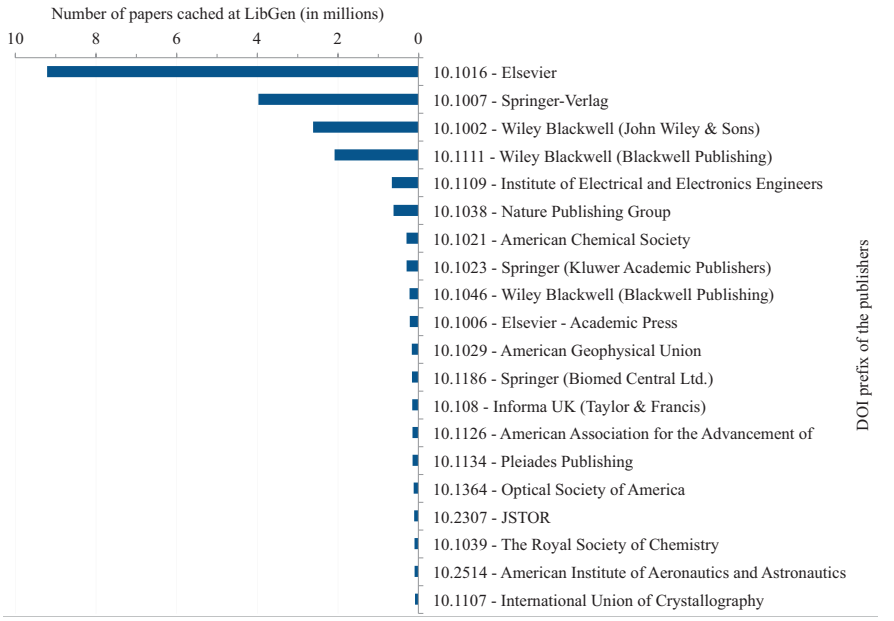


**Fig. 3** Distribution of the 22,829,088 papers available from LibGen, for each registrant (DOI prefix and associated publishing house). Only the top 20 publishers in number of cached papers are showed, totaling 94% of all cached papers.

digital libraries (i.e., the number of assigned DOIs given by CrossRef). Note that all the DOIs assigned by CrossRef might not have been used by registrants yet (e.g., papers to appear), which implies that the results reported here are conservative. LibGen hosts 36% of all papers with DOIs. Its coverage is higher for the three aforementioned publishers with an average of 68% of all papers published by Elsevier, Springer, and Wiley that are distributed at LibGen.

**Table 4** Number of papers available at LibGen *versus* registered with a DOI at CrossRef.

| Publisher | Number of Papers | | Coverage of LibGen |
|---|---|---|---|
| | LibGen | CrossRef | |
| All | 22,829,088 | 63,580,196 | 36% |
| Elsevier | 9,579,795 | 12,398,807 | 77% |
| Springer | 4,504,256 | 8,538,817 | 53% |
| Wiley | 4,973,954 | 6,848,146 | 73% |

Wondering about the topics of the papers available from LibGen, I labeled each paper with the research field of the journal it was included in, as found in the *Essential Science Indicators*[6] (ESI). This database published by Thomson Reuters lists 11,155 journals, each one being classified into one of 22 research fields. About one-third of the LibGen papers appeared in journals not included in the ESI (Figure 4). Clinical medicine and chemistry

[6] http://about.esi.incites.thomsonreuters.com

feature more than 10% of the remaining papers, while the other research fields represent 5% or less of all LibGen papers.

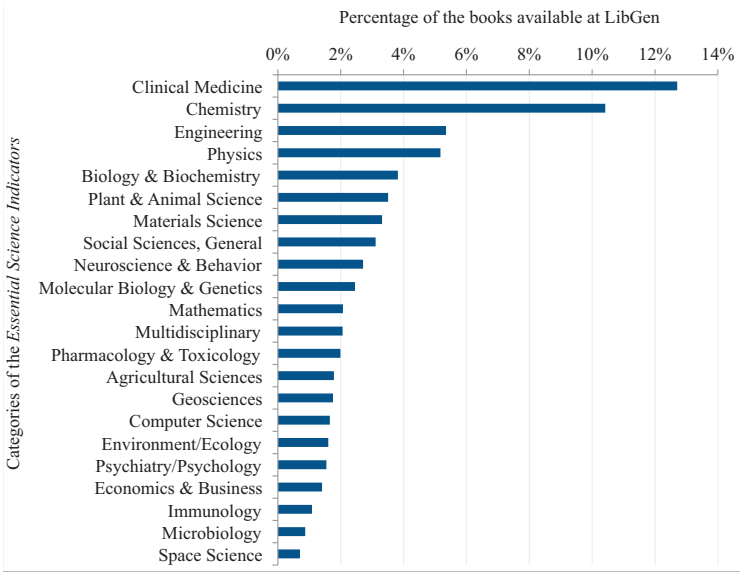


**Fig. 4** Distribution of the 22,829,088 papers available at LibGen across the 22 research fields of the *Essential Science Indicators* database. Papers with unknown category (i.e., 29% of all the 22,829,088 recorded papers) are not included in this figure.

The collection distributed at LibGen is up-to-date, as the most recent papers are published in 2014. The oldest paper available (Oldenburg, 1665) appeared in the first issue of the *Philosophical Transactions of the Royal Society of London*, which is recognized as the first scientific journal (Singleton, 2014). Figure 5 shows the distribution and exponential model of the number of cached papers published during the last 60 years. Half of the papers where published after year 2000. The decline in 2013 with 'only' 601,969 papers available (compared to 1,156,322 in 2012) looks like an interruption of service. However, more likely, it might reflect the delay of paper caching.

### *Scientific Books and Textbooks*

LibGen also distributes 1,126,091 scientific books and textbooks, 88% of which were published between 1953 and 2013 (Figure 6). Most of the books and textbooks are written in English (65%) and Russian (22%). PDF is the most frequent file format (71% of `.pdf` files), followed by DjVu (16% of `.djvu` files), and electronic publication (5% of `.epub` files). Some books are even available in multiple formats.

The `bookwarrior` database features a list of topics available for indexing purposes. The topic of 44% of the books and textbooks is known, although it is unclear how topics were assigned. Figure 7 shows the distribution of books and textbooks across these topics, with mathematics being the most frequent topic. Both the natural sciences and the social sciences are represented.

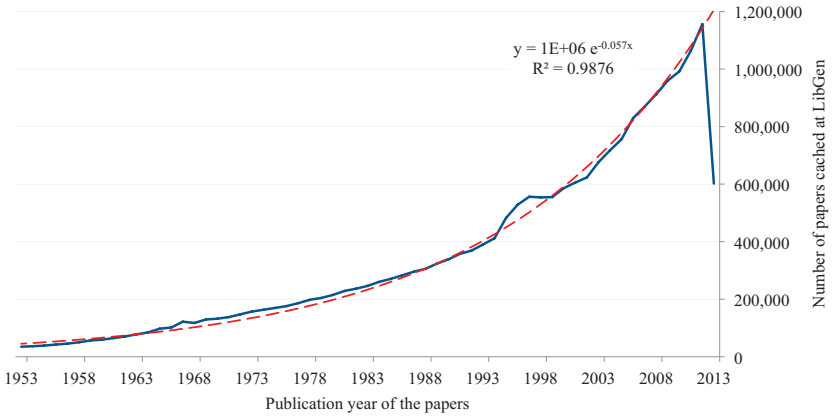


**Fig. 5** Exponential distribution of the number of papers available at LibGen that were published between 1953 and 2013. This sample represents 95% of all cached papers published between 1665 and 2014 ($N$ = 22,829,088).

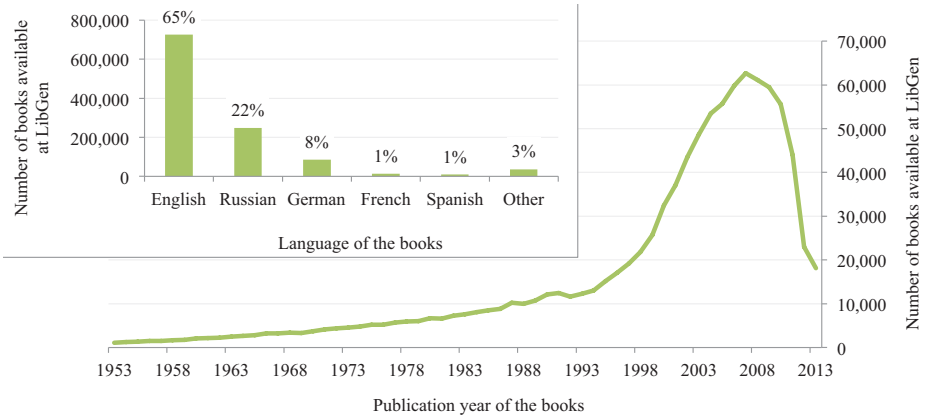


**Fig. 6** Distribution of the publication years and languages recorded for the scientific books and textbooks available at LibGen. Books with unknown publication date (i.e., 9% of all the 1,126,091 recorded books) and unknown language (1%) are not included in this figure.

### *Fiction books*

In the category of recreational materials, LibGen distributes 932,307 fiction books. The publication date is known for only 46% of the books, which were mainly published after 2008 (Figure 8). This period corresponds to an increased availability of e-books (Walters, 2014). Most of the fiction books are written in English (74%). Electronic publication is the most frequent file format (30% of `.epub` files), followed by PDF (20% of `.pdf` files), and Mobipocket (12% of `.mobi` files).

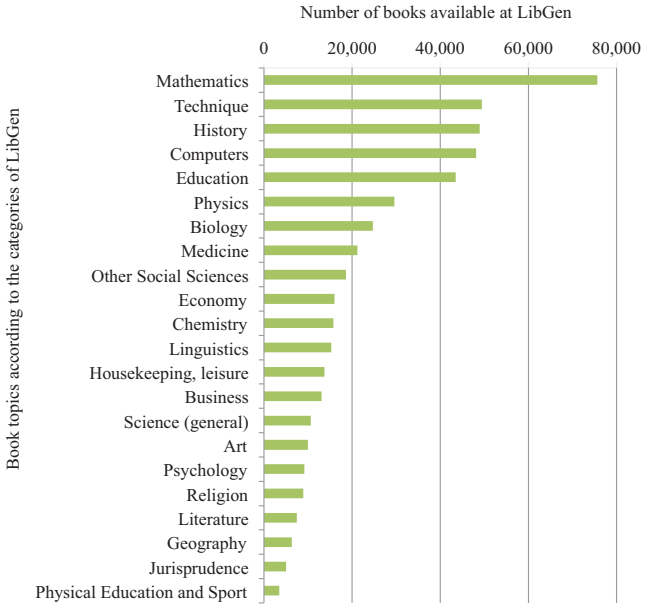


**Fig. 7** Topics of the scientific books and textbooks available at LibGen. Books with unknown topic (i.e., 56% of all the 1,126,091 recorded books) are not included in this figure.

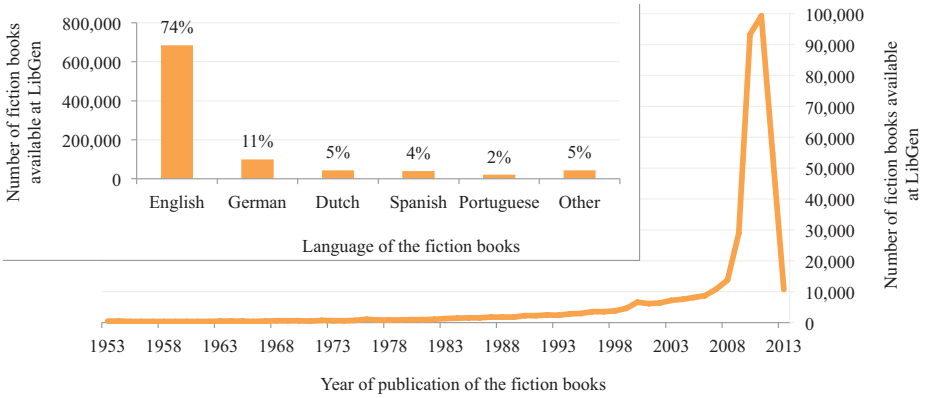


**Fig. 8** Distribution of the publication years recorded for the fiction books available at LibGen. Those with unknown publication date (i.e., 54% of all 932,307 recorded fiction books) are not included in this figure.

### *Comics*

Comics are another type of recreational material distributed at LibGen. Each of the 472,269 files distributed comes with its cover and title, such as "Superman The Dailies 1939-1942 (2006)." The Comic Book Archive is the most popular format (68% of `.cbr` files and 26% of `.cbz` files that is, compressed `.cbr` files), followed by PDF (2% of `.pdf` files). Unfortunately, fields documenting the other types of materials (e.g., publication date, topics, language) are not provided in the `comics` database.

**Population of LibGen with Biblioleaks and Crowdsourcing**

The growth of LibGen suggests that it has benefited from a few isolated but massive additions of scientific papers to its cache (Figure 1). For instance, 71% of the paper collection was uploaded in 13 days at a rate of 100,000+ papers a day. It is likely that such massive collections of papers result from biblioleaks (Dunn et al., 2014), but one can only speculate about this because of the undocumented source of each file cached at LibGen.

With a median of 2,720 new papers uploaded a day, most additions to the text-sharing platform are not massive. People crowdsource papers to LibGen directly or indirectly. Direct uploading from a user provided PDF file and DOI is available from a regular webform hosted at `libgen.org`. Indirect uploading occurs when the papers that users request (*via* services such as Reddit Scholar and Sci-Hub — discussed in the following sections) get automatically uploaded to LibGen, thus contributing to its growth.

*Crowdsourcing of Papers via the Scholar Subreddit*

The `reddit.com` website is used to share (and vote for) hyperlinks to contents organized in categories called 'subreddits' (Weninger, Zhu, & Han, 2013). The Scholar subreddit[7] was launched in June 2009 "for requesting and sharing articles available in various databases, as well as discussion relating to the material." Users are advised to "see if [the] article is already available by checking LibGen, Google, and Google Scholar" before requesting it. Another caveat is displayed on the front-page: "If the request isn't urgent, please try an interlibrary loan (ILL) first. ILL avoids potential copyright issues and lets libraries know which subscriptions are useful." Users of the service submit the metadata and hyperlink (DOI) of the requested papers, hoping that another user entitled to download the paper for free will share the PDF file by uploading it to LibGen and posting the hyperlink of the cached file on the text-sharing platform. Once a request is fulfilled, the Reddit user is advised to "mark [his/her] request as NSFW after it's been found." Posts marked NSFW (not safe for work) are then hidden from the other users to prevent them from seeking any paper already found. It is thus impossible to see the history of all requests on this subreddit. Nonetheless, I estimated that there are 17,000 posts per year requesting the PDF of a paper, based on my systematic monitoring of Reddit Scholar for a 3-month period (see code in Appendix). Unfortunately, I do not know the number of fulfilled requests, and the number of subsequent PDF downloads.

*Crowdsourcing of Papers via Sci-Hub*

The `sci-hub.org` website is used to download paywalled papers without paying a subscription to subscription-based publishers (Figure 9). Users requesting papers not present in LibGen are advised to download them from Sci-Hub ("Nothing was found: search in `sci-hub.org`?").

On Sci-Hub, papers are requested by submitting their DOIs *via* the form displayed on the front-page of its website. Sci-Hub is connected to computers worldwide that are running on networks entitled to access to subscription-based publishers. The user's browser is redirected to one of these computers, which acts as a proxy server to show the publisher's page for the requested DOI. The name and logo of the compromised institution are displayed — mostly

[7] `http://www.reddit.com/r/Scholar`

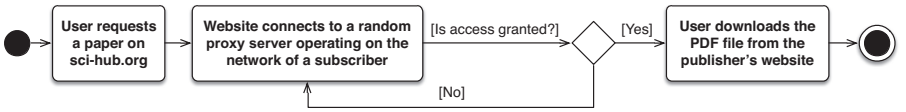


**Fig. 9** This activity diagram shows how Sci-Hub serves PDF files of papers through proxy servers operating on the networks of worldwide institutions. The user randomly switches proxy servers until he/she get connected to a proxy entitled to download the PDF of the requested paper.

universities and libraries — as if the user was operating from this distant location. When the proxy is operating on a network entitled to access the paper in full-text version, the user can download the associated PDF file. These PDF files bear the marks added by most publishers: Date and time of download, originating IP, and name of the (distant) institution. If the full-text version is not available from the current location, the user can randomly switch to another proxy server operating in a different institution, until he/she eventually finds a proxy entitled to download the PDF of the coveted paper. Sci-Hub also acts as a proxy to subscription-based services, such as the *Web of Knowledge* and the *Journal Citation Reports* run by Thomson Reuters (see screenshots in Appendix).

PDF files downloaded from Sci-Hub are served to users *and* automatically added to LibGen (if not already present). Further requests for already downloaded DOIs are fulfilled by serving the PDF file directly from the LibGen cache. This is how LibGen grows daily by adding the papers crowdsourced by the users of Sci-Hub.

## Related Work

Volentine and Tenopir (2013) studied the value and outcome of scholarly reading for academic staff. Academics described scholarly articles as "critical," "essential," or "vital" to their work but, in the meantime, they complained about not having access to all the articles that they would like to read. Academics even expressed frustration and stated that they could not do research without the availability of scholarly articles.

Scientists and the general public (Davis & Walters, 2011) rely on a wide range of channels to access research literature. Paper-based journals once embodied the dominant channel and one can still read papers in printed issues of journals or request hard copies *via* interlibrary loan. However, scholars are increasingly downloading papers and reading them on-screen, or printing them off (Tenopir, King, Edwards, & Wu, 2009; Volentine & Tenopir, 2013). Papers are now available online from a variety of platforms. On the one hand, publishing houses run digital libraries, such as Elsevier ScienceDirect, and the Wiley Online Library. For paywalled papers (i.e., non open-access), one needs to pay a fee or to be affiliated to a subscribing institution. For open-access papers, there is no such barrier and papers are freely available to anyone to read (Harnad et al., 2008). On the other hand, one can read papers for free from preprint repositories, such as ArXiv (Davis & Fromerth, 2007) and institutional websites where authors self-archive their papers (Harnad, 2001).

People failing to retrieve papers through these channels have recourse to another more direct strategy: paper requests. There is a long tradition of readers requesting paper reprints from authors by post or email (Hartley, 2004a, 2004b). But the success of this strategy depends on various factors, such as authors' willingness to honor such requests, and their swiftness in finding and sending the paper.

In a reflexive and autobiographic paper, Veletsianos (2013) considers how researchers participate in social media. He discusses how they defy restrictions to get papers through crowdsourcing. Research-oriented platforms (e.g., Academia.edu, Mendeley, and ResearchGate) now support readers who connect with a larger audience than the paper's author(s) who may or may not answer requests. Now, readers who use social media ask their (dozens of) followers/connections to check if they have access to the coveted paper for them. This practice is popular on Twitter, the leading microblogging service where users post short messages called 'tweets' (Efron, 2011). The #icanhazpdf hashtag was coined in 2011 to mark tweets requesting the PDF file of a paper (Bond, 2013; Dunn et al., 2014; Kroll, 2011). These tweets invite users entitled to access the requested paper to send it by email or direct message (i.e., non-public). Figure 10 illustrates this strategy. The first tweet recommends to use #icanhazpdf with the metadata of the requested paper, and advises the deletion of the tweet-request once the PDF is received. The second tweet requests a paper with #icanhazpdf and a link to the paper. While #icanhazpdf was far from being mainstream with about 1,000 requests posted between May 2012 and April 2013 (Liu, 2013), its use is clearly increasing.[8]

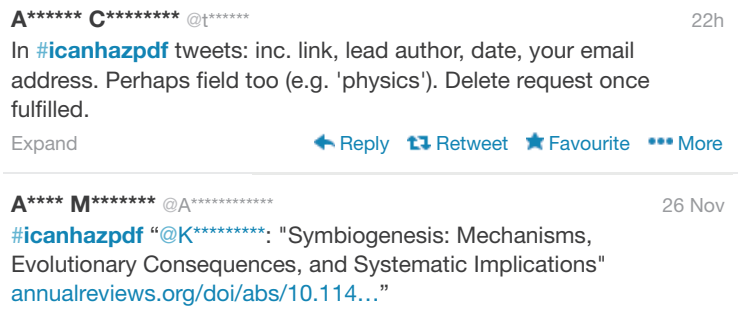


**Fig. 10** Example of two tweets published with the #icanhazpdf hashtag on Twitter (user accounts are anonymized). The first tweet recalls informal guidelines when requesting papers. The second tweet requests a paper by mentioning its title and hyperlink.

A final note relates to methodological discrepancies found in several studies that estimate the share of the research literature available online for free. Van Noorden (2013) summarized these in a *Nature* paper headlined "Half of 2011 papers now free to read" based on a report to the European Commission (Archambault et al., 2013). According to Khabsa and Lee Giles (2014, p. e93949):

> Our estimates show that at least 114 million English-language scholarly documents are accessible on the web, of which Google Scholar has nearly 100 million. Of these, we estimate that at least 27 million (24%) are freely available since they do not require a subscription or payment of any kind.

Prior work by Gargouri, Larivière, Gingras, and Harnad (2012, p. 286) relied on a program to "estimate what percentage of journal articles [. . . ] were freely available on the Web (OA) in September 2011." Khabsa and Lee Giles (2014) used specialized search engines (i.e., Google Scholar and Microsoft Academic Search) to search for a set of papers and estimate the number of scholarly documents available on the web. Björk et al. (2010) designed a hybrid method combining the use of a standard search engine (i.e., Google) and manual

[8] The query #icanhazpdf on Twitter retrieves many tweets posted every day about paper requests, see https://twitter.com/search?q=%23icanhazpdf.

validation. This allowed them to discard articles "openly accessible by accident, clearly against the [subscription] site policy" (Björk et al., 2010, p. e11273). While acknowledging this, Archambault et al. (2013, p. 5) "made no such attempt as [their] research question seeks to find what is available for free at a given time."

I believe that future studies should stress that apparent bibliogifts (i.e., papers available online for free) come from a variety of channels, such as publishers' libraries, institutional repositories, preprint repositories, personal websites, and text-sharing platforms. The scholarly documents available from LibGen through a URL (and thus retrievable from a search engine[9]) are free to download, but they are certainly not always distributed as open-access.

## Conclusion

Access to the research literature is essential to the work of researchers (Volentine & Tenopir, 2013) and the education of the general public (Davis & Walters, 2011). Gigantic text-sharing platforms supported by new practices on social media (e.g., #icanhazpdf on Twitter) are currently emerging to bypass any barriers preventing readers from accessing published research papers. This phenomenon is so recent that it has hardly been studied to date.

The present study focused on the LibGen text-sharing platform hosting 25 million documents (42 terabytes in size). This collection contains 95% of educational materials, such as scientific articles, books, and textbooks. It also distributes recreational materials, such as fiction books, and comics. Most of the documents were added *en masse*, as in the hypothetical 'biblioleaks' scenario imagined by Dunn et al. (2014). No less than 71% of the collection was collected at a rate of 100,000+ papers a day. Documents were also crowdsourced: directly on LibGen or indirectly via Sci-Hub. This latter service automatically adds to LibGen all of the retrieved papers from the proxy servers it operates in worldwide institutions. These figures reveal the extraordinary breadth of LibGen in academic publishing. As a comparison, only 1.3 million journal papers were published in 2006 (Björk et al., 2009). Another point of comparison lies in the *United States v. Aaron Swartz* case (Atkinson & Fitzgerald, 2014; Sims, 2011) which was concerned about the legality of 4.8 million papers downloaded from JSTOR — "about [their] entire database."[10]

Legal issues certainly apply to text-sharing platforms like LibGen for operating paper caching (Borrull & Oppenheim, 2004; Oppenheim, 2008). Unfortunately, there is no systematic way to separate the documents that are in the public domain or published in open-access from those that are not. Reddit Scholar calls users' attention to legal issues: "Please be aware of copyright issues and Fair Use Copyright".[11] Varian (2005, p. 125) elaborates on the concept of Fair Use in the U.S. as follows:

> "Furthermore, under certain conditions, extracts from works that have been copyrighted may be reproduced. The U.S. Copyright Act of 1976 indicates that reproductions for purposes such as "criticism, comment, news reporting, teaching (including multiple copies for classroom use), scholarship, or research, is not an infringement of copyright." This "fair use doctrine" is essentially a *defense* against an infringement claim, and U.S. law indicates several factors that can be taken into account in such a defense, including the purpose of the use, the nature of the work, the proportion of the work copied

[9] While `http://libgen.org/robots.txt` currently forbids robots to crawl LibGen, not every mirror enforces the same policy. Moreover, documents hosted at LibGen can also be indexed when a crawler follows a direct link to it (e.g., `http://lib.gen.in/36ba9ab556f46fcfcb52d37756b26891.pdf`) from any online document. See for instance the result of the query "`filetype:pdf site:libgen.org OR site:lib.gen.in OR site:gen.lib.rus.ec`" submitted to Google in Appendix.

[10] JSTOR Evidence in *United States vs. Aaron Swartz* `http://docs.jstor.org/`.

[11] Including a link to `http://www.lib.purdue.edu/uco/CopyrightBasics/fair_use.html`

> and the economic impact of the use on the market. The fair use exemption is notoriously vague, but perhaps intentionally so, as it allows the law to deal flexibly with cases as they arise."

More research is needed to clarify the consequences of enjoying these bibliogifts. What are the penalties (if any?) incurred by the various worldwide users of text-sharing platforms, including the owners running the platforms, the users populating them *en masse* with biblioleaks, the users crowdsourcing one paper at a time, and the readers who download papers for free?

The availability of text-sharing platforms such as LibGen raises a variety of questions. How many people use this service? How did they become acquainted with it? What is the balance of users between scientists and the general public? For what purposes do users need the papers? Are they for educational purposes only? Are they for research-related projects related with, for example, text-mining (Van Noorden, 2012)? These open questions, and many more, require further research and studies involving a variety of users familiar with text-sharing platforms.

## Appendix: Supporting Information

Additional supporting information may be found in the online version of this article at the publisher's website or at http://www.irit.fr/publis/IRIS/2016_JASIST_C.zip.

**Acknowledgements** I am grateful to Jérémie Clos, Adam Dunn, Yves Gingras, James Hartley, Marie-Dominique Heusse, Gilles Hubert, and András Schubert for their feedback on previous versions of the paper. I also acknowledge the anonymous reviewers and the Editor-in-Chief of *JASIST* for their valuable comments.

## References

Archambault, E., Amyot, D., Deschamps, P., Nicol, A., Rebout, L., & Roberge, G. (2013). *Proportion of Open Access Peer-Reviewed Papers at the European and World Levels—2004-2011* (Report to the European Commission). Science-Metrix.

Atkinson, B., & Fitzgerald, B. (2014). The meaning and future of copyright. In *A short history of copyright* (pp. 129–136). Springer. doi: 10.1007/978-3-319-02075-4_14

Björk, B.-C., Roos, A., & Lauri, M. (2009). Scientific journal publishing: Yearly volume and open access availability. *Information Research*, *14*(1).

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

Bond, A. (2013, October 5). *How #icanhazpdf can hurt our academic libraries.* (Retrieved March 14, 2016, from http://labandfield.wordpress.com/2013/10/05/how-icanhazpdf-can-hurt-our-academic-libraries)

Borrull, A. L., & Oppenheim, C. (2004). Legal aspects of the Web. *Annual Review of Information Science and Technology*, *38*(1), 483–548. doi: 10.1002/aris.1440380111

Davidson, L. A., & Douglas, K. (1998). Digital Object Identifiers: Promise and problems for scholarly publishing. *Journal of Electronic Publishing*, *4*(2). doi: 10.3998/3336451.0004.203

Davis, P. M., & Fromerth, M. J. (2007). Does the arXiv lead to higher citations and reduced publisher downloads for mathematics articles? *Scientometrics*, *71*(2), 203–215. doi: 10.1007/s11192-007-1661-8

Davis, P. M., & Walters, W. H. (2011). The impact of free access to the scientific literature: A review of recent research. *Journal of Medical Library Association*, *99*(3), 208–217. doi: 10.3163/1536-5050.99.3.008

Doan, A., Ramakrishnan, R., & Halevy, A. Y. (2011). Crowdsourcing systems on the World-Wide Web. *Communications of the ACM*, *54*(4), 86–96. doi: 10.1145/1924421.1924442

Dunn, A. G., Coiera, E., & Mandl, K. D. (2014). Is Biblioleaks inevitable? *Journal of Medical Internet Research*, *16*(4), e112. doi: 10.2196/jmir.3331

Efron, M. (2011). Information search and retrieval in microblogs. *Journal of the American Society for Information Science and Technology*, *62*(6), 996–1008. doi: 10.1002/asi.21512

Egorov, V. S. (2013). The specifics of access to scholarly journal literature in the digital society. *Scientific and Technical Information Processing*, *40*(1), 46–57. doi: 10.3103/s0147688213010103

Estellés-Arolas, E., & González-Ladrón-de-Guevara, F. (2012). Towards an integrated crowdsourcing definition. *Journal of Information Science*, *38*(2), 189–200. doi: 10.1177/0165551512437638

Gargouri, Y., Larivière, V., Gingras, Y., & Harnad, S. (2012). Green and gold open access percentages and growth, by discipline. In E. Archambault, Y. Gingras, & V. Larivière (Eds.), *Proceedings of 17th international conference on science and technology indicators* (pp. 285–292). Montréal: Science-Metrix and OST.

Harnad, S. (2001). The self-archiving initiative [Commentary]. *Nature*, *410*(6832), 1024–1025. doi: 10.1038/35074210

Harnad, S., Brody, T., Vallières, F., Carr, L., Hitchcock, S., Gingras, Y., … Hilf, E. R. (2008). The access/impact problem and the Green and Gold roads to Open Access: An update. *Serials Review*, *34*(1), 36–40. doi: 10.1016/j.serrev.2007.12.005

Hartley, J. (2004a). On requesting conference papers electronically. *Journal of Information Science*, *30*(5), 475–479. doi: 10.1177/0165551504047826

Hartley, J. (2004b). On requesting re-prints electronically. *Journal of Information Science*, *30*(3), 280–284. doi: 10.1177/0165551504044671

Khabsa, M., & Lee Giles, C. (2014). The number of scholarly documents on the public Web. *PLoS ONE*, *9*(5), e93949. doi: 10.1371/journal.pone.0093949

Kroll, D. J. (2011, December 22). *#icanhazpdf: Civil disobedience?* (Retrieved March 14, 2016, from http://cenblog.org/terra-sigillata/2011/12/22/icanhazpdf-civil-disobedience)

Liu, J. (2013, May 9). *Interactions: The numbers behind #icanhazpdf.* (Retrieved March 14, 2016, from http://www.altmetric.com/blog/interactions-the-numbers-behind-icanhazpdf)

Oldenburg, H. (1665). Epistle dedicatory. *Philosophical Transactions of the Royal Society of London*, *1*(1–22). doi: 10.1098/rstl.1665.0001

Oppenheim, C. (2008). Legal issues for information professionals IX: An overview of recent developments in the law, in relation to the Internet. *Journal of Documentation*, *64*(6), 938-955. doi: 10.1108/00220410810912460

Pickard, V., & Williams, A. T. (2014). Salvation or folly? The promises and perils of digital paywalls. *Digital Journalism*, *2*(2), 195–213. doi: 10.1080/21670811.2013.865967

Sims, N. (2011). Library licensing and criminal law: The Aaron Swartz case. *College & Research Libraries News*, *72*(9), 534–537.

Singleton, A. (2014). The first scientific journal. *Learned Publishing*, *27*(1), 2–4. doi: 10.1087/20140101

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

Van Noorden, R. (2012). Trouble at the text mine [News In Focus]. *Nature*, *483*(7388), 134–135. doi: 10.1038/483134a

Van Noorden, R. (2013). Half of 2011 papers now free to read [News In Focus]. *Nature*, *500*(7463), 386–387. doi: 10.1038/500386a

Varian, H. R. (2005). Copying and copyright. *Journal of Economic Perspectives*, *19*(2), 121–138. doi: 10.1257/0895330054048768

Veletsianos, G. (2013). Open practices and identity: Evidence from researchers and educators social media participation. *British Journal of Educational Technology*, *44*(4), 639–651. doi: 10.1111/bjet.12052

Volentine, R., & Tenopir, C. (2013). Value of academic reading and value of the library in academics' own words. *Aslib Proceedings*, *65*(4), 425–440. doi: 10.1108/AP-03-2012-0025

Walters, W. H. (2014). E-books in academic libraries: Challenges for sharing and use. *Journal of Librarianship and Information Science*, *46*(2), 85–95. doi: 10.1177/0961000612470279

Wark, M. (2007). Copyright, copyleft, copygift. *Open*, *12*, 22–29.

Weninger, T., Zhu, X. A., & Han, J. (2013). An exploration of discussion threads in social news sites: A case study of the Reddit community. In *ASONAM'13: Proceedings of the 5th international conference on advances in social networks analysis and mining* (pp. 579–583). ACM Press. doi: 10.1145/2492517.2492646