3 Library Genesis in Numbers: Mapping the Underground Flow of Knowledge

Balázs Bodó

Chapter 2 documented the largely Russian social history of pirate book sites. This chapter explores the question of the growth and impact of the Library Genesis (or LibGen) network, via a close look at its collections and traffic. This quantitative analysis clarifies how these services operate, what publics they serve, and ultimately what harms to publishers and authors can be reasonably attributed to them. LibGen and its mirror sites infringe the copyrights on hundreds of thousands of works, potentially undercutting the market for those works. But they also respond to clear (and sometimes not so clear) market failures where work is unavailable or unaffordable, and they play a role in expanding global access to scientific and scholarly work. On what basis can we evaluate these trade-offs? To date, there has been no substantive account of the shape, reach, or impact of these archives. This chapter takes some steps in that direction.

The first section reconstructs the growth of the LibGen collection through an examination of changes in its catalog over time—mapping it by language and subject matter, and evaluating how much of it is accessible through legal alternatives. The second section discusses the demand for books on these sites, based on download data acquired from one of the LibGen mirror sites. Here we look at what is being downloaded and by whom. The third section connects the supply and demand discussions to reflections on the wider impact of these pirate archives on libraries, higher education institutions, and authors.[1]

The Supply of Documents in Library Genesis

Between 2008 (the start of LibGen), and April 2014 (the end of our analysis), the size of the LibGen catalog grew from nearly 34,000 items to almost 1.2 million records.[2] Figure 3.1 shows the number of documents added to the collection each month between January 2008 and April 2014.

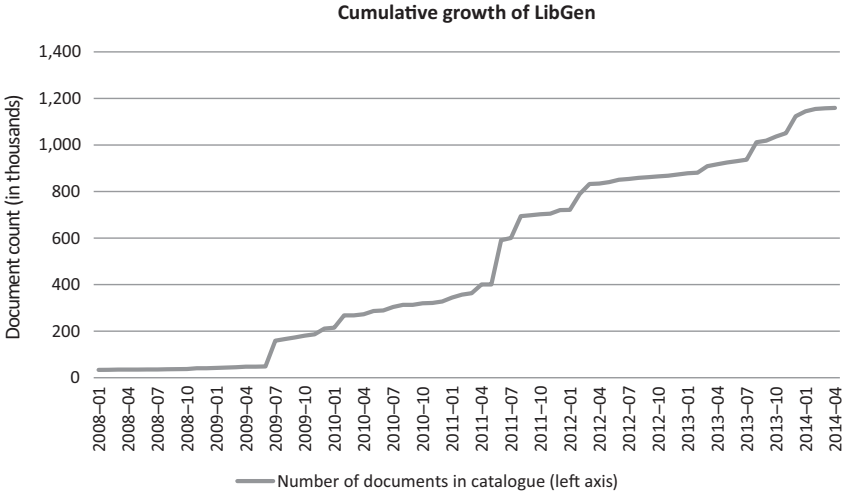

Figure 3.1
The cumulative growth of Library Genesis between January 2008 and April 2014 (full catalog).

Most shadow libraries are thought to be "peer-produced commons" in the sense that they are built from the contributions of many individual users. One example of such peer production is the Gigapedia/library.nu collection, which contained one-half million documents assembled from contributions by thirty major contributors (together responsible for adding a little more than 50 percent of all books), and nearly nine thousand small contributors, who usually uploaded only one or two contributions each. In contrast, LibGen's growth (82 percent of all the records) came from huge, single-day additions of tens of thousands of documents each. These occasions most likely mark the integration of large, preexisting collections into the LibGen collection. Although there are a variety of methods in use in the file sharing community to encourage users to contribute (Bodó 2014), such as social or financial rewards for uploaders, LibGen unlike Gigapedia uses none of these. Individuals can submit documents to the collection, but LibGen does not encourage and definitely does not reward such submissions. Typically, individual submissions add only a few thousand documents each month, accounting for a total of around 18 percent of the collection.

Preexisting Collections

Because the LibGen community is very conscious of its history as an aggregator of collections, data on the provenance of source collections is usually maintained within the database. This allows for a relatively clear picture of the expansion of the collection.

Prior to 2011, Library Genesis was one of several large, predominantly Russian-language archives. It grew through aggressively integrating other, primarily Russian corpuses developed in academic networks in the early and mid-2000s, such as the Kolkhoz collection described in chapter 2. Altogether, LibGen added 330,000 documents in those years. By 2011, however, the preexisting Russian sources were largely exhausted. The corpus of valuable Russian scientific and classic literature was increasingly complete. Then the game changed. Gigapedia/Library.nu began by copying and cataloging English-language texts from the LibGen collection, which it built into a much larger English-language catalog. As publisher-led enforcement pressure on Library.nu grew in 2011, LibGen returned the favor. Between mid-2011 and mid-2012, LibGen integrated nearly half a million new books—by all appearances nearly all from the Gigapedia archive prior to its shutdown. A third wave of growth in 2013 is attributable to the integration of publisher-produced electronic text repositories.

Linguistic and Thematic Expansion of Library Genesis

The integration of the Gigapedia material transformed LibGen from a predominantly Russian, natural sciences-focused collection into a predominantly English-language multidisciplinary shadow library. Since the LibGen records contain document metadata, such as the document language, subject matter, and the date of addition to the archive, it is relatively easy to map how the focus of the collection shifted over time.

Figure 3.2 suggests that the majority of Russian-language documents were added in 2008–2010, whereas around 80 percent of the English language documents arrived in 2011 and after, beginning with the Gigapedia/Library.nu collection in 2011.

The linguistic composition of the database continues to change. German, the third most common language in the collection, representing 8.5 percent of the full catalog, emerged only in 2013, fueled by large, single-day additions of documents from the same publisher. The German additions very likely represent the start of a new trend. As large, peer-produced free-floating text archives are slowly exhausted, and as publisher-developed digital archives grow and become more widely accessible, the major opportunities for expansion will come from the latter. In most cases, such expansion represents a process of leakage, in small and large quantities, from universities and other institutions with legal access to publisher catalogs—a process we see repeatedly in the history of developing-country shadow libraries. Over time, such downloaded collections find their way to LibGen.

Other major languages, such as French, Spanish, and Mandarin are strikingly underrepresented in the collection. Forum discussions on LibGen offer various explanations for the omission of Chinese documents, which on balance appears to be based on

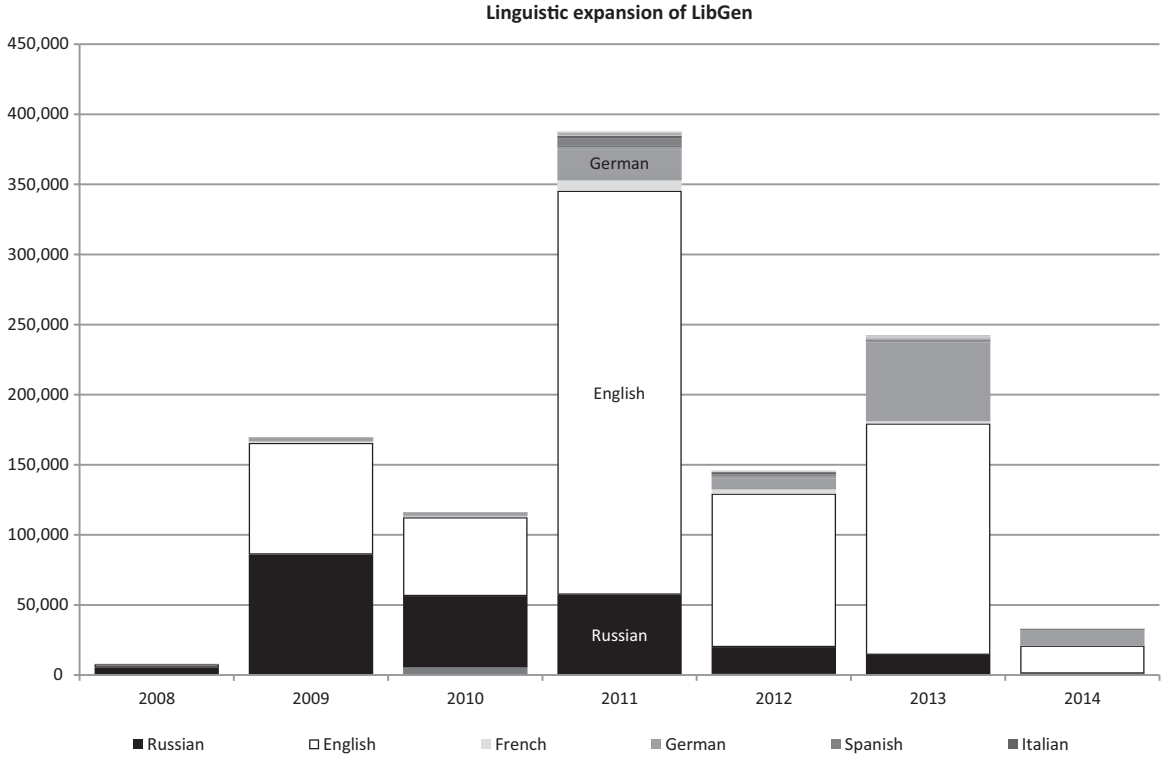

Figure 3.2
Language of documents added to the Library Genesis collection each year (full catalog, document/language > 1,000).

a decision by the LibGen administrators to avoid content that they have no capacity to manage. To date, LibGen has not integrated any of the large Chinese-language shadow libraries available on the web. The lack of scholarship in other major European languages is more puzzling and likely reflects a combination of factors. There appear to be few large, persistent shadow libraries in French or Spanish, and—to the best of our knowledge—fewer for other languages. Where digital collections are available, the social and curatorial networks that underpin the creation of large, online English and German collections do not appear to have developed. To date, LibGen has not become a repository for archive communities in other languages, nor have LibGen administrators sought to significantly expand their linguistic coverage. Such expansion remains opportunistic.

As figure 3.3 suggests, the majority of works in the natural sciences, mathematics, and computer science were added in 2009–2010. The 2011 integration of Gigapedia also substantially changed the thematic focus of the library,[3] with LibGen absorbing the overwhelming majority of works in other disciplines in 2011 or later. Before the Gigapedia material arrived, LibGen was a mostly Russian, natural sciences-focused collection that incorporated the various scientific corpuses developed in Russian universities and scientific institutes. The post-Gigapedia LibGen became a much broader archive with reach into the much larger English-reading public for scholarly work.

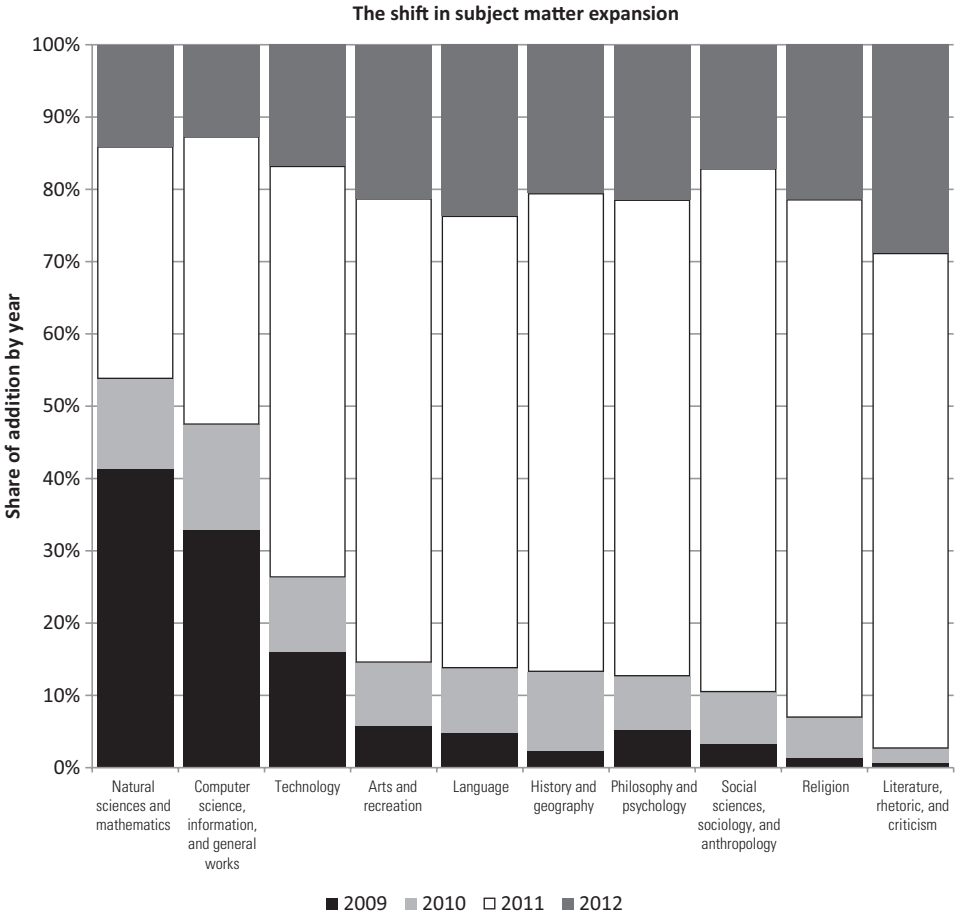

Figure 3.3
Each column represents one top-level Dewey subject category. The shading shows what percentage of all documents was added to the catalog in a given year (identified dataset).

Publishers

More than fifty-five thousand publishers are represented in the LibGen collection, though the exact number is difficult to pinpoint due to both the large number of records without publisher information (in the full dataset: 27 percent; among the texts we've identified: 3.2 percent) and the noise in the existing data. The distribution, as expected, is very concentrated, with the top 100 publishers accounting for somewhere between one-third and one-half of all documents in the catalog (full dataset: 34 percent, identified dataset 50 percent). The top ten publishers' share of the identified catalog and the average downloads per document are visible in table 3.1.

The major Western academic publishers dominate the catalog. Nevertheless, we the catalog also contains thousands of smaller publishers, with just a few titles each, and although there are documents in more than a hundred different languages, the collection predominantly represents the Western, English-language, scholarly mainstream. This focus has an impact on demand, as we will discuss later.

As the last column of table 3.1 shows, publishers with the highest number of works in the catalog are not necessarily the most popular ones. Supply and the demand do not perfectly overlap. The ten most popular publishers in terms of the number of downloads per title (based on publishers with more than a hundred titles in the catalog) account for only less than 0.8 percent of the catalog, but more than 2.2 percent of

Table 3.1
The document share of the top ten publishers in the identified dataset, with average downloads/title figures per publisher (the average downloads/title in the whole identified dataset is 3.1)

Publisher (ISBN based)	Share of catalog	Downloads/title (catalog average: 3.1)
Springer	6%	3
Cambridge University Press	4%	6
Routledge	4%	5
Wiley	3%	4
Oxford University Press	3%	5
Palgrave Macmillan	1%	3
Harper & Row	1%	2
Springer Verlag	1%	6
McGraw-Hill	1%	4
Academic Press	1%	4

all the downloads. These publishers are among the smaller ones with, on average, only 300 works each in LibGen. Most specialize in mathematics and social sciences: Verso (12.58 average downloads per document), The Society for Industrial and Applied Mathematics (10.76), Benjamin/Cummings Pub. Co. (9.81), The Mathematical Association of America (9.76), Попурри (9.70), Polity Press (9.58), John Benjamins Publishing Company (8.74), Blackwell Publishers (8.26), The American Mathematical Society (8.18), and Birkhäuser (7.92).

The same divergence between supply and demand is present in subject matter, as seen in table 3.2. Social sciences are the leading category in the archive, both in terms of volume and demand, representing 15 percent of identified titles, and with slightly higher-than-average downloads per title. Social sciences are followed by technology and engineering texts (14.5 percent), natural sciences and mathematics (9.3 percent), and literature and criticism (8.6 percent). While these latter two categories account for more or less the same share of the catalog, they cannot differ more in terms of demand. Natural science titles on average see almost three times higher demand than literary works.

Drilling down further into the second- and third-level Dewey Decimal Classification (DDC) classes offers a more detailed map of the thematic composition of the collection and the focus of demand. Due to their length, we limit the lists to the ten most frequent classes in tables 3.3 and 3.4.

Table 3.2
Subject matter share and demand in Library Genesis by top-level DDC classes

Top-level DDC classes	Share of titles	Downloads/title
Unclassified	31%	3
Social sciences, sociology, and anthropology	15%	3
Technology	14%	3
Natural sciences and mathematics	9%	5
Literature, rhetoric, and criticism	9%	2
Computer science, information, and general works	6%	3
History and geography	4%	2
Arts and recreation	3%	2
Philosophy and psychology	3%	5
Religion	2%	3
Language	2%	6

Table 3.3
The thematic composition of Library Genesis by second-level DDC classes

Second-level DDC classes	Share of identified dataset	Downloads/title
Medicine and health	6%	2
Computer science, information, and general works	5%	3
American literature in English	4%	1
Economics	4%	3
Mathematics	4%	8
Engineering and allied operations	3%	4
Social sciences, sociology, and anthropology	3%	4
Management and public relations	3%	3
Social problems and social services	2%	2
English and Old English literatures	2%	2

Table 3.4
The thematic composition of Library Genesis by third-level DDC classes

Third-level DDC classes	Share of identified titles	Average downloads/title
American fiction in English	4%	1
Diseases	3%	2
Computer programming, programs, and data	3%	3
General management	2%	3
Applied physics	2%	4
English fiction	1%	2
Special computer methods	1%	3
Data processing and computer science	1%	2
Production	1%	3
Culture and institutions	1%	4

Based on the Dewey subject categories, LibGen has a wide supply of works in American fiction, health, computer science, and natural sciences. It is also apparent that the most populous subsections are not necessarily the most popular ones. The most popular subject matter in terms of average downloads per title are: English grammar (10.29 downloads per title), standard usage and applied linguistics (10.13), analysis (9.87), French philosophy (9.30), algebra (8.67), numerical analysis (8.05), general principles of mathematics (7.99), topology (7.99), probabilities and applied mathematics (7.44), geometry (7.34), German and Austrian philosophy (7.33), modern Western philosophy (7.33), other philosophical systems (7.25), philosophy and theory (7.22), social sciences, sociology and anthropology (7.19), and logic (7.07).

The data indicates pretty clearly that the subjects in highest demand in the LibGen shadow library are books used for learning or working in English, mathematics, and philosophy. English language resources point to the international reach of the archive. As we discussed in chapter 2, mathematics was one of the first disciplines to be extensively digitized and the first discipline to be integrated into LibGen. These parts of the collection were probably more carefully selected and curated by a specialist group than, for example, those that were ingested en masse from publisher e-libraries. LibGen probably also inherited the readers along with the collections, leading to relatively steady demand. Readers of Western philosophy probably arrived later, when the relevant works were integrated from the Gigapedia collection. Whether the high level of interest in Western philosophy is a function of the quality of the collection, of broader awareness of LibGen in these fields, of ethical norms specific to these fields—as one commentator has suggested (Schwitzgebel 2009)—or some combination of the three is a question we must leave open.

The Age of Works in Library Genesis

LibGen also contains information about the date of publication of the documents in its collection, allowing us to make some observations about the age of the collection and the factors that affect it. As seen in figure 3.4, although the collection has a large number of classics, it is heavily skewed toward recent work, which is more likely to have a digital version and thus easier to include than scanning a version by hand.

The Legal Supply of Works in Library Genesis

We measured the legal availability of the titles in the LibGen catalog by collecting data from two additional sources: Amazon.com (in September and October 2013) and

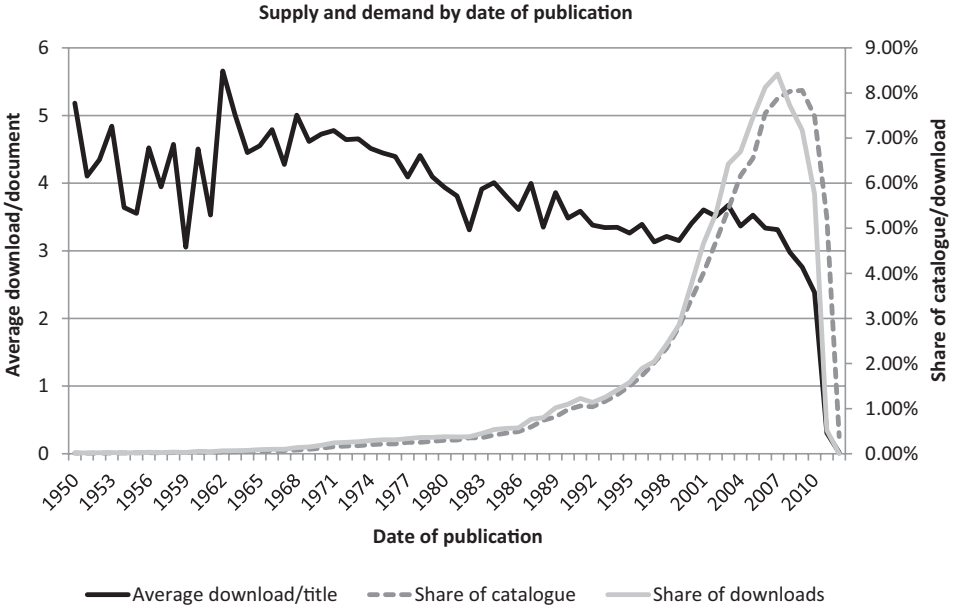

Figure 3.4
The share of catalog and the share of downloads by the date of publication (right axis), and the average download/title/date of publication (identified dataset, average download/title: 3.01).

WorldCat.org (November 2013). We used Amazon.com for data on legal market access, and WorldCat.org for e-library availability. Price information in some categories (such as used book prices or rental prices) should be treated with caution due to their extreme volatility on Amazon.[4] Table 3.5 shows the availability and price information for all the identified documents in all categories.

Based on the Amazon data, it is clear that while print availability is generally high, with nearly 83 percent of titles in LibGen available in some sort of print format (new or used, purchase or rental), there are huge gaps in electronic availability.[5] As figure 3.5 shows, electronic availability figures are dramatically improving for works published more recently. Still, on average, only a third of the identified catalog is available as a Kindle e-book (to buy or rent). E-repository availability seems to be higher, but this result should be treated with caution.[6]

Further analysis reveals that different subject matter has different legal availability rates: Natural sciences and mathematics titles, which form the core of the LibGen collection, have much lower e-book availability rates than literary works, for example.

E-libraries could, in theory, successfully compete with shadow libraries. Institutional subscriptions allow affiliated individuals to access a relatively wide range of titles, at no

Table 3.5
Price and availability information for the identified dataset, based on Amazon.com (prices in USD)

	List price	Sold by Amazon	Sold second hand	Sold as new	Available for print rent	Available on Kindle	Available for Kindle rent	Available through e-libraries
Available	46.74%	53.82%	80.93%	79.07%	1.46%	31.62%	6.26%	64.83%
Not available	53.26%	46.18%	19.07%	20.93%	98.54%	68.38%	93.74%	35.17%
Mean price	87.19	69.43	54.44	62.46	30.73	46.64	28.07	
Median price	57.00	41.40	15.98	29.19	24.94	23.99	14.93	
Mode price	24.95	7.19	0.01	0.01	17.00	9.99	11.61	
25 percentile	28.95	20.95	2.50	11.43	17.18	9.99	10.23	
75 percentile	125.00	89.99	48.24	63.75	37.51	59.99	34.26	

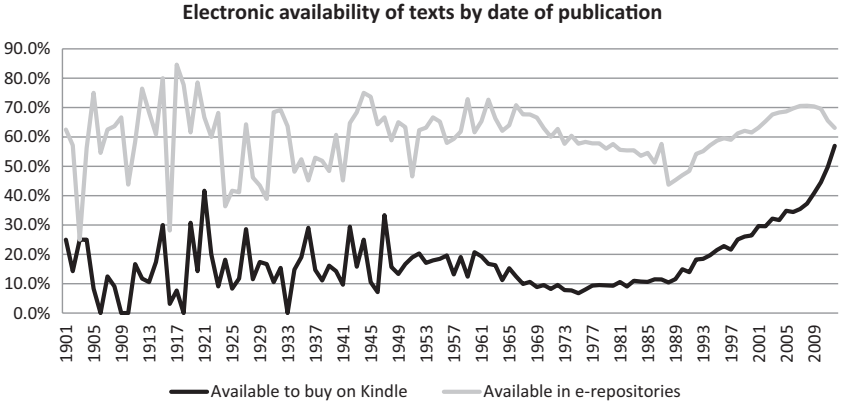

Figure 3.5
Electronic availability of titles in the identified dataset by date of publication.
Note: The sudden drop in shares between 1988 and 1989 can be attributed to political change in the Soviet Union. In 1989, Perestroika was in full swing, resulting in the publication of important long-suppressed works in Russian. Few of these are translated or available in digital formats.

direct cost. In principle, e-library availability is outstanding compared to other forms of electronic access. But actual access to these repositories is sharply limited by a number of factors, beginning with the cost of institutional subscription, and including the necessity of being affiliated with an institutional subscriber, either as faculty or as a student. Basic technical difficulties in accessing e-library catalogs also remain commonplace, making crude but free an effective competitor to even subsidized legal channels.

The analysis of prices suggests that academic publishers tend to price their titles with the library market in mind. A quarter of the titles have a list price over $125, and both the mean and the median prices are well above the $20 to $40 range, which is the usual price for a fiction title. The secondhand and e-book market prices (both targeting individual rather than institutional buyers) are much closer to this price range, suggesting that the primary target for print editions is not the individual buyer and that, accordingly, the effect of pirated copies on sales is not readily measured by conventional estimates of "substitution effects."

The Demand Side

Who uses these shadow libraries? To what extent do they compete with legal sources? There are many theories that link the demand for pirated content to the availability of legal alternatives. Theories of substitution argue that unauthorized file sharing services directly compete with legal alternatives (Dejean 2009; Fink Maskus, and Qian 2010; OECD 2009; Smith and Telang 2012). Other studies find evidence that unauthorized file sharing networks correct the shortcomings of legal markets by providing access to otherwise inaccessible works (Bodó and Lakatos 2012; Bodó 2011; Karaganis 2011). The two accounts are not incompatible, but have tended to be very difficult to reconcile empirically. Markets for media goods are changing rapidly as technologies enable both new forms of intermediation and access (including, in the publishing field, the emergence of a superintermediary in the form of Amazon.com) as well as new practices of consumption (such as bibliophilia freed from the constraints of income and shelf space). The majority of studies from the last decade have focused on disentangling these issues in the music and audiovisual sector. Although there has been some recent work on the unavailability of copyrighted works on legal markets—the so-called "orphan works" problem (Heald 2014; Rosen 2013)—studies of unauthorized downloading in the book market have been few and focused primarily on trade sales (Hardy, Krawczyk, and Tyrowicz 2014; Reimers 2016). Most of the evidence on the effects of piracy in the book industry remain anecdotal (Laskow 2013; Pogue 2013).

Among the academic communities that form the primary audience for the LibGen sites, we are clearly discussing a phenomenon of some global size: on average: 43,500 documents per day were downloaded from B—one of the many mirror sites that incorporate the LibGen catalog—during the three-month period of study in 2012.[7] Positively identified LibGen items were downloaded on average 24,000 times a day—indicating substantial demand for titles from B's large catalog of popular, non-LibGen materials. Since B is only one of the many mirrors of LibGen, overall use within the ecosystem can be assumed to be much higher.[8]

One of the most persistent questions about digital piracy is its impact on legal markets. Demand for pirated materials can compete with legal sales, or it can be driven by market unavailability. If we compare the average download figures for works (un) available in various formats (table 3.6), we can make two claims. First, LibGen clearly plays an archival function in contexts where works are out of print. Although the absolute number of such titles is relatively low, our dataset from B records hundreds of thousands of downloads of such texts. This archival function is almost certainly more pronounced for the nonidentified part of the collection (some 30 percent), which is made up of predominantly harder-to-access, older, non-English works.

Table 3.6
Descriptive statistics of global downloads by legal availability (all means have a statistically significant difference on a 0.05 level)

		Share of titles	Average downloads/title
Available in used copy?	No	19.10%	2.28
	Yes	80.90%	3.29
Available in new copy?	No	20.90%	2.27
	Yes	79.10%	3.32
Available to buy on Kindle?	No	68.40%	3
	Yes	31.60%	3.36
Available to rent on Kindle?	No	93.70%	3.01
	Yes	6.30%	4.51
Available to rent in print?	No	98.50%	3.05
	Yes	1.50%	6.28
Available in e-repositories?	No	35.20%	2.55
	Yes	64.80%	3.4

Yet, in general, as table 3.6 shows, demand on LibGen correlates with legal availability: if a title is legally available in any format, it enjoys higher downloads. The explanation of this correlation is, in our view, unremarkable and somewhat circular: texts are both kept in print by publishers and downloaded via LibGen in function of demand. By the same token, texts are more likely to appear on LibGen when they are in publication in print or digital form. This correlation is consistent but shows some noteworthy variations depending on the nature of the supply channel. The very high per-title demand for titles available as rentals, for example, probably denotes high student demand for textbooks. The high average demand for titles not available on Kindle probably reflects the fact that relatively few scientific books and articles are available in this format. The relatively high demand for titles that are also available through institutional archives suggests the importance of the academic and scientific user community in institutions and countries with little access to paywall services. Through these partial indicators, a picture of the LibGen community begins to emerge.

Library Genesis's administrators stress that they focus on collecting only works that are relevant to the heavily academic community they serve, irrespective of their legal availability. Although large categories of popular work are excluded from these criteria, the definition of relevance clearly piggybacks on the gatekeeping function of publishing itself. What's relevant, broadly speaking, is what's in print. Both the high degree of availability of in-print (if not digitally available) titles and the higher demand for those titles support this general connection. While LibGen certainly has a strong archival function, its main function is to address the lack of access to digital copies, especially outside the communities that have access to large university libraries and publisher e-catalogs.

Demand by Country

This role in expanding access beyond privileged universities is reflected in differences in country-level demand. Table 3.7 contains country-level transaction data for both the B dataset overall and for the identified documents within it.[9]

We will make no strong effort here to disentangle the developmental issues, cultural issues, and other factors that might account for these differences. At a very basic level, B may simply be better known in some national academic communities than in others. But we will venture some observations. We see three broad categories of countries among the largest downloaders.

Table 3.7
Top users of the Library Genesis catalog via the B mirror

	All B downloads			Identified document downloads		
Country	(1) net downloads (without proxy traffic)	(2) share of proxy traffic in country traffic	(3) country share of all net downloads	(4) net downloads (without proxy traffic)	(5) share of proxy traffic in country traffic	(6) country share of all net downloads
Russia	861 865	1%	31%	168 863	1%	12.8%
Indonesia	175 234	2%	6%	135 961	2%	10.3%
United States	222 373	5%	8%	133 827	4%	10.2%
India	129 679	6%	5%	86 817	6%	6.6%
Iran	96 836	1%	3%	67 084	1%	5.1%
Egypt	96 302	0%	3%	55 468	0%	4.2%
China	77 065	0%	3%	55 458	0%	4.2%
Germany	96 618	35%	3%	54 516	33%	4.1%
United Kingdom	61 772	10%	2%	41 065	6%	3.1%
Ukraine	135 726	2%	5%	32 246	2%	2.5%
Turkey	42 637	0%	2%	31 836	0%	2.4%
France	56 131	13%	2%	31 720	10%	2.4%
Poland	48 525	0%	2%	27 925	1%	2.1%
Italy	41 659	0%	2%	26 550	0%	2.0%
Canada	34 393	5%	1%	21 400	3%	1.6%
Spain	30 874	2%	1%	19 691	1%	1.5%
Sweden	35 117	5%	1%	18 229	5%	1.4%
Romania	26 419	3%	1%	18 159	2%	1.4%
Greece	25 161	8%	1%	17 791	5%	1.4%
Netherlands	29 405	45%	1%	16 306	42%	1.2%
Australia	19 988	1%	1%	12 002	1%	0.9%
Algeria	17 747	0%	1%	11 772	0%	0.9%
Hungary	13 988	0%	1%	10 072	0%	0.8%
Czech Republic	17 762	39%	1%	9 431	36%	0.7%

First, Russia and other post-Soviet countries are, predictably, heavy traffic sources, with significantly more downloading of Russian-language content than of material from the rest of the collection.

Second, developing countries such as Indonesia, India, and Iran are also major traffic sources. These countries have in common relatively low per-capita GDP, underdeveloped electronic text markets, and rapidly growing student populations—all factors that we would associate with high shadow library use.

Third, developed countries such as the United States, Germany, and the UK are also represented at or near the top, and require a somewhat different explanation. All of these countries have highly developed print markets, comparatively well-developed electronic book markets, dense and accessible library systems, and otherwise good infrastructures for higher education, science, and research. Nevertheless, for many categories of both scholarly works and users, similar barriers of price and availability come into play: legal access to scholarly works in digital formats is still generally poor and pricing (for any format) is often set at levels that target libraries rather than individual buyers. From the perspective of students, the conflict between personal library building and economic constraints are particularly sharp. As we see elsewhere in this report (and parallel to developments in music downloading), collecting is a powerful motivation in and of itself, and in the downloading era has become increasingly divorced from intentions to read or consume.

A somewhat different global picture emerges if we adjust these results for population size, and only account for the identified documents (see table 3.8). The top of this list is dominated by small, relatively poor countries at the edges of the European Union. All have highly educated populations, dense cultural, political, and economic ties with the West, and—in the case of the Eastern European countries and crisis-ravaged Greece—diminished resources and educational infrastructure compared to the core European countries. Most, moreover, are under obligations to implement EU educational standards established by the 1999 Bologna Accords, which promote compatibility with Western European and North American degrees (Keeling 2006; Reinalda and Kulesza-Mietkowski 2005). The effort to establish such degree and accreditation systems, in turn, has required the rapid transformation of the *content* of education within these systems, ranging from the curricula, to the acquisition policies of university libraries, to the corpuses of knowledge that faculty and students need to be competitive in Western-centric disciplines.[10] Given the limited financial (and sometimes also human) resources available for such transitions, many libraries cannot meet faculty and student demand. Such contexts provide fertile ground for shadow libraries like Library Genesis.[11]

Table 3.8
Document downloads per 1,000 inhabitants (without proxies)

Country	All B downloads per 1,000 persons	Identified document downloads per 1,000 persons
Lithuania	5.5	2.9
Estonia	4.2	2.3
Sweden	3.7	1.9
Greece	2.2	1.6
Barbados	2.9	1.5
Latvia	3.4	1.5
Slovenia	2.2	1.5
Iceland	2.6	1.3
Luxembourg	2.5	1.3
Croatia	1.6	1.2
Russian Federation	6.0	1.2
Macedonia, Fyr	1.5	1.1
Hungary	1.4	1.0
Bulgaria	1.8	1.0
Netherlands	1.8	1.0
Israel	2.1	0.9
Armenia	2.0	0.9
Czech Republic	1.7	0.9
Iran	1.3	0.9
Romania	1.2	0.8
Montenegro	1.1	0.8
Cyprus	1.4	0.8
Malta	1.1	0.8
Finland	1.4	0.8
Poland	1.3	0.7
Portugal	1.0	0.7
Ukraine	3.0	0.7
Egypt	1.2	0.7
Moldova	1.8	0.7
Germany	1.2	0.7
Albania	0.8	0.7
United Kingdom	1.0	0.7

Country-Level Knowledge Diets

Finally, and more speculatively, we can look at the distribution of top-level Dewey subject headings in country-level downloading—a step that allows us to develop a rough sense of the "knowledge diet" of LibGen users in different countries.[12] This analysis revealed three major clusters (mapped to geography in figure 3.6), representing significantly different consumption patterns of subject matter (mapped to subject matter in figure 3.7).

The simple clustering approach had some surprising results. It clearly identified the post-Soviet republics as one group (cluster 3). These countries are differentiated by their large share of unidentified documents in their diet. As we have indicated earlier, the documents we were not able to identify via ISBN-based WorldCat services tend to be older, Russian-language titles. LibGen's Russian collection is actively used by countries that share a common Soviet past.

We also find significant differences among the rest of the countries in the analysis. The clustering algorithm identified two relatively homogenous groups. Countries belonging to cluster 1 have higher levels of social sciences, literature, history, and philosophy, and lower levels of natural sciences and technology in their overall consumption than countries that belong to cluster 2.

There might be many reasons why a country would prefer downloading social sciences to downloading hard sciences, or the other way around. One such explanation is

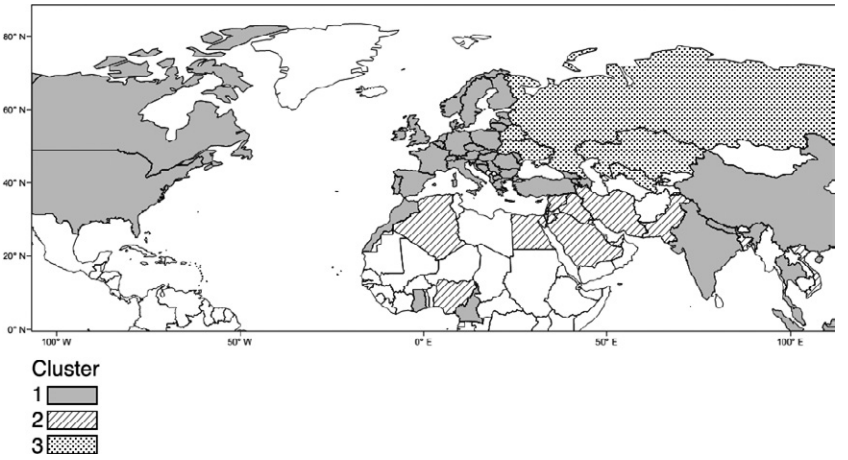

Figure 3.6
Country clusters based on their "knowledge diet." Shading corresponds to that used in figure 3.7.

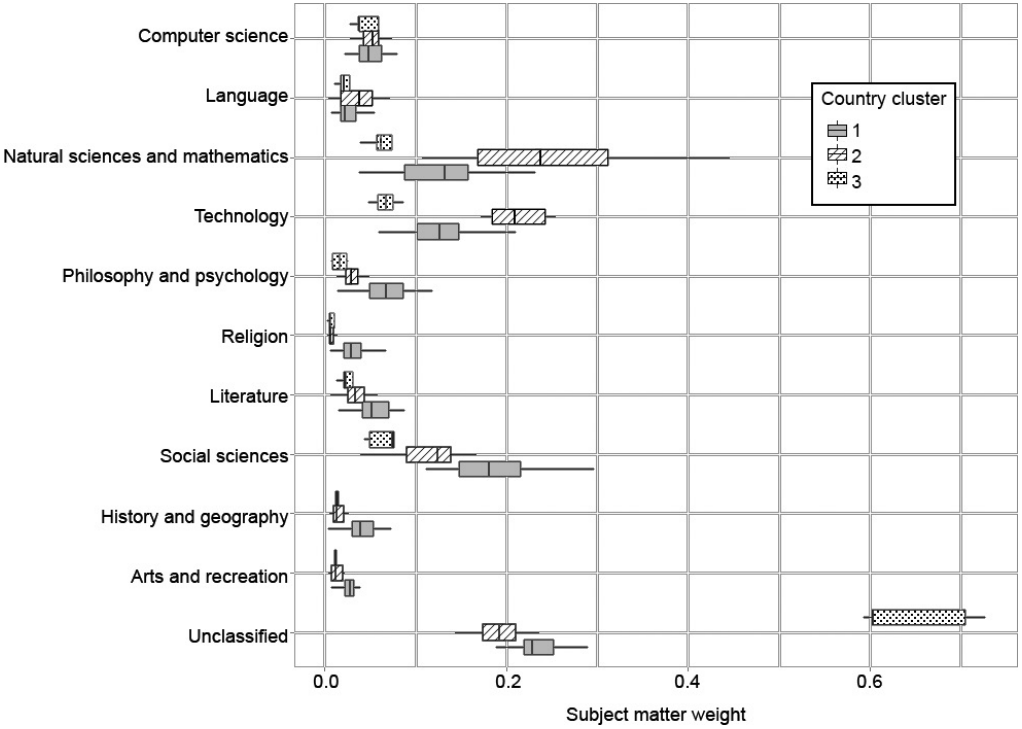

Figure 3.7
The mean weight (and quartiles) of each subject matter in the science diet of the three country clusters. Shading corresponds to that used in figure 3.6.

institutional: based on UNESCO and OECD data, the share of social science, business, and law graduates in the cluster 1 is nearly twice the share of social science graduates in cluster 2 (7.85 percent vs. 4.58 percent), while the difference in the share of science graduates is significantly smaller (OECD 2016; UNESCO 2016). But the existence of the two clusters may also reflect some inherent internal characteristics of the two types of scientific discourse. On the one hand, the hard sciences of use and interpretation relies on the lingua franca of logics and mathematics, which are the least determined by the cultural context in which such interpretation takes place. The social sciences, literature, history, philosophy, and psychology sections of LibGen, on the other hand, are made up of the mainstream of Western thought, and they strongly reflect the conditions that produced that corpus of knowledge. This corpus, contrary to hard sciences, does not in itself constitute a universal interpretative frame, nor can it rely on one. The culturally strongly situated Western social science corpus may not enjoy the frictionless

diffusion that hard sciences rely on due to the existence of the universal language of mathematics.

Most of the Latin American and African countries are conspicuously missing from this chart. We could not find any data-related explanation for this phenomenon, so we have to assume other factors explain the dearth of users, such as the lack of substantive Spanish and Portuguese collections in LibGen, and the Russian, Eastern European, and respective ex-pat social networks through which LibGen and similar sites operate.

Because these libraries are frequently penalized by or excluded from search engine indexes, these networks depend heavily on dedicated online discussions and word of mouth. The strong presence of Russian-speaking users may be a self-limiting factor in this regard—as visible to Russian-speaking users as it is invisible to Latin American ones. Other factors, such as differences in Internet penetration and the nature of other informal distribution channels (such as photocopying) almost certainly play a role as well.

Conclusion

In key respects, Library Genesis is the product of social, cultural, and historical circumstances specific to post-Soviet Russia. These circumstances initially gave rise to a shadow library that catered primarily to Russian materials and users—one of many such libraries that digitized and collected books and made them freely available in the 1990s and early 2000s in Russia. By 2014, LibGen was the leading shadow library in both Russian and English for Western science, complemented by a sizable collection in German and smaller collections in other languages.

Eventually, most such libraries must either limit their growth, reach, and relevance, or accept a higher profile and increased risk of prosecution. The current shadow library landscape has many small, specialized collections that operate largely under the radar of the major copyright enforcement efforts in publishing.

Library Genesis is an exception in that it is both big and, to date, enduring. The generally permissive legal environment in the early and mid-2000s in Russia provides some of the explanation for this persistence. And despite signs of stronger enforcement, this is probably still a factor: the limited reach of Internet enforcement into not just Russia but also the array of post-Soviet states linked by Russian-language social and academic networks still provides a wider margin for gray and illegal services than the core European countries. The social norms and legal disarray that shaped LibGen's open policies may or may not ensure its long-term survival, but as a manifestation of broader social pressures they are almost certain to ensure its reproduction.

LibGen's open approach suggests affinities with bottom-up collaborative projects that rely on many small contributors—the classic conception of the peer-produced commons. In reality, and like many of these projects, LibGen neither pursues nor realizes this vision. The impressive growth of the collection is driven by the efforts of a very small community that seeks out and integrates other digital collections en masse, whether derived from long-term scanning and collecting by other academic and quasi-academic communities or, more recently, from the large-scale copying of publisher catalogs.

B and other mirror sites of the LibGen network have developed substantial side interests that extend beyond the original LibGen collection (of which B's massive collection of literary works is perhaps the most prominent example). LibGen itself does not purport to be a universal library—rather, it is strongly grounded in a conception of quality and relevance to academic disciplines, which in turn maps closely to the gatekeeping role of the major publishers. Accordingly, LibGen is made up of mostly in-print but undigitized works.

Given the rapid pace of digitization and the porous borders of the academic community in the United States and Europe, continued leakage of publisher catalogs into shadow libraries is a virtual certainty. Furthermore, given the expansion of Internet access and markets for cheap readers into large parts of the developing world (and the comparatively slow pace of expanded site licensing of publisher databases), we should expect continued high demand for these works at the peripheries of these university and publisher ecosystems. The role that these services play will continue to depend on a balance of forces between legal market development, the viability of highly organized libraries like Library Genesis in the face of stronger enforcement, and the back-up plan when both fail: the "sneaker net" of portable media libraries and small-group student and faculty exchanges. Inside the United States and EU, where most of the academically "relevant" work is at least in print, there is still considerable scope for the improvement of digital catalogs, expanded site licensing, and open access models, which can undercut the main functions of the pirate library.

Notes

1. The analysis is based on multiple datasets from different sources. The LibGen catalog is freely accessible through its website, including many bibliographic and file-related metadata. We refer to this dataset in the subsequent analysis as the full catalog. This data was subsequently enriched with metadata from the WorldCat database, and market accessibility data (prices and formats) by collecting additional data from Amazon.com. In the analysis, this enriched, positively identified, de-duplicated subset of the catalog from the second half of 2012 is referred to as the *identified*

dataset. The demand-side analysis is based on web server logs acquired from the administrators of B, a commercial LibGen mirror. It contains author, title, and partially redacted IP address information from between March 2, 2012 and May 27, 2012. The cleansed log data is referred to as *all B downloads*, while the dataset in which the log records were linked to the identified dataset is referred to as *all identified downloads*. Data on the catalog of Gigapedia/Library.nu come from late 2011. The author would like to thank the Online Computer Library Center (OCLC) for providing access to the WorldCat services, and the B administrators for sharing the web server logs.

2. For the sake of comparison, at the time of the last review of this chapter, in February 2016, LibGen contained more than 1.6 million records.

3. We limited the analysis to the identified dataset, since Dewey subject categories are not reliably present in the full LibGen catalog.

4. Availability is subject to strong seasonal fluctuations as semesters start and end, while price, especially for the used book market, is subject to strong, often software-aided competition among different sellers, resulting in constant adjustments and discounting.

5. Because we rely here only on the U.S.-based Amazon, the actual availability rates in all categories (new, secondhand, e-book) are probably overestimated. Though Amazon ships new books globally, it is used only infrequently by consumers outside of the United States—primarily in contexts where the title is not available through local retailers. For our purposes, we assume that if a title is not available on Amazon, it is less likely to be available via other, local channels, especially for English-language titles. For other formats, such as Kindle, book rental, and used books, Amazon has an even more limited global reach. E-book distribution rights are regional: even if there is a Kindle version in the Amazon store, it may not be available beyond U.S. borders. Many second-hand book dealers who offer used books do not ship outside the United States, and textbook rental (both electronic and print) is certainly unavailable for most markets. In these cases, Amazon-based accessibility data represents the best-case scenario, and almost certainly overestimates the actual availability of titles in most local markets.

To further explore these estimation errors, we compared the harvested data with a dataset provided to us by a prominent academic publisher with a significant number of publications in the LibGen collection. We harvested list prices with near perfection. E-book availability was correctly identified in 69 percent of the cases, while the share of false positives and negatives was around 15 percent for each. Since publisher-provided e-book availability data does not perfectly coincide with the date of data collection from Amazon, and includes other e-book providers besides Amazon, we concluded that the Amazon-gathered data adequately represents the actual facts on the ground in terms of theoretical availability, but overestimates actual availability in local, non-U.S. markets.

6. E-libraries are electronic text collections available through university libraries or publisher portals, such as Oxford Scholarship Online. The 60 percent number requires some methodological caveats. E-repository availability is based on WorldCat library records, which may note if a book has an electronic document version. On manual inspection of the records, it turned out that many of the links to electronic versions point to a limited preview Google Books entry, or a

"table of contents" published as a PDF on the publisher's website. The comparison of our collected data with a publisher-provided dataset showed that we falsely assumed the existence of an e-repository copy in 14 percent of the titles. As a result, e-repository availability in fact may be much lower than indicated.

7. The demand-side analysis of LibGen is based on a log file we acquired from the administrators of the B mirror. The log contains 7.990.130 records from between March 2, 2012 and May 25, 2012. The records contain a document identifier unique to B, the title and author information as well as the partially redacted IP address of the downloader. We discarded log records that could not be positively and unambiguously associated with an LibGen catalog entry. We successfully mapped 54 percent of the identified dataset to the cleaned B log, accounting for 1.399.278 (47 percent) of the transactions.

After cleaning the dataset from bot traffic, we identified the countries and ISPs associated with the IP numbers, we marked those records that could be associated with known proxies (such as Tor and VPN exit nodes, Opera mini proxies), and anonymized the dataset by discarding the IP addresses. We then matched the author and title information with the appropriate fields in the LibGen catalog.

Excluded log entries are either Russian-language scientific books/periodicals (without or with more than one corresponding item in the identified dataset) or Russian- and English-language nonscientific material (such as song lyrics, comics, and literary works) included in the B database, but not included in the LibGen scientific catalog.

8. Each successful LibGen search lists LibGen as well as the official LibGen mirrors as download options. Since the download links that point to B are second behind LibGen's own (but superior in download speeds), we are safe to assume that the analysis based on the B logs correctly represents the structure of demand, and seriously underestimates its size. We don't have up-to-date usage numbers from LibGen, but forum discussions suggest that in June 2013, a year after our observation period, LibGen registered 40,000 daily users and 1,230,000 page views.

9. The first thing to note about table 3.7 is that a substantial share of traffic for certain countries comes from *proxy relayed traffic*—i.e., the use of Tor exit nodes or other VPN services to disguise the user's IP address. Luxembourg (44 percent), the Netherlands (42 percent), Denmark (41 percent), Germany (33 percent), and Switzerland (29 percent) all have high shares of proxy traffic, due to the many Tor exit nodes located at local ISPs. Iceland (86 percent proxy traffic) is a special case, as the traffic of the mobile version of the Opera browser flows through proxy servers with Icelandic IP addresses. For our purposes, we have subtracted proxy traffic from country traffic, since a request made via a Germany-based Tor exit node, or an Iceland-based Opera mini proxy most probably does not originate in those countries. Overall, 6 percent of the traffic comes through known proxies. This finding fits in the more general trend of pirate traffic being increasingly conducted through VPNs and other privacy-enhancing technologies (Bodó 2015).

10. See, for example, Abramitzky and Sin 2014 on how these demands play out in relation to legal publishing.

11. The place of Sweden probably requires a different explanation. One obvious factor might be Sweden's pioneering role in file sharing, grounded in the creation of services like The Pirate Bay

in the early 2000s and in wider norms that made file sharing the basis of an actual political movement (The Pirate Party). The other reason for Sweden's high rank might be that the actual share of proxy traffic is higher than what we were able to detect. In large part because of the prominence of file sharing, Sweden is a market leader in VPN adoption, and non-Swedish traffic may inflate the Swedish numbers to a considerable extent.

12. We used hierarchical clustering to check whether there are significant differences between the diffusion of different subject matter. For the process, we only included countries with more than a thousand nonproxy downloads from the identified subset of the catalog.

References

Abramitzky, R., and I. Sin. 2014. "Book Translations as Idea Flows: The Effects of the Collapse of Communism on the Diffusion of Knowledge." NBER Working Paper No. w20023. http://papers.ssrn.com/abstract=2421123 (accessed August 18, 2017).

Bodó, B. 2011. "Coda: A Short History of Book Piracy." In *Media Piracy in Emerging Economies*, ed. J. Karaganis, 399–413. New York: Social Science Research Council.

Bodó, B. 2015. "Piracy versus Privacy: An Analysis of Values Encoded in the PirateBrowser." *International Journal of Communication* 9:818–838.

Bodó, B. 2014. "Set the Fox to Watch the Geese: Voluntary IP Regimes in Piratical File-sharing Communities. In *Piracy: Leakages from Modernity*, ed. M. Fredriksson and J. Arvanitakis, 241–264. Sacramento, CA: Litwin Books.

Bodó, B., and Z. Lakatos. 2012. "P2P and Cinematographic Movie Distribution in Hungary." *International Journal of Communication* 6:413–445.

Dejean, S. 2009. "What Can We Learn from Empirical Studies About Piracy?" *CESifo Economic Studies* 55 (2): 326–352. http://ssrn.com/paper=1219442 (accessed August 18, 2017).

Fink, C., K. Maskus, and Y. Qian. 2010. *The Economic Effects of Counterfeiting and Piracy: A Literature Review*. *Advisory Committee on Enforcement*. Geneva: WIPO.

Hardy, W., M. Krawczyk, and J. Tyrowicz. 2014. "Internet Piracy and Book Sales: A Field Experiment." Faculty of Economic Sciences, University of Warsaw Working Papers, no. 23.

Heald, P. J. 2014. "How Copyright Keeps Works Disappeared." *Journal of Empirical Legal Studies* 11 (4): 829–866.

Karaganis, J. 2011. *Media Piracy in Emerging Economies*. New York: Social Science Research Council.

Keeling, R. 2006. "The Bologna Process and the Lisbon Research Agenda: The European Commission's Expanding Role in Higher Education Discourse." *European Journal of Education* 41 (2): 203–223.

Laskow, S. 2013. "Book 'em: Piracy.lab Is Gathering Data on Digital Book Sharing." October 2. *Columbia Journalism Review*. http://www.cjr.org/cloud_control/piracylab.php?page=all (accessed June 30, 2014).

OECD. 2009. *Piracy of Digital Content*. Paris: OECD Publishing.

OECD. 2016. "Education Database: Enrollment by Field." OECD Education Statistics (database). doi:10.1787/33c390e6-en (accessed January 25, 2016).

Pogue, D. 2013. "The E-Book Piracy Debate, Revisited." *Pogue's Posts,* May 9. http://pogue.blogs.nytimes.com/2013/05/09/the-e-book-piracy-debate-revisited/ (accessed June 30, 2014).

Reimers, I. 2016. "Can Private Copyright Protection Be Effective? Evidence from Book Publishing." *Journal of Law and Economics* 59: 411–440. doi:10.1086/687521.

Reinalda, B., and E. Kulesza-Mietkowski. 2005. *The Bologna Process: Harmonizing Europe's Higher Education*. Farmington Hills, MI: Barbara Budrich Publishers.

Rosen, R. J. 2013. "The Hole in Our Collective Memory: How Copyright Made Mid-Century Books Vanish." *The Atlantic*, July 30. http://www.theatlantic.com/technology/archive/2013/07/the-hole-in-our-collective-memory-how-copyright-made-mid-century-books-vanish/278209/ (accessed August 18, 2017).

Schwitzgebel, E. 2009. "Do Ethicists Steal More Books?" *Philosophical Psychology* 22 (6): 711–725.

Smith, M. D., and R. Telang. 2012. "Assessing the Academic Literature Regarding the Impact of Media Piracy on Sales." SSRN. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2132153 (accessed August 18, 2017).

UNESCO. 2016. "Distribution of Tertiary Graduates by Field of Study." UNESCO Institute for Statistics (database). http://data.uis.unesco.org/index.aspx?queryid=163 (accessed January 25, 2016).