DHQ: Digital Humanities Quarterly
2022
Volume 16 Number 2

Lessons from the Library: Extreme Minimalist Scaling at Pirate Ebook Platforms

Martin Paul Eve <martin_dot_eve_at_bbk_dot_ac_dot_uk>, Birkbeck College, University of London
https://orcid.org/0000-0002-5589-8511

Abstract

At 33TB of data in its main collection, the highly illegal Library Genesis project is one of the largest repositories of copyright-violating educational ebooks ever created. Established over a decade ago in 2008, the goal of Library Genesis is nothing short of a modern Library of Alexandria, albeit without anyone's legal sanction. As one of its administrators wrote: "within decades, generations of people everywhere in the world will grow up with access to the best scientific texts of all time. [...] [T]he quality and accessibility of education to the poor will grow dramatically too. Frankly, I see this as the only way to naturally improve mankind: we need to make all the information available to them at any time" [Bodó 2018b]. Rooted in its homeland's Russian communist principles and particularly the Soviet isolationist copyright policies of the twentieth century, Library Genesis is a formidable resource and threat to conventional academic publishers.

The Library Genesis database had just short of 1.2m records (books) in 2014 [Bodó 2018a]. As of January 2020, this capacity has doubled to 2.5m books. In this article, I examine the minimal computational design choices taken by this maximal-in-intent, illicit archive of epistemological dissent and how such decisions have shaped the scalability and growth of the platform. This includes Library Genesis's numerical subdivision of record identifiers into "buckets" to work around directory file limitations in the GNU/Linux operating system; its use of md5 hashing of filenames within directories capped at 1,000 files to avoid future hashing collisions while allowing for on-disk integrity checking; and its use of the MySQL socket/network server as opposed to SQLite or similar disk-based database.

Beyond these computational details, though, the theoretical tension that this article highlights is the path dependencies that are set in (illegal) computational projects that have goals of absolute abundance and maximalist capacity, and the minimalist design principles that they must instigate at the outset to ensure a degree of scalability. I also query the ways in which the project's contested mission statements target an economic (geographic) audience demographic with only minimalist access to high-capacity computing resources. I finally examine the limits on scalability of the distribution of the Library Genesis through its torrent archive and other distributed networking technologies such as IPFS, which despite their promise of peer-to-peer redundancy fall down on an archive of this size.

Minimalist computing principles in scholarly communications focus on the moderation of digital resource consumption and global social equity [Sayers 2016].[1] Such an application of minimalist principles to computing is designed both to make the global flow of digital information accessible *to* the widest audiences and to allow the broadest set of people to participate in the creation *of* such resources. By minimizing the resource constraints required to run computational architectures, minimalist computing principles work to ensure that the world is not falsely separated into a Global "expert" North and a Global "apprentice" South, in which those with resources merely export digital expertise outwards from a claimed center to the supposed margins [Gil and Ortega 2016, 29, 23]. Minimalist computational principles are designed for inclusivity. 1

By contrast, contemporary models of scholarly communication — the composite systems through which academics and other intellectuals disseminate their findings for global consumption — appear to be the opposite of minimalist. They cover a full and ever-expanding disciplinary range, from natural scientific outputs, to humanistic arguments, to social scientific research. Their social processes and techniques are vast, involving peer review systems, citations, references, and footnotes [Grafton 1999] [Fyfe 2015] [Moxham and Fyfe 2018]. The technical elements of such systems also tend to proliferate, including platforms, persistent identifiers, XML standards, digital preservation systems, and manuscript management technologies [Gray 2020] [Andrews 2020]. 2

3

The costs involved in these dissemination systems are also maximal. Since the 1980s, rising in line with the expansion of higher education, the cost of subscribing to academic serials has outpaced inflation by approximately 6% every year [Bosch and Henderson 2018], while library budgets have faced cuts or remained flat [Jurchen 2020, 161]. Over forty years, this amounts to a several-hundredfold price increase. A discourse of maximalism and growth seems omnipresent in scholarly communications.

4 These maximalist principles in scholarly communications have resulted in well-known damaging consequences and inequalities. For instance, Thomas Mboa Nkoudou notes the spread of an "epistemic alienation" when researchers at the "margins" of the Global North's publication systems must conform to external pressures and norms [Mboa Nkoudou 2020]. As another example, peer review is supposed to be a system that distinguishes work based solely on merit rather than any identity characteristic of the author. However, critics of this system note that pre-publication peer review favors those for whom English is their first language [Moore et al. 2017] [Eve 2021]. Further evidence of biases from peer review includes the extraordinary fact that just 1.5% of economics articles in highly ranked journals were about countries other than the United States [Das et al. 2013, 112] [Roh et al. 2020, 43].

5 The spiraling, maximalist costs of the for-profit subscription scholarly publication ecosystem also price out much of the world's population [Kapczynski and Krikorian 2010] [Andročec 2017] [Boudry et al. 2019]. It is in part this inequality that has spurred the open access movement (OA) for research, which seeks to abolish paywalls [Chan et al. 2002] [Berlin Declaration 2003] [Suber et al. 2003] [Fitzpatrick 2011] [Suber 2012] [Eve 2014]. Yet even this movement has come under fire. Built on the premise that nobody, worldwide, should be excluded from access to the scholarly record by their (in)ability to pay and that the benefits of re-use should be broadly extended to third parties using Creative Commons licenses, the road to OA was paved with noble intentions. As Ulrich Herb notes, open access was originally "embedded in a conceptual ensemble of participation, democratisation, digital commons and equality" [Herb 2018, 69]. However, parts of this vision have died over time. "Nowadays," Herb writes, "Open access seems to be exclusive: to the extent that commercial players have discovered it as a business model and article fees have become a defining feature of gold open access, open access has increasingly transformed into a distinguishing feature and an exclusive element" [Herb 2018, 69].

6 Thus, the contemporary scholarly communications environment has many defects that pertain to inequality and that intersect with its maximalist characteristics. Given the goals of minimalist computing, rooted in equity and diversity, might it be possible for our scholarly communications to learn from such principles? Is there a way in which ideas of "scaling small" [Adema and Moore 2021] might help these systems to contend, as Roopika Risam puts it, "Not only with the colonial hangovers from the cultural record, but also with forces that are actively constructing the medium of the digital cultural record — the Internet — as a hostile environment" [Risam 2018, 6]?

7 In the remainder of this article I turn to a specific illicit scholarly communications practice that has emerged in recent years: the idea of the "pirate" shadow library. These archives, which violate copyright, work around paywalls and provide access to all comers. Although frequently on the wrong side of the law, shadow library operators conceive of their sites in ethical terms. Framing their banditry in terms of a Robin Hood-esque outfit, such sites believe they are robbing from the rich to give to the scholarly poor [Faust 2016] (see also [Hobsbawm 1981]). Importantly for the themes of this special issue, I believe that such archives can also be understood in certain terms of computational minimalism. As I will go on to detail, various technical design principles of these archives lower social barriers to participation. This article thus sets out a fresh theoretical terrain for understanding what I call "minimal-maximal tensions" in computational architectures and projects. While this discourse is rooted in older information technology debates about microkernels versus macrokernels [Tanenbaum and Torvalds 2008], through my case study of Library Genesis — the shadow library that is the main focus of this article — I seek to unpack a new framework for thinking about the relationships between minimal components and maximal outcomes. While Library Genesis seeks to "shadow" an enormous archive, the constraints under which it operates lead to a series of unexpected minimalist design principles.

The Growth and Emergence of Shadow Libraries

8 Shadow libraries — pirate archives of copyrighted scholarly publications — emerged in response to a frustration at the slow growth of open access [Green 2017] [Brembs 2020]. Nonetheless, open-access advocates remain divided as to whether these "guerilla" libraries are a solution to, or merely a symptom of, the ills of scholarly communication [Swartz 2015] [Hockenberry 2013] [Faust 2016] [Machin-Mastromatteo et al. 2016]. The two most famous (and interlinked) of these systems for illicit access to pirate scholarship as of 2022 are Sci-Hub and Library Genesis.

9 The former, Sci-Hub, is a credential-proxying site that bypasses publisher paywalls, primarily for journal articles. Founded by Alexandra Elbakyan in 2011, the site has grown to provide access to at least 68.9% of the 81.6 million scholarly articles registered with Crossref and to 85.1% of all articles published in subscription journals [Himmelstein et

al. 2018]. The site is based in Kazakhstan, which introduces complex jurisdictional legal issues. It works by collecting credentials from academic institutions, possibly by conducting phishing attacks [Russell and Sanchez 2017]. These credentials are then used to fetch any article requested by an end-user. In order to reduce the number of requests to publishers' sites — and thereby evade detection and a ban on its credentials — Sci-Hub caches fetched journal articles in an archive called Library Genesis. Sci-Hub has faced and lost several court cases, particularly in the U.S. where the publisher Elsevier has been awarded millions of dollars in damages [Schiermeier 2017]. Because the site sits outside of U.S. jurisdiction and because Elbakyan has no means of paying such damages, it is unlikely that publishers will see any financial return from these lawsuits. Nonetheless, for much of 2021, Elbakyan paused the ingestion of new articles while awaiting the results of a court case in India, which could rule in Sci-Hub's favor [Reddy et al. 2021].

10 Library Genesis, by contrast, is the largest and oldest shadow archive on the Internet. With over thirty-three terabytes of data in its primary book collection (and more than sixty terabytes in its pool of scientific journal articles powered by its aforementioned sister project, Sci-Hub), the project is one of the largest repositories of copyright-violating educational ebooks ever created [Bodó 2020c]. Established in 2008, the goal of Library Genesis is nothing short of a totalizing modern Library of Alexandria, albeit without legal sanction. As one of its administrators wrote, emphasizing its extralegal, yet claimed moral and ethical, status: "Within decades, generations of people everywhere in the world will grow up with access to the best scientific texts of all time.... [T]he quality and accessibility of education to the poor will grow dramatically too. Frankly, I see this as the only way to naturally improve mankind: we need to make all the information available to them at any time" [Bodó 2018b, 25].[2] Philosophically rooted in the communist principles of its homeland, Russia, and particularly in the Soviet isolationist copyright policies of the 20th century [Bodó 2018b], Library Genesis is a formidable resource and a large-scale threat to conventional academic publishers [Green 2017].

11 Shadow libraries differ in their models, usage patterns, and effects. The Z-library system, for instance, charges for specific formats, such as Amazon Kindle conversion [Dulong de Rosnay 2021]. Other pirate libraries such as Library Genesis have no charges at all and their funding mechanisms are unknown, leading to accusations of state subterfuge, although they do use advertising [Harris and Barrett 2019]. While the total percentage of the market eaten by piracy is unknown, in the general ebook space some studies have shown that as many as 35.1% of books are downloaded illegally [Camarero et al. 2014]. What is clear from existing studies of Library Genesis is that it is used by participants worldwide, including in wealthier regions of the Global North [Bohannon 2016] [Till et al. 2019]. Various studies have also shown a citation advantage to papers that appear in Sci-Hub [Correa et al. 2021]. However, considering the prevalence of material in this archive, this effect could simply be due to the difficulty of obtaining the original papers.

12 Given that Sci-Hub and Library Genesis are *shadows* of the formalised academic publication system, we might expect them to share the maximalist tendencies of mainstream scholarly communications systems. These libraries are, indeed, enormous. Yet what it means to be a "shadow" has changed over time. It was not until the 18th century that the shadow became defined as the colorless inverse image of the object itself. Before this point, shadows were represented, in art and heraldry, as a partial transparency and outline. In heraldic terms, shadows represent an outlined shape that reveals a hidden part of the family tree, not a mirror of it [Pastoureau 2003, 26–7]. That is, shadows can be seen as a shameful family link, rather than a precise formal reflection, and such a history reminds us, as Nanna Bonde Thylstrup puts it, of the "inherently unstable form of shadow libraries as a cultural construct" [Thylstrup 2018, 98]. In this light, there are several characteristics of shadow libraries that present minimalist design principles and that we can take as refractive and instructive lenses for understanding mainstream practices.

Minimalist Shadows

13 Just what, then, is minimalist about these shadow libraries? Certainly, they are minimalist in terms of the "minimal barriers" that they present to their readership for access to scholarly publications. In eliminating paywalls and presenting only a flat search box, with no authentication mechanisms, Sci-Hub and Library Genesis are far simpler than the systems used by formal publishers. By contrast, these archives are *not* minimal in terms of the "minimal space" that they consume [Sayers 2016]. Nonetheless, I will argue that we can understand shadow libraries in minimal terms along a number of axes: minimalist surface exposure, minimalist metadata design principles, and minimalist distributional principles. I will cover each of these in turn, primarily with respect to Library Genesis.

14 To begin with the minimal surface exposure of the site, as a highly illegal, copyright-violating initiative that also wants to achieve worldwide transformation of educational potential, Library Genesis finds itself in a minimal-maximal double bind. On one hand, it must remain difficult to access, hidden, and must lie low to evade law enforcement. It must be a resource with a minimal surface. On the other hand, to achieve its stated goals, this resource must be accessible and known to as many people as possible. This double bind is not unique to these libraries. Similar minimal-maximal tensions exist within legal digital social justice projects that seek to criticize powerful governmental norms, such as the

Global Detention Project. Such platforms are not illegal like Library Genesis, but they seek to criticize often-hostile regimes for a broad audience while avoiding tyrannical government crackdowns.

15 Nonetheless, Library Genesis has several technical hurdles to overcome in its quest to retain its minimal surface presence on the Internet. Two of these are the use of the Domain Name System (DNS) and the threat of (distributed) denial of service (DDOS) attacks. On the first of these fronts, to provide a memorable location for the archive, Library Genesis uses DNS [Mockapetris 1987]. This is the system that translates an address such as gen.lib.rus.ec into an Internet Protocol (IP) address (198.167.223.167, for example). This is a useful system because IP addresses are not easily memorable for humans. However, DNS addresses are subject to takedowns and blocking by Internet Service Providers (ISP). That is, if an allegedly infringed party can persuade a court of law or an ISP that a site's sole purpose is copyright infringement, the domain name (or IP address) can be blocked for large swathes of users [Bambauer 2012]. It is, of course, possible to circumvent such blocks by a variety of techniques, such as the use of a Virtual Private Network (VPN) that routes traffic through a different ISP in a friendlier jurisdiction. Systems such as The Onion Router (Tor) are another way of evading these blocks [Farnan et al. 2019].

16 Library Genesis plays the games of whack-a-mole and hide-and-seek with DNS [Schiermeier 2015]. Its mirrors rapidly switch between addresses in an attempt to avoid takedowns [Sar 2015]. Given international jurisdictional challenges, it is very hard for countries that *do* care about copyright infringement to shut down Library Genesis permanently. The distributed nature of DNS acts in the interests of projects such as Library Genesis as it is simply impossible, at present, to garner the level of international legal compliance that would be necessary to shut down its DNS records permanently. Advances in DNS privacy and encryption are only likely to make this problem more difficult to combat (see, for instance, [Schmitt et al. 2018].) That said, Library Genesis also has a communication problem with respect to DNS. If the goal of DNS is to provide memorable addresses for sites on the Internet then changing these addresses within a narrow time window frustrates the ultimate purpose of the system. Sites such as Reddit — a contemporary bulletin board system — spread news of new DNS mirrors, but, of course, once these addresses are public, the takedown process can begin anew. It is a war of attrition and, thus far, Library Genesis is winning the war through its guerilla tactics.

17 The small visible central surface of sites such as Library Genesis, only accessible through circumvention technologies such as VPNs, reflects a phenomenon seen in Alternate Reality Games (ARGs). Such hiddenness is characteristic of ARGs, which often have obscure entry points, conventionally referred to as "rabbit holes": points of ingress that lure in new users searching for clues [Szulborski 2005, 49]. For Garcia and Niemeyer, "a 'good' rabbit hole is one that, for those not looking for clues, blends into the background and noise of the world" [Garcia and Niemeyer 2017, 15]. An example of how Library Genesis seeds such a trail, rather than blatantly advertising itself, can be seen in its metadata. Searching for "Library Genesis" on Google yields a link but its description only reads, "No information is available for this page." Additionally, the site does not feature in the top results of "download scientific books for free." Instead, one needs to be told about the site and then visit it directly.

18 This minimal surface principle is also clear in the site's upload procedures, which intersect with its minimal metadata implementations. Although Library Genesis has a prominent upload link on its homepage, it requires a password to proceed. This gives the impression that uploading is a private activity, conducted only by an elite cadre of individuals who know the magic word.[3] In reality, though, users who follow the "rabbit hole" into the "forum" on the site and then register can readily find the requisite username and password to begin uploading.

19 There are several reasons why Library Genesis presents such a minimal surface. While, surely, the core reason for Library Genesis' minimal surface is its illegality, this feature also lowers the potential for rights-holders to flood the site with false metadata and uploads. Indeed, the gravest threat to Library Genesis's operation would be contamination of its records with inaccurate files, which could cause a denial-of-service attack against the archive. While the simple reason is that all uploads must be vetted, it would also be surprising if the use of the password seriously deters a concerted effort to pollute the library. As we will also see, though, the minimalist metadata principles of Library Genesis actively *encourage* participation.

20 In terms of minimal metadata, it is worth first considering the scale of Library Genesis. The main Library Genesis database collection had just short of 1.2 million records (books) in 2014 [Bodó 2018a, 53]. As of January 2020, this capacity had more than doubled to 2.5 million books. Clearly, a database of epistemological dissent at this size and with this scale of growth requires a lightweight — or, as we might say, "*minimal*" metadata scheme if it is not to collapse under the strain of its size. Although the ext2 and ext3 filesystems that were commonly used on Linux systems at the time of the database's inception have a sizable upper limit of 1.3×1020 files per directory, there are known performance issues in handling more than 10,000 [Poirier 2001]. There are also issues of integrity monitoring at this scale. When dealing with 2.5 million records, how can one ensure that physical media degradation — "bit rot" — does not lead to

corrupted files on the disk that may not come to light for substantial periods? That said, the scale of Library Genesis pales in comparison to formal archives. For instance, a report from 2009 noted that there are over 9,000 missing volumes in the British Library's main collection, illustrating that this problem exists in both the legal and illegal spaces [Dawar and Kennedy 2009]. The British Library is also substantially larger than Library Genesis, with a main collection of approximately 170 million items total and 13.5 million books, with a digital collection that is over a petabyte in size [The British Library 2021].

21 Library Genesis is an enormous archive in terms of total size (maximal), but it is composed of files that are relatively small (minimal). The average (mean) file size in the database is 13.90MB, with a significant portion of these files being less than 5MB. The file size distribution varies for the different file types within the database. Despite the accessibility challenges of the format — the Portable Document Format (PDF) is not the best format for screen readers, for instance — PDFs dominate the archive. For PDFs (n=1,697,927), for instance, the tail is longer to trail off and the average file size is higher at 16.49MB. By contrast, EPUB (n=179,926; average: 7.32 MB) and Mobi (n=23,947; average: 4.33MB) bring the mean file sizes down as they can jettison the formatting information inherent in PDF files. Nonetheless, with this volume of small files, addressability is a core performance concern.

22 The solution that Library Genesis devised for handling such a vast, proliferating archive composed of relatively small files is based on minimalist principles of metadata, distribution, and hashing. Files are *distributed* over multiple directories (called "buckets") while being referenced within a database that includes metadata about the file, as well as a message-digest algorithm v5 (MD5) hash of the file, which serves as a filename. There are 47 metadata fields within the database for each entry, which can be subdivided into categories of file and record information, file and record properties, and external identifiers.[4] Importantly, though, very few of these fields are *required*. The subset that are absolutely necessary are kept to a minimum.

23 The first of these categories — file and record information — pertains to local lookup of the file. For instance, to retrieve a local file for a record, one simply divides the ID field by 1,000 to get the bucket directory, then retrieves the object with the specified MD5 record from that directory (e.g., "243000/9e107d9d372bb6826bd81d3542a419d6"). To specify the content type, a user can append the "Extension" field (e.g., "pdf" or "epub") to the delivered file (and could also infer the multipurpose Internet mail extensions or MIME type if necessary). In addition to using the "Filesize" field to check a download, it is also possible to verify that the contents of the file have not been corrupted at any time. This verification is achieved by computing the file's MD5 hash and then comparing this to the database/filename. Although MD5 is a very old (and possibly even "broken") hashing algorithm, it has a 1.47x10-29 chance of a random collision (that is, of two files sharing the same hash) [Ramirez 2015]. There are no two records within the Library Genesis database that share an MD5 hash as of March 1, 2020.[5] Nonetheless, the subdivision into "bucket" directories of 1,000 files makes the on-disk likelihood of two files sharing a hash/filename extremely unlikely and further reduces this risk. The verification algorithm means that the detection of corrupt files can be handled automatically rather than on a reporting basis from users (although computing all MD5 hashes in the database is a computationally demanding task given the scale). These field and record information fields are the bare minimum required to retrieve a file from the filestore.

24 The second category of metadata field — file and record properties — gives specific metadata for a record, such as the work's title and authors. Interestingly, for the principles of minimal computational design, the database does not store authors in a structured and linked form, but rather as free-text. Traditionally, if one were designing a relational database where an author could be ascribed to more than one book, one would create a separate database table called "author" that had properties such as "first name," "last name," and "ORCID ID," and then, in the "books" table, link to the author. This would create a mechanism to query an author called, say, "Joe Bloggs," and to retrieve all books written by that specific Joe Bloggs. Such a schema would provide high-quality structured data. This is not what Library Genesis does. Instead, it simply stores author names as "plaintext." Sometimes this means that authors and editors are listed in a long string, and it is not clear whether authors of the same name are the same person.

25 Such an approach comes with several minimalist computational advantages. The clearest of these are: (1) there is no need to maintain the structural integrity of the database between tables, and (2) the overhead for entering metadata is greatly simplified by flattening the input. In this way, Library Genesis lowers both the computational requirements for maintenance and the barriers for entry/participation by allowing freeform textual input. Given the size of the database, and the fact that a free-text search can take a long time, the single-field, free-text approach to "author" rather than a linked record also allows interested parties to create their own indexes of the database with relative ease.

26 The third and final metadata field type (external identifiers) — also entirely optional — demonstrates further minimalist design principles. Instead of storing all metadata locally, the database points to offsite management of such data, allowing for others, likely specialists, to focus entirely on metadata curation, collection, and provision. While this carries

some initial lookup overhead for the pirate entering the data, the labor of maintaining this metadata is then outsourced and can be resolved on demand. This all points towards a minimalist labor approach and an awareness that by lowering the threshold for participating to the minimum, the maximalist goals of the project are more likely to be achieved.

27 If Library Genesis encodes many of its minimalist functions in its metadata design principles, its distributional characteristics are also emblematic of such operations. The distributional minimalism principle can be seen in discussions about the preservation of the archive and the future of sites such as Library Genesis. On the first front of digital preservation, the challenge is that this archive is maximal, not minimal. A reasonable cost estimate for simply the hard disks to mirror the entire 33TB archive as of March 2020 is $1,200. This does not include either redundancy in terms of drive failure or the server hardware in which the hard drives would be housed. In terms of minimal cost scaling, the sheer size of the archive makes for a difficult environment if the local copy is to be complete and usable.

28 That said, full preservation in every replica may not be the aim of distribution, and it is of course possible for partial replicas of the database to exist worldwide [Menasché et al. 2013] [Neglia et al. 2007]. Clients such as aria2 also make it possible to download a single file from a torrent swarm, which in theory makes single books within the Library Genesis archive addressable and retrievable in distributed form. However, the torrent swarm is not ideal for Library Genesis's use case. Anybody seeding on the network will have an exposed IP address, and it will be clear that they are participating, with legal risk. For this reason it is likely that many users in the swarm will be connecting using VPNs or so-called seedboxes (remote high-bandwidth servers) in order not only to protect their identity, but also to ensure the efficient, high-speed distribution of the material.

29 Indeed, in late 2019, coordinating around a pirate archiving initiative known as "The Eye," a group of individuals took it upon themselves to ensure the full torrent availability of the Library Genesis filestore. Under this initiative, "swarm peers increased from 3,000 seeders to 30,000 seeders," and "speeds increased from about 60KB/s on most torrents to over 100MB/s" [u/shrine 2020]. The users who undertook this illegal initiative viewed their work as "charitable," presumably under the rubric of "educational advancement," which has long held eleemosynary status in many jurisdictions. The amateur "archivists" who seeded Library Genesis believe that the "initiative fulfils United Nations/UNESCO world development goals that mandate the removal of restrictions on access to science" and that "[l]imiting and delaying humanity's access to science isn't a business, it's a crime, one with an untold number of victims and preventable deaths" [u/shrine 2020]. Somewhat boldly, two seedbox companies — Seedbox.io and UltraSeedbox.com — offered their servers for this avowedly illegal project, thereby providing the bandwidth to achieve the aforementioned speeds [Maxwell 2019]. That said, the torrent archive is not viewed as a long-term viable mechanism for distribution, according to these individuals: "It obviously isn't sane to store 33TB long-term, we just want to push this out to archivers" [u/shrine 2019].

30 A potentially more viable distribution, storage, and retrieval mechanism proposed for Library Genesis could be the InterPlanetary File System (IPFS) protocol [Rahalkar and Gujar 2020]. IPFS presents an addressable, distributed system in which objects are assigned a hash:

> IPFS provides a high throughput content-addressed block storage model, with content addressed hyper links [sic]. This forms a generalized Merkle DAG, a data structure upon which one can build versioned file systems, blockchains, and even a Permanent Web. IPFS combines a distributed hashtable [sic], an incentivized block exchange, and a self-certifying namespace. IPFS has no single point of failure, and nodes do not need to trust each other. [Benet 2014]

31 Further, IPFS has a system called "pinning," in which objects are immutable and pulled down to clients, thus ensuring their permanent availability: "This also makes IPFS a Web where links are permanent, and Objects can ensure the survival of others they point to" [Benet 2014].

32 IPFS does not address the anonymity of nodes that participate, meaning that it would still be possible to locate a serving entity by IP address, which could have legal implications. Such problems could, again, be mitigated by the use of anonymity networks such as Tor. The scalability of IPFS to 33TB of material, spread over 2.5 or so million unique entities of small file sizes is also unproven. An intermediate address lookup table would be needed to translate between the Library Genesis structure and any IPFS version of the platform. However, IPFS would present a minimalist design platform with the distributed preservation of a maximalist scholarly archive, in which many actors all share a small proportion of the total output, contributing towards a holistic totality in fragmentary participation. Nonetheless, once again, the fragmentary approach, in which smaller portions of the archive are distributed between many actors, rather than in a centralised store, yields a minimal approach to the function of a maximal archival.

Minimum-Maximum

Throughout this Library Genesis case study, I have been proposing a framework for thinking about minimal-maximal tensions. It is tempting to consider minimalist computing paradigms as the opposite to proliferating maximal systems. In reality, instead, minimalist paradigms can work as fragments of a distributed whole; this is one of the lessons that we can learn from the pirate archive. 33

These debates are not abstract for scholarly communications. At present, various research funders across the globe are embarking on the establishment of their own platforms to replace the journal ecosystem. Spurred by the same motivation as Library Genesis and Sci-Hub — the slower-than-desired growth of open access — the Open Research Europe, Wellcome Open Research, and Gates Open Research platforms represent significant centralizations of research outputs under funder control, as opposed to a distributed journal publishing architecture. These platforms, running on the for-profit F1000 platform owned by Taylor & Francis, bill themselves as "megajournals," following the terminology developed by the world's largest journal, *PLOS ONE*. There are many competing interests in these systems. Even as such initiatives represent centralizations and maximal architectures, funders insist on open licenses for re-use and re-distribution. Hence, we have centralization efforts mandating licenses that allow for de-centralization and micro/minimal circulation. 34

Library Genesis also yields various principles of participation and consumption that are minimal in their design, even while they yield access to a maximal resource. By requiring only the minimum of metadata provision, Library Genesis lowers barriers to participation, an aspect also bolstered by the off-shoring and distribution of its identifier systems. In terms of consumption, the single search box that returns all results for everything, ever, represents a maximal centralization. In terms of ease of use and accessibility, this one-stop-shop demonstrates a concision and minimalism that is lacking in most scholarly communications systems. 35

Finally, Library Genesis demonstrates a minimally exposed surface, mostly due to its illegality. It would be churlish to pretend that there are huge implications for legitimate scholarly communications infrastructures or lessons to be learned from this. The more people have access to scholarship and research, the better. The minimally exposed surface of the shadow archive is precisely — and only — because it must remain hidden. The strait that this archive must navigate, though, is the space between its maximal presence and its minimal entrance, the rabbit holes of the alternate reality game. 36

Nonetheless, in all, I have attempted in this article to document several features of Library Genesis that serve as an index of minimal-maximal infrastructures. At least some of these features could be adopted by mainstream scholarly communications providers to increase not only the resilience of, but also global equity of access to, our publication ecosystems. While shadows are not equal to mirrors, in these senses I contend that there are lessons to be learned from the (shadow) library. 37

Notes

[1] I would like to extend my thanks in this article to the anonymous readers at *DHQ* who prompted me to rethink the scope and scale of this work, as well as to the editors of the special issue. Work on this article was funded by a Philip Leverhulme Prize from the Leverhulme Trust.

[2] It is worth noting that some of the secondary accounts of Library Genesis can tip over into hagiographic commentary, an angle that I seek to avoid here.

[3] For more on passwords, see [Eve 2016].

[4] The metadata fields are ID, Title, VolumeInfo, Series, Periodical, Author, Year, Edition, Publisher, City, Pages, PagesInFile, Language, Topic, Library, Issue, Identifier, ISSN, ASIN, UDC, LBC, DDC, LCC, Doi, Googlebookid, OpenLibraryID, Commentary, DPI, Color, Cleaned, Orientation, Paginated, Scanned, Bookmarked, Searchable, Filesize, Extension, MD5, Generic, Visible, Locator, Local, TimeAdded, TimeLastModified, Coverurl, Tags, IdentifierWODash.

[5] This was verified by the MySQL query: "SELECT MD5, COUNT(MD5) FROM updated GROUP BY MD5 HAVING COUNT(MD5) > 1;".

Works Cited

Adema and Moore 2021 Adema, Janneke and Samuel A. Moore. "Scaling Small; Or How to Envision New Relationalities for Knowledge Production." *Westminster Papers in Communication and Culture* vol. 16.1 (2021). https://doi.org/10.16997/wpcc.918.

Andrews 2020 Andrews, Penny C. S. "The Platformization of Open." In *Reassembling Scholarly Communications: Histories, Infrastructures, and Global Politics of Open Access*, edited by Martin Paul Eve and Jonathan Gray, 265-276. Cambridge, MA: The MIT Press, 2020.

Andročec 2017 Andročec, Darko. “Analysis of Sci-Hub Downloads of Computer Science Papers.” *Acta Universitatis Sapientiae, Informatica* vol. 9.1 (2017): 83–96. https://doi.org/10.1515/ausi-2017-0006.

Bambauer 2012 Bambauer, Derek E. “Orwell’s Armchair.” *The University of Chicago Law Review* vol. 79.3 (2012): 863-944.

Benet 2014 Benet, Juan. (2014). “IPFS - Content Addressed, Versioned, P2P File System.” *ArXiv:1407.3561 [Cs]*, July. https://arxiv.org/abs/1407.3561.

Berlin Declaration 2003 “Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities.” October 22, 2003. https://oa.mpg.de/lang/en-uk/berlin-prozess/berliner-erklarung.

Bodó 2018a Bodó, Balázs. “Library Genesis in Numbers: Mapping the Underground Flow of Knowledge.” In *Shadow Libraries: Access to Educational Materials in Global Higher Education*, edited by Joe Karaganis, 53–78. Cambridge, MA: The MIT Press, 2018.

Bodó 2018b Bodó, Balázs. “The Genesis of Library Genesis: The Birth of a Global Scholarly Shadow Library.” In *Shadow Libraries: Access to Educational Materials in Global Higher Education*, edited by Joe Karaganis, 25–52. Cambridge, MA: The MIT Press, 2018.

Bodó 2020c Bodó, Balázs, Dániel Antal and Zoltán Puha. “Can Scholarly Pirate Libraries Bridge the Knowledge Access Gap? An Empirical Study on the Structural Conditions of Book Piracy in Global and European Academia.” *PLOS ONE* vol. 15.12 (2020): e0242509. https://doi.org/10.1371/journal.pone.0242509.

Bohannon 2016 Bohannon, John. “Who’s Downloading Pirated Papers? Everyone.” *Science,* April 25, 2016. https://www.sciencemag.org/news/2016/04/whos-downloading-pirated-papers-everyone.

Bosch and Henderson 2018 Bosch, Stephen and Kittie Henderson. “Predicting the Future in 3,000 Words and Charts: The Library Journal Serials Pricing Article.” *The Serials Librarian* vol. 74.1–4 (2018): 224–27. https://doi.org/10.1080/0361526X.2018.1430442.

Boudry et al. 2019 Boudry, Christophe, Patricio Alvarez-Muñoz, Ricardo Arencibia-Jorge, Didier Ayena, Niels J. Brouwer, Zia Chaudhuri, Brenda Chawner, et al. “Worldwide Inequality in Access to Full Text Scientific Articles: The Example of Ophthalmology.” *PeerJ* 7 (2019): e7850. https://doi.org/10.7717/peerj.7850.

Brembs 2020 Brembs, Björn. “The Ultimate Open Access Timeline.” *Bjoern.Brembs.Blog*, March 3, 2020. https://bjoern.brembs.net/2020/03/the-ultimate-open-access-timeline/.

Camarero et al. 2014 Camarero, Carmen, Carmen Antón and Javier Rodríguez. “Technological and Ethical Antecedents of E-Book Piracy and Price Acceptance: Evidence from the Spanish Case.” *The Electronic Library* vol. 32.4 (2014): 542–66. https://doi.org/10.1108/EL-11-2012-0149.

Chan et al. 2002 Chan, Leslie, Darius Cuplinskas, Michael Eisen, Fred Friend, Yana Genova, Jean-Claude Guédon, Melissa Hagemann, et al. “Budapest Open Access Initiative.” February 14, 2002. https://www.soros.org/openaccess/read.shtml.

Correa et al. 2021 Correa, Juan C., Henry Laverde-Rojas, Julian Tejada and Fernando Marmolejo-Ramos. “The Sci-Hub Effect on Papers’ Citations.” *Scientometrics*, January 2021. https://doi.org/10.1007/s11192-020-03806-w.

Das et al. 2013 Das, Jishnu, Quy-Toan Do, Karen Shaines and Sowmya Srikant. “U.S. and Them: The Geography of Academic Research.” *Journal of Development Economics* 105 (2013): 112–30. https://doi.org/10.1016/j.jdeveco.2013.07.010.

Dawar and Kennedy 2009 Dawar, Anil and Maev Kennedy. “British Library Mislays 9,000 Books.” *The Guardian*, March 17, 2009. . https://www.theguardian.com/uk/2009/mar/17/british-library-books-mein-kampf.

Dulong de Rosnay 2021 Dulong de Rosnay, Melanie. “Open Access Models, Pirate Libraries and Advocacy Repertoires: Policy Options for Academics to Construct and Govern Knowledge Commons.” *Westminster Papers in Communication and Culture* vol. 16.1 (2021). https://doi.org/10.16997/wpcc.913.

Eve 2014 Eve, Martin Paul. *Open Access and the Humanities: Contexts, Controversies and the Future*. Cambridge: Cambridge University Press, 2014. https://doi.org/10.1017/CBO9781316161012.

Eve 2016 Eve, Martin Paul. *Password*. NY: Bloomsbury, 2016.

Eve 2021 Eve, Martin Paul, Cameron Neylon, Daniel O’Donnell, Samuel Moore, Robert Gadie, Victoria Odeniyi and Shahina Parvin.*Reading Peer Review:* PLOS ONE *and Institutional Change in Academia*. Cambridge: Cambridge University Press, 2021.

Farnan et al. 2019 Farnan, Oliver, Alexander Darer and Joss Wright. “Analysing Censorship Circumvention with VPNs via DNS Cache Snooping.” *ArXiv:1907.04023 [Cs]*, July 2019. https://arxiv.org/abs/1907.04023.

Faust 2016 Faust, Jeremy S. “Sci-Hub: A Solution to the Problem of Paywalls, or Merely a Diagnosis of a Broken System?” *Annals of Emergency Medicine* vol. 68.1 (2016): A15–17. https://doi.org/10.1016/j.annemergmed.2016.05.010.

Fitzpatrick 2011 Fitzpatrick, Kathleen. *Planned Obsolescence: Publishing, Technology, and the Future of the Academy*. NY: New York University Press, 2011.

Fyfe 2015 Fyfe, Aileen. “Journals, Learned Societies and Money: Philosophical Transactions, ca. 1750–1900.” *Notes and Records: The Royal Society Journal of the History of Science* vol. 69.3 (2015): 277–99. https://doi.org/10.1098/rsnr.2015.0032.

Garcia and Niemeyer 2017 Garcia, Antero and Greg Niemeyer. “Introduction.” In *Alternate Reality Games and the Cusp of Digital Gameplay*, edited by Antero Garcia and Greg Niemeyer, 1–28. NY: Bloomsbury Academic, 2017.

Gil and Ortega 2016 Gil, Alex and Élika Ortega. “Global Outlooks in Digital Humanities: Multilingual Practices and Minimal Computing.” In *Doing Digital Humanities: Practice, Training, Research*, edited by Crompton Constance, Richard J. Lane, and Raymond G. Siemens, 22–34. NY: Routledge, 2016.

Grafton 1999 Grafton, Anthony. *The Footnote: A Curious History*. Cambridge, MA: Harvard University Press, 1999.

Gray 2020 Gray, Jonathan. “Infrastructural Experiments and the Politics of Open Access.” In *Reassembling Scholarly Communications: Histories, Infrastructures, and Global Politics of Open Access*, edited by Martin Paul Eve and Jonathan Gray, 251–64. Cambridge, MA: The MIT Press, 2020.

Green 2017 Green, Toby. “We’ve Failed: Pirate Black Open Access Is Trumping Green and Gold and We Must Change Our Approach.” *Learned Publishing* vol. 30.4 (2017) 325–29. https://doi.org/10.1002/leap.1116.

Harris and Barrett 2019 Harris, Shane and Devlin Barrett. “Justice Department Investigates Sci-Hub Founder on Suspicion of Working for Russian Intelligence.” *The Washington Post*, December 19, 2019. https://www.washingtonpost.com/national-security/justice-department-investigates-sci-hub-founder-on-suspicion-of-working-for-russian-intelligence/2019/12/19/9dbcb6e6-2277-11ea-a153-dce4b94e4249_story.html.

Herb 2018 Herb, Ulrich. “Open Access And Symbolic Gift Giving.” In *Open Divide: Critical Studies on Open Access*, edited by Joachim Schöpfel and Ulrich Herb, 69–81. Sacramento, CA: Library Juice Press, 2018. https://doi.org/10.5281/zenodo.1206377.

Himmelstein et al. 2018 Himmelstein, Daniel S, Ariel Rodriguez Romero, Jacob G Levernier, Thomas Anthony Munro, Stephen Reid McLaughlin, Bastian Greshake Tzovaras and Casey S Greene. “Sci-Hub Provides Access to Nearly All Scholarly Literature.” *ELife* 7 (February 2018): e32822. https://doi.org/10.7554/eLife.32822.

Hobsbawm 1981 Hobsbawm, Eric. *Bandits*. NY: Pantheon Books, 1981.

Hockenberry 2013 Hockenberry, Benjamin. “The Guerilla Open Access Manifesto: Aaron Swartz, Open Access and the Sharing Imperative.” *Lavery Library Faculty/Staff Publications*, November 2013, 1–7.

Jurchen 2020 Jurchen, Sarah. “Open Access and the Serials Crisis: The Role of Academic Libraries.” *Technical Services Quarterly* vol. 37.2 (2020): 160–70. https://doi.org/10.1080/07317131.2020.1728136.

Kapczynski and Krikorian 2010 Kapczynski, Amy and Gaëlle Krikorian, eds. *Access to Knowledge in the Age of Intellectual Property*. NY: Zone Books, 2010.

Machin-Mastromatteo et al. 2016 Machin-Mastromatteo, Juan D, Alejandro Uribe-Tirado and Maria E Romero-Ortiz. “Piracy of Scientific Papers in Latin America: An Analysis of Sci-Hub Usage Data.” *Information Development* vol. 32.5 (2016): 1806–14. https://doi.org/10.1177/0266666916671080.

Maxwell 2019 Maxwell, Andy. “Meet the Guy Behind the Libgen Torrent Seeding Movement.” *TorrentFreak,* December 5, 2019. https://torrentfreak.com/meet-the-guy-behind-the-libgen-torrent-seeding-movement-191205/.

Mboa Nkoudou 2020 Mboa Nkoudou, Thomas Hervé. “Epistemic Alienation in African Scholarly Communications: Open Access as a *Pharmakon.*” In *Reassembling Scholarly Communications: Histories, Infrastructures, and Global Politics of Open Access*, edited by Martin Paul Eve and Jonathan Gray, 25–40. Cambridge, MA: The MIT Press, 2020.

Menasché et al. 2013 Menasché, Daniel S., Antonio A. de A. Rocha, Bin Li, Don Towsley and Arun Venkataramani. “Content Availability and Bundling in Swarming Systems.” *IEEE/ACM Transactions on Networking* vol. 21.2 (2013): 580–93. https://doi.org/10.1109/TNET.2012.2212205.

Mockapetris 1987 Mockapetris, P. V. “Domain Names - Concepts and Facilities.” *Internet Engineering Task Force*. November 1987. https://tools.ietf.org/html/rfc1034.

Moore et al. 2017 Moore, Samuel, Cameron Neylon, Martin Paul Eve, Daniel O’Donnell and Damian Pattinson. “Excellence R Us: University Research and the Fetishisation of Excellence.” *Palgrave Communications* 3 (2017). https://doi.org/10.1057/palcomms.2016.105.

Moxham and Fyfe 2018 Moxham, Noah and Aileen Fyfe. “The Royal Society and the Prehistory of Peer Review, 1665–1965.” *The Historical Journal* vol. 61.4 (2018): 863–89. https://doi.org/10.1017/S0018246X17000334.

Neglia et al. 2007 Neglia, G., G. Reina, H. Zhang, D. Towsley, A. Venkataramani and J. Danaher. “Availability in BitTorrent Systems.” In *IEEE INFOCOM 2007 - 26th IEEE International Conference on Computer Communications*, 2007, 2216–24. Anchorage, AK, USA: IEEE. https://doi.org/10.1109/INFCOM.2007.256.

Pastoureau 2003 Pastoureau, Michel. “L’incolore n’existe Pas.” In *Points de Vue: Pour Philippe Junod*, edited by Danielle Chaperon and Philippe Kaenel, 21–36. Champs Visuels. Paris: L’Harmattan, 2003.

Poirier 2001 Poirier, Dave. “The Second Extended File System.” Savannah. 2001. https://www.nongnu.org/ext2-doc/ext2.html.

Rahalkar and Gujar 2020 Rahalkar, Chaitanya, and Dhaval Gujar. “Content Addressed P2P File System for the Web with Blockchain-Based Meta-Data Integrity.” *ArXiv:1912.10298 [Cs]*, January 2020. http://arxiv.org/abs/1912.10298.

Ramirez 2015 Ramirez, Gorka. “MD5: The Broken Algorithm.” *Avira Blog*, July 28, 2015. https://blog.avira.com/md5-the-broken-algorithm/.

Reddy et al. 2021 Reddy, Hrishikesh and Shivang Mishra. “Sci-Hub Case: Legally Removing the Barriers in the Way of Science.” *NLUJ Law Review,* April 29, 2021. http://www.nlujlawreview.in/sci-hub-case-legally-removing-the-barriers-in-the-way-of-science/.

Risam 2018 Risam, Roopika. *New Digital Worlds: Postcolonial Digital Humanities in Theory, Praxis, and Pedagogy*. Evanston, IL: Northwestern University Press, 2018.

Roh et al. 2020 Roh, Charlotte, Harrison W. Inefuku and Emily Drabinski. “Scholarly Communications and Social Justice.” In *Reassembling Scholarly Communications: Histories, Infrastructures, and Global Politics of Open Access*, edited by Martin Paul Eve and Jonathan Gray, 41–52. Cambridge, MA: The MIT Press, 2020.

Russell and Sanchez 2017 Russell, Carrie and Ed Sanchez. “Sci-Hub Unmasked: Piracy, Information Policy, and Your Library | Russell | College & Research Libraries News.” *College & Research Libraries News* vol. 77.3 (2017): 122–25. https://doi.org/10.5860/crln.77.3.9457.

Sar 2015 Sar, Ernesto Van Der. “Sci-Hub, BookFi and LibGen Resurface After Being Shut Down.” *TorrentFreak,* November 21, 2015. https://torrentfreak.com/sci-hub-and-libgen-resurface-after-being-shut-down-151121/.

Sayers 2016 Sayers, Jentery. “Minimal Definitions.” *Minimal Computing Working Group,* October 2, 2016. http://go-dh.github.io/mincomp/thoughts/2016/10/02/minimal-definitions/.

Schiermeier 2015 Schiermeier, Quirin. “Pirate Research-Paper Sites Play Hide-and-Seek with Publishers.” *Nature*, December 2015. https://doi.org/10.1038/nature.2015.18876.

Schiermeier 2017 Schiermeier, Quirin. “US Court Grants Elsevier Millions in Damages from Sci-Hub.” *Nature*, June 2017. https://doi.org/10.1038/nature.2017.22196.

Schmitt et al. 2018 Schmitt, Paul, Anne Edmundson, and Nick Feamster. “Oblivious DNS: Practical Privacy for DNS Queries.” *ArXiv:1806.00276 [Cs]*, December 2018. http://arxiv.org/abs/1806.00276.

Suber 2012 Suber, Peter. *Open Access*. Cambridge, MA: The MIT Press, 2012. http://bit.ly/oa-book.

Suber et al. 2003 Suber, Peter, Patrick O. Brown, Diane Cabell, Aravinda Chakravarti, Barbara Cohen, Tony Delamothe, Michael Eisen, et al. “Bethesda Statement on Open Access Publishing,” 2013 http://dash.harvard.edu/handle/1/4725199.

Swartz 2015 Swartz, Aaron. “Guerilla Open Access Manifesto.” In *The Boy Who Could Change the World*, 26–27. London: Verso, 2015.

Szulborski 2005 Szulborski, Dave. *This Is Not A Game: A Guide To Alternate Reality Gaming*. Macungie, PA: New-Fiction Pub, 2005.

Tanenbaum and Torvalds 2008 Tanenbaum, Andrew S. and Linus Torvalds. “Appendix A: The Tanenbaum-Torvalds Debate.” In *Open Sources: Voices from the Open Source Revolution*, edited by Chris DiBona, Sam Ockman, and Mark Stone, 221–51. Sebastopol: O’Reilly Media, Inc., 2008. http://public.ebookcentral.proquest.com/choice/publicfullrecord.aspx?p=443191.

The British Library 2021 “Facts and Figures of the British Library.” *The British Library*. Accessed September 23, 2021. https://www.bl.uk/about-us/our-story/facts-and-figures-of-the-british-library.

Thylstrup 2018 Thylstrup, Nanna Bonde. *The Politics of Mass Digitization*. Cambridge, MA: The MIT Press, 2018.

Till et al. 2019 Till, Brian M., Niclas Rudolfson, Saurabh Saluja, Jesudian Gnanaraj, Lubna Samad, David Ljungman and Mark Shrime. “Who Is Pirating Medical Literature? A Bibliometric Review of 28 Million Sci-Hub Downloads.” *The Lancet Global Health* vol. 7.1 (2019): e30–31. https://doi.org/10.1016/S2214-109X(18)30388-7.

u/shrine 2019 u/shrine. “Charitable Seeding Update: 10 Terabytes and 900,000 Scientific Books in a Week with Seedbox.Io and UltraSeedbox.” Reddit - r/Seedboxes. Accessed September 28, 2021. https://www.reddit.com/r/seedboxes/comments/e3yl23/charitable_seeding_update_10_terabytes_and_900000/.

u/shrine 2020 u/shrine. “Library Genesis Project Update: 2.5 Million Books Seeded with the World, 80 Million Scientific Articles Next.” Reddit - r/DataHoarder. Accessed September 28, 2021. https://www.reddit.com/r/DataHoarder/comments/ed9byj/library_genesis_project_update_25_million_books/.