# Book catalog evidence in AI copyright cases

How dated library catalogs and copyright registrations can support research into the acquisition of books for AI development.

Text revised October 20, 2026. Dataset payauthors-2026-10-01-m64, built October 5, 2026. Case summaries through October 5, 2026. Validation results updated October 20, 2026.

Read online, with live charts and links: payauthors.com/white-paper

## Overview

Court orders and company filings in several AI copyright cases describe the acquisition of books from Library Genesis, the Pirate Library Mirror, and other collections. Authors and their lawyers still need to establish which works were involved, who holds the relevant rights, and how publication and registration dates affect the available remedies.

Dated catalogs provide part of that record. They can show that a library listed a particular edition by a particular date. Where the catalog supplies a file identifier or checksum, it can also help compare that listing with files disclosed by a company. Establishing that the company acquired the file requires evidence of its acquisition, such as a manifest, logs, testimony, or a stipulation.

PayAuthors compiles catalog metadata and matches it to candidate U.S. Copyright Office registrations. Reports record the sources, matching methods, and date comparisons for each result. They do not determine infringement, current ownership, or the value of a claim.

**22.4M**
catalog records across overlapping collections (records, not unique books)

**31**
dated catalogs and lists: LibGen 25, Z-Library 4, Books3 1, Internet Archive 1

**8.03M**
U.S. literary-work registrations covered, 1978 to 2025

**1.52M**
distinct registrations with at least one candidate catalog match

Record counts include overlap between collections and should not be read as counts of unique books. Source details are in the methodology.

# 1. Why the catalogs matter

In *Bartz v. Anthropic*, the court described downloads of millions of books from LibGen and the Pirate Library Mirror. Meta has acknowledged acquiring LibGen and Anna's Archive data in its own filings. A discovery order in the OpenAI litigation described an employee's 2018 LibGen download as undisputed. The scope and legal consequences differ between cases.

A catalog helps an author investigate whether a relevant collection listed their work. It may distinguish a particular edition, identify a file, or establish that the listing existed before a reported acquisition. It does not establish which files were acquired when a company took only part of a collection.

The *Bartz* settlement illustrates the importance of identifying individual works. Its final Works List covered 482,460 works under a $1.5 billion settlement. Eligibility depended on the list and the settlement's criteria, including identifiers and registration timing. Some books did not qualify because of missing identifiers, registration dates, or mismatches between records. Whether a particular work was released must be checked against the official settlement record. PayAuthors' exclusion screen is an estimate, not the administrator's determination.

# 2. The collections

### Library Genesis

Library Genesis, usually called LibGen, maintains separate fiction and non-fiction catalogs. Entries commonly include a title, author, ISBN, publication year, file format, size, and catalog-reported MD5 checksum. Archived catalog dumps preserve those entries at particular dates.

PayAuthors holds 25 LibGen snapshots dated November 14, 2015 through June 3, 2025. Coverage is uneven. Between the May 2017 and February 2021 snapshots, our source set contains April 2019 and February 2020 snapshots; the April 2019 snapshot contains non-fiction only. A fiction title absent from that snapshot cannot be treated as absent from LibGen.

### Z-Library and its mirrors

The Pirate Library Mirror, or PiLiMi, copied Z-Library and released two indexes in 2022: approximately 2.18 million records in the first release and 3.82 million in the second. Anna's Archive combines and mirrors several collections, including LibGen and Z-Library.

PayAuthors uses four Z-Library catalogs: the two PiLiMi indexes, records added between August 2022 and August 2023, and a January 2025 catalog. Some records include the library's added date. Others establish presence only at the snapshot date. Reports distinguish the available dates rather than supplying an acquisition date the source does not contain.

### Books3

Books3 was a collection of approximately 196,640 books, released in 2020 and included in The Pile training dataset. Its texts reportedly came from Bibliotik, a private torrent site.

PayAuthors holds a published list of 196,608 file names and a separate list of 129,635 ISBNs extracted from the texts. An ISBN appearing in a book's text can refer to another book, so the ISBN list is checked against the file-name list before a work is counted as present in Books3. PayAuthors does not hold the texts.

### Internet Archive lending books

The Internet Archive's lending-library catalog is included separately because pending AI cases allege that companies acquired books from that collection. An entry establishes a lending-library listing, not an AI company's acquisition of it.

The Archive is a nonprofit library. In a separate case, a 2024 appellate ruling rejected its fair-use defense for the scanning and lending of the 127 books at issue. That ruling does not establish that an AI developer downloaded a particular Archive book.

PayAuthors works from catalog metadata only. It does not download, hold, or open any book file.

## 3. What the cases establish

This paper identifies the basis for each statement. A court finding or order, a company's account of its conduct, and an allegation in a party's brief carry different evidentiary weight. PayAuthors' calculations are identified separately as our analysis.

**● COURT RECORD**

What a judge found or ordered, or a fact an order calls undisputed.

**○ ALLEGATION**

What a complaint, brief, or statement of facts asserts. Not decided.

**■ OUR ANALYSIS**

PayAuthors' counts and matches from public catalogs and Copyright Office records.

**◆ COMPANY FILING**

A company's own answer, filing, or paper about its conduct, or its argument.

**GOVERNMENT FILING**

A position filed by a government agency that is not a party.

## Anthropic

● COURT RECORD The June 23, 2025 order in *Bartz* described Anthropic's acquisition of Books3 in early 2021, at least five million LibGen books in June 2021, and at least two million PiLiMi books in July 2022. The PiLiMi acquisition concerned books Anthropic did not already have from LibGen; presence in PiLiMi alone does not establish that a particular copy was included.

◆ COMPANY FILING Anthropic's own interrogatory answer gives the days: Books3 on or around February 25, 2021; LibGen torrented June 8–19, 2021; the Pirate Library Mirror torrented July 8 to August 10, 2022 (*Bartz*, Dkt. 554-6). That is the company's statement, not a court finding.

● COURT RECORD Final approval of the $1.5 billion settlement was granted on July 20, 2026. The timely opt-outs covered 1,802 works associated with 350 owners.

◆ COMPANY FILING Anthropic argues that claims in the opt-out suits are barred by the limitations period. Its September 24, 2026 motion in *Cambronne* is an argument for dismissal, not a ruling. A hearing is scheduled for December 17, 2026.

Sources: *Bartz*, Dkts. 244, 680, 619-3 and 692; *Cambronne*, Dkt. 228. See the Bartz docket and the official settlement website.

## Meta

● COURT RECORD The June 25, 2025 *Kadrey* order describes Meta's first shadow-library use in October 2022 and its acquisition of Anna's Archive data in early 2024. The court granted Meta summary judgment on fair use on the named plaintiffs' record, and the ruling should be reported with that scope.

◆ COMPANY FILING Meta's answers in *Sullivan* and *Elsevier* acknowledge acquiring LibGen and Anna's Archive data, using portions for Llama models, and uploading some data while torrenting. Meta characterizes the uploading as de minimis.

○ ALLEGATION Publishers allege that Meta torrented 134.6 terabytes and uploaded 40.42 terabytes between April and July 2024. Textbook authors allege that about 4,500 users obtained LibGen files from Meta in 2022. Those quantities and distribution claims should not be presented as court findings.

The first trial in the related suits is set for May 3, 2027: Entrepreneur Media's, or another of the individual suits if it settles. Sources: *Kadrey*, Dkt. 598; *Kadrey* Dkt. 862; *Sullivan* answer, Dkt. 85; *Elsevier* answer, Dkt. 78; the *Elsevier* and *Sullivan* complaints. See the Kadrey docket.

## OpenAI

**COURT RECORD** A November 24, 2025 discovery order in *Authors Guild v. OpenAI*, Dkt. 782, describes an employee's 2018 LibGen download as undisputed.

**ALLEGATION** The class plaintiffs' September 17, 2026 statement of facts alleges that two employees torrented approximately 35 terabytes of LibGen material between September 2019 and January 2020. It also describes the renaming of LibGen1 and LibGen2 as Books1 and Books2, and the deletion of those datasets in 2022. These are the plaintiffs' assertions.

**GOVERNMENT FILING** The Justice Department's September 2026 statement of interest argues that model training on copyrighted text is fair use. It did not address how training copies were obtained. It is a government position filed in the case, not the court's decision. The news plaintiffs responded on September 28, 2026.

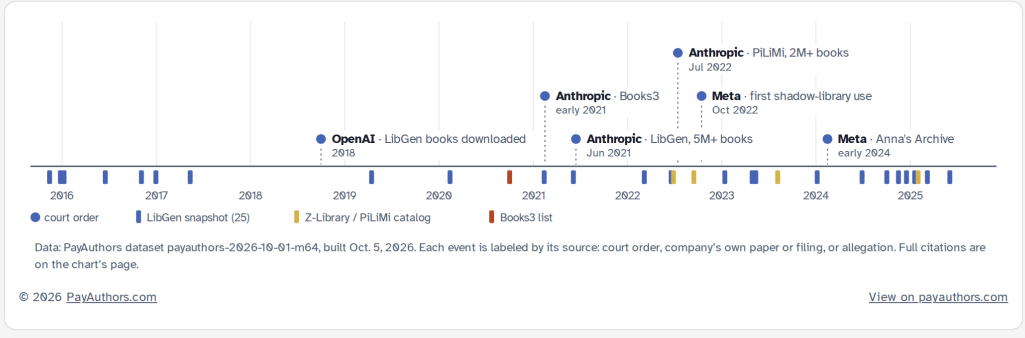
Sources: *Authors Guild*, Dkt. 782; consolidated litigation, Dkts. 1987, 1682 and 2071. See the plaintiffs' statement and the consolidated docket.

## Other companies

**COURT RECORD** In *Nazemian v. NVIDIA*, a May 2026 ruling allowed claims involving PiLiMi, Bibliotik, and BitTorrent to proceed. It was a ruling on the pleadings.

**COMPANY FILING** NVIDIA's June 2026 answer admits acquiring data from Anna's Archive's datasets page and libgen.rs without specifying dates. Bloomberg's BloombergGPT paper identifies Books3 among its training data. DeepSeek's DeepSeek-VL paper describes using e-books from Anna's Archive. A paper's publication date is an outer bound on prior use, not the date of a download.

**ALLEGATION** Other NVIDIA sources and dates remain allegations, as do the claims concerning Google, Apple, xAI, and Perplexity. Sources and later developments are collected on the cases and news pages.



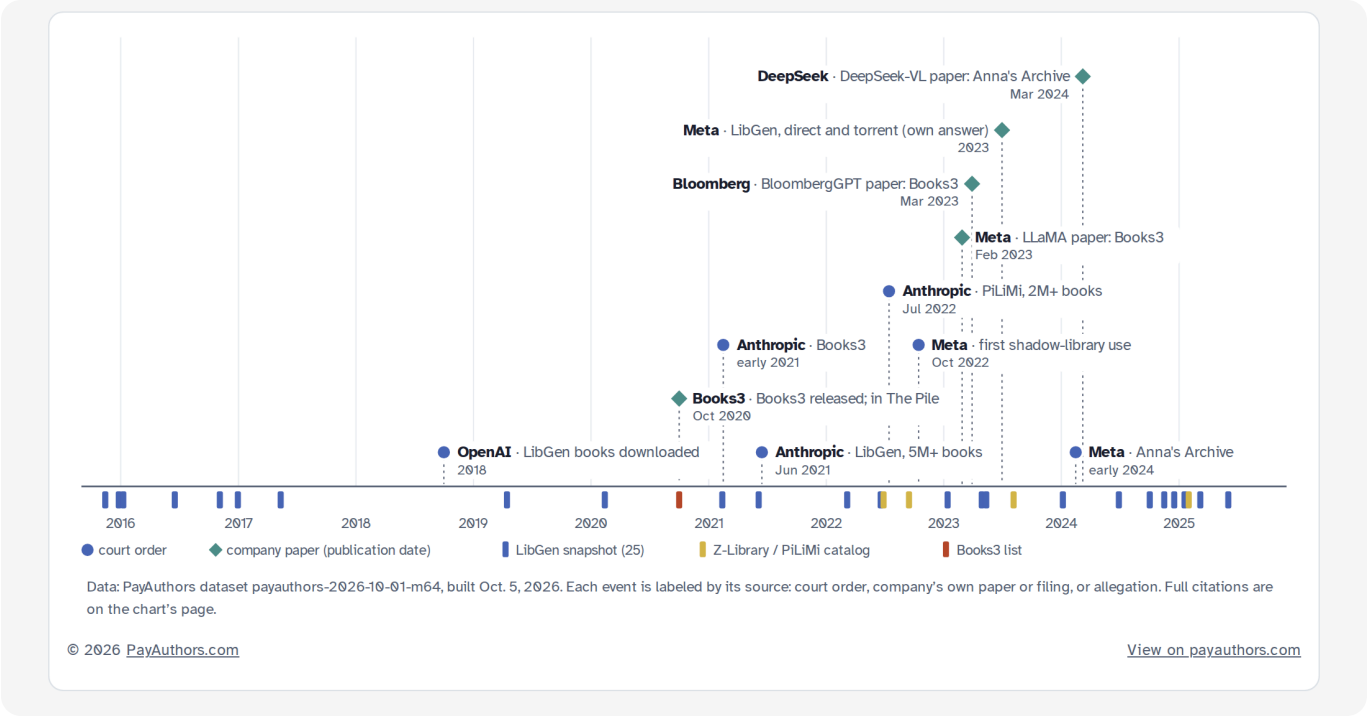

Court record only: the downloads a judge has found or called undisputed. Live version and citations: payauthors.com.

# 4. Comparing catalog and acquisition dates

A snapshot dated before an acquisition shows that the catalog listed a candidate record by that date. A snapshot from the same month does not establish the sequence within the month. For example, the June 5, 2021 LibGen snapshot falls in the month of Anthropic's LibGen acquisition as the court described it; the month alone does not establish whether that snapshot preceded it.

Catalog added dates can supply additional context, but they are the library's own metadata. They are not independently verified acquisition logs. Gaps between snapshots may hide entries that were added and removed between the surviving copies.

A work-specific acquisition record is still needed. Useful evidence can include directory listings, a file manifest, hashes, logs, testimony, or an agreed identification of particular works. The timeline shows the source and precision of the dates used in reports.




Documented downloads, 2016 to 2024, with our dated catalogs. The full timeline, with allegations and how precisely each date is known, is at payauthors.com/timeline.

# 5. Matching a work to its registration

A name search checks the parts of an author's name within catalog author fields. Prefix matching means that Ann can find Anna; a common name can also return another person's books. Records filed under a pen name are linked to a legal name when Copyright Office records support that connection. Other spellings may require separate searches.

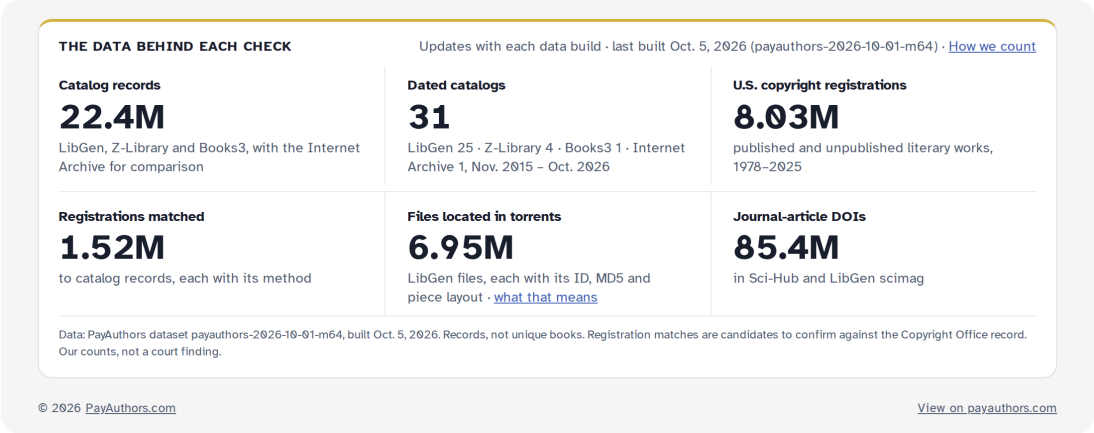
Entries describing the same edition are combined, with the underlying record count retained. Where available, a line includes the library's file ID, catalog-reported MD5, format, and size. Archived torrent metadata can add

the bundle name and the file's position within its pieces. PayAuthors has located 6.95 million LibGen files in this way. These identifiers support comparison with produced files; PayAuthors has not opened the books or recomputed their catalog-reported MD5s.

Registration matching begins with the strongest available correspondence, such as an ISBN supported by a matching name or title. Other methods compare the full title, main title, title words, and author, claimant, or publisher. Conflicting volumes, editions, translations, adaptations, and study guides are examined rather than assumed to be the registered work.

Automatic checks test the candidate against the Office record. Matches the automatic rules can't settle get a second, software review; a person checks samples, and published accuracy figures come only from a person's checks. Paid reports are issued initially in preliminary form. A person then checks the registration matches affecting the counts against the full Office record, and that person's decision overrides the software's. Unconfirmed matches are identified and excluded from those counts.

Reports provide registration numbers and links to the Office record. Registration Details reports also show the full Office record for each match and the recorded documents that cite it, such as transfers. A claimant named at registration may not own the relevant rights today. Transfers must be checked, and the bulk recordation source does not include every later or unlinked document.




The data at a glance. Sources and dates for every figure: payauthors.com/methodology.

# 6. Registration timing and remedies

For a United States work, registration is generally a prerequisite to an infringement action under 17 U.S.C. § 411(a), subject to the statutory exceptions and the procedure following a refusal. Whether a work is a United States work is a legal classification; publication abroad alone does not settle every case.

Section 412 generally limits statutory damages and attorney's fees when infringement began before registration. For published works, registration within three months of first publication can preserve access to those remedies. The statute contains exceptions. A passing date comparison addresses this timing condition only; it does not establish ownership, infringement, registration coverage, or an entitlement to an award.

Statutory damages ordinarily range from $750 to $30,000 per work, with a maximum of $150,000 for willful infringement. The amount is set by the court. The statute generally provides one award per work for the infringements covered by the action, with its own rules for jointly liable infringers and compilations. Actual damages and attributable profits are an alternative that must be proved. The court may award fees and costs under § 505.

PayAuthors compares registration dates with the event dates and catalog metadata described in each report. It considers earlier acquisitions attributed to the same company because, under the continuing-infringement analysis applied in cases such as *Derek Andrew, Inc. v. Poof Apparel Corp.*, 528 F.3d 696 (9th Cir. 2008), a later registration may not cure an earlier commencement. The date screen does not decide how that analysis applies to a particular claim. Unresolved or overlapping dates are shown separately and excluded from the screen's counts.

Settlement and proposed-class criteria are separate questions. The *Bartz* Works List used specified identifier and registration conditions, including a five-year publication window and an agreed Download Date of August 10, 2022. A general registration-date screen should not be described as proof of settlement or class eligibility.

Legal references: 17 U.S.C. §§ 411 and 412, §§ 504 and 505, and the definition of a United States work in § 101.

# 7. Dataset results

The following figures are PayAuthors' analysis of catalog listings and candidate registration matches. They are not findings that the companies acquired every listed work. Copies and editions assigned to the same work are grouped, and overlap between a company's source lists is removed. A work is on a company's list if it was in that collection at the time, whatever its registration date; it passes the date screen only if it was registered before the earliest of that company's copying events it appears in (or within three months of a publication that came first). They were computed across the whole dataset (payauthors-2026-10-01-m64), not through individual reports.

| COMPANY | COLLECTIONS | WORKS ON ITS LISTS | ON TWO OR MORE LISTS | PASS THE DATE SCREEN | PRE-1978 WORKS (CANDIDATES) |
|---|---|---|---|---|---|
| **Anthropic** | Books3, LibGen and PiLiMi works not on the settlement Works List; works whose only copy is a LibGen torrent made through Aug. 10, 2022 are on its list but not counted (caution) | 62,537 | 8,176 | **34,330** | 21,792 |
| **Meta** | LibGen, Anna's Archive and Books3 | 625,735 | 148,757 | **617,167** | 41,130 |
| **NVIDIA** | Anna's Archive (acquisition admitted, date alleged) and Books3 (use alleged) | 274,043 | 29,992 | **267,678** | 57,884 |
| **OpenAI** | The 2018 LibGen event, alleged 2019–2020 torrenting, and a February 2021 Books3 acquisition | 432,371 | 337,675 | **422,577** | 24,321 |

The separate Anthropic source lists contain 5,419 Books3, 28,750 LibGen, 6,787 PiLiMi and 18,904 later-LibGen-torrent passing works, and 21,806 pre-1978 works; those lists overlap. Anthropic's rows count every work with a copy in its downloads that is not on the Works List as collected from the settlement's lookup site, under any registration or edition we can link to it, and that was registered before Anthropic's copying could have begun. Anthropic's count is 34,330 works with a TX or TXu registration and a copy in Anthropic's stated download windows. Another 18,093 such works have a copy only in a LibGen torrent made after Anthropic's stated LibGen window (June 8–19, 2021) and on or before Aug. 10, 2022; they are on its list but are reported only, not counted. **Caution:** Aug. 10, 2022 is the date by which the *Bartz* parties agreed every Works List work was downloaded, for the settlement only and only for Works List works. Anthropic's stated window is an estimate from code dates, and no record shows Anthropic downloaded those later torrents. Works first registered from 1931 to 1977 are shown in their own column for every company. They are matched by title and author to the original registrations for 1964–1977 and to the renewal records for earlier works, both as transcribed by the New York Public Library from the Catalog of Copyright Entries; each is counted once per company and left out if a copy carries a later TX or TXu registration, and for Anthropic also if the work, its renewal or another edition is on the Works List. They are candidates to confirm: in a check of 50 Anthropic matches, 44 were the same work. Registrations that cover only new matter, such as a revision, a translation or illustrations, are included and protect only that new matter. Another 136,767 (Meta) and 119,381 (NVIDIA) pre-1978 works are in Internet Archive lending books those companies are alleged to have downloaded; they are not counted. Other counts cover TX and TXu registrations from 1978 on. The rows should not be added to produce a total across companies. They are counts from the specified dataset and method, not estimates of damages. Official settlement records determine whether a particular work was released as to Anthropic.

**OUR ANALYSIS** Registrations matched to dated catalog records; each match is a candidate to confirm against the Office record.

# 8. Validation and limitations

Registration matching is tested against court-filed lists pairing titles with registration numbers. For *Bartz* opt-out works found in the catalogs, the matcher linked 94.0% to the exact listed registration and 96.0% to some registration for the work. The exhibit contains 1,749 registrations; those percentages are conditional on the work being found in the catalogs, not on the entire exhibit.

For textbook works lists containing 7,001 registrations, mainly from *Cengage Learning v. Google*, the corresponding rates were 93.1% and 96.9%. These tests measure registration matching on those lists. They do not measure every source, name-search error, or ownership question.

In the human review, rescored on October 23, 2026 under analysis revision 33, all 310 sampled matches that the rules would count were judged to concern the same work. PayAuthors reports a 95% interval of 98.8% to 100%. The first scoring, reported October 13, 2026, compared some rows with another catalog listing of the same title and gave 228 of 228; each row is now scored on the listing the reviewer saw. A second reviewer, working separately without the first reviewer's answers or the software's verdict, has so far judged 191 of 194 counted matches to concern the same work, could not decide on 3, and judged none to concern a different work; the first reviewer has also checked all 281 of those rows, and the two gave the same answer on 266. This is a sample result, not a guarantee of perfect accuracy. Neither reviewer had written the matching rules. A separate blind check of 226 counted matches was still in progress when these results were reported.

An earlier software review of 100 candidate matches found 91% to concern the same work. It used a different sample, including matches held out of report counts, and should not be compared directly with the later human precision figure. The methodology describes the samples, exclusions, rules evaluated, and published review records. These are PayAuthors' tests, not an independent audit.

Readers should also account for these limits:

- Names can refer to several people. Catalog author fields may include editors, translators, or other contributors. Review the titles and roles before relying on a result.
- A listing may be missed because of another spelling, a missing author field, or gaps between catalog snapshots. About 178,000 LibGen entries lack an author field usable by the name search.
- A registration may cover only material added to an edition, such as an introduction or translation. A shared title is not enough to establish coverage of the listed text.
- The registration source covers TX and TXu literary works from 1978 through 2025. It does not provide comprehensive coverage of earlier registrations, renewals, or audiobooks. A missing match does not establish that a work is unregistered.
- Unregistered works are outside the registration-based counts. Owners of some non-U.S. works may still pursue remedies without meeting the pre-suit registration requirement.
- A catalog match does not resolve fair use, ownership, authorization, a settlement release, or a filing deadline. Copyright status must be assessed for the relevant work and time, including any separately protected edition material.

# 9. Sources and use of reports

Source files include archived LibGen catalog dumps, PiLiMi and later Z-Library indexes, published Books3 file-name and ISBN lists, the Internet Archive lending catalog, and public Copyright Office data. PayAuthors preserves approximately 61 GB of source material and records checksums, archive locations, and file sizes. Reports identify their dataset build and analysis revision so a result can be traced to the inputs and rules used.

The source records identify what a library described and when. A checksum supports comparison with a produced file; it does not independently establish the accuracy of the catalog description or who acquired the file. Match decisions and their stated bases are retained for review.

For reporting, identify the kind of source supporting each assertion. Use a court's finding within the scope of that ruling, attribute a company's account to the company, and describe an unresolved claim as an allegation. Describe a work as listed in a catalog unless evidence supports the stronger statement that a company acquired that work.

Dates and procedural developments change. The official settlement website provides payment and claim information. The case pages, timeline, and news archive provide further references. Reporters may request a press code from the press page. Codes are shared by an outlet and cover 10 Registration Details reports.

**Suggested attribution:** Source: PayAuthors.com analysis of dated LibGen, Z-Library and Books3 catalog records, the Internet Archive lending-library catalog, and U.S. Copyright Office registrations. Identify the dataset or report version when quoting a specific figure.

© 2026 PayAuthors.com. Charts, figures and tables may be shared unaltered with credit and a link: payauthors.com/terms#reuse.

PayAuthors provides litigation support and holds catalog metadata, not book files. It is not a law firm, does not refer clients to lawyers, and is not affiliated with or endorsed by the U.S. Copyright Office. Questions about ownership, legal claims, and deadlines should be addressed to counsel. Contact: support@payauthors.com. Reuse terms: payauthors.com/terms#reuse.