Journal of Cultural Analytics

Mazières, Antoine and Thierry Poibeau. 2026. MajinBook: An Open Catalogue of Digitally Mediated World Literature. *Journal of Cultural Analytics*, 11(1): pp. 1–21. DOI: https://doi.org/10.22148/jca.1164

MajinBook: An Open Catalogue of Digitally Mediated World Literature

Antoine Mazières, Lattice (ENS-PSL) and LEDa (Dauphine-PSL), France, antoine.mazieres@cnrs.fr
Thierry Poibeau, Lattice (ENS-PSL), France

This data paper introduces MajinBook, an open catalogue designed to facilitate the use of shadow libraries, such as Library Genesis and Z-Library, for computational social science and cultural analytics. By linking metadata from these vast, crowd-sourced archives with structured bibliographic data from Goodreads, we create a high-precision corpus of over 539,000 references to digitally mediated English-language books. Spanning three centuries and reflecting a contemporary selection bias, these entries are enriched with first publication dates, genres, and popularity metrics such as ratings and reviews. Our methodology prioritises natively digital EPUB files to ensure machine-readable quality, while addressing biases in traditional corpora such as HathiTrust, and includes secondary datasets for French-, German-, and Spanish-language works. We evaluate the linkage strategy for accuracy, release all underlying data openly, and discuss the project's legal permissibility under EU and U.S. frameworks for text and data mining in research.

Journal of Cultural Analytics is a peer-reviewed open access journal published by the Center for Digital Humanities. © 2026 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See http://creativecommons.org/licenses/by/4.0/.
OPEN ACCESS

Introduction

The use of text collections as a basis for linguistic inquiry has a long history that far predates modern computing. However, it was the computational turn of the late 1950s that established this practice as a fully-fledged academic discipline, later known as corpus linguistics. Dependent on technology capable of managing large quantities of machine-readable text (McEnery and Hardie), early corpus linguistics often aimed to study the building blocks of language (e.g., grammar and lexis). With the increasing availability of computational power, other subfields soon broadened the practice's aspirations. Sociolinguistics and discourse analysis are prominent illustrations of this methodological borrowing, applying corpus techniques to the study of social structures and norms. Similarly, the field of distant reading applies these methods to uncover patterns in literary history. In this context, the recent progression of Computational Social Science (CSS) can be seen as an advance in scale (i.e., utilising millions of books and social media data) and in tools (i.e., applying advanced machine learning and network analysis) rather than a fundamental change in the long-standing goal of understanding culture through texts.

The vision of using extensive book corpora to represent culture at a large scale was significantly advanced by mass digitisation projects. Beginning in 2002, the scanning effort Google undertook culminated in the Google Books project and yielded, among other resources, the dataset for a foundational "quantitative analysis of culture" (Michel et al.). Shortly thereafter, the HathiTrust Digital Library (Christenson) was formed as a partnership between Google and major research libraries. Now containing over 18 million volumes, HathiTrust has become a go-to resource for supporting broad claims about culture through the computational analysis of books. Its utility is demonstrated by its application across diverse fields, including the history of science (Murdock et al.), literary history (Underwood), gender studies (Underwood et al.), and musicology (Downie et al.).

Despite its scale and utility, HathiTrust is not without significant limitations and biases, many of which stem from its origins. The collection itself is not a neutral representation of world literature; its composition privileges large research universities, while effectively excluding smaller, nonaffiliated institutions and other forms of participation. Furthermore, as a collection built from scanned physical books, the quality of the underlying text can be inconsistent due to data integrity issues and errors from the Optical Character Recognition (OCR) process. The most significant challenge, however, relates to access. The "dark history" (Centivany) of HathiTrust's formation was fraught with legal and political tensions surrounding copyright law. Consequently, a large portion of the collection containing in-copyright

works remains inaccessible for direct reading. This forces researchers to operate within the controlled environment of their research center's secure "data capsule," which not only presents a steeper learning curve but can also inhibit the broader, collaborative evolution of new computational tools and methods that thrive on open data. At the time of writing, the HathiTrust Research Center, which operates this data capsule environment, is scheduled to lose its funding at the end of 2026, with no confirmed successor for computational access to in-copyright materials. These combined issues of representational bias, data quality, and the constraints of a closed or even closing research environment highlight the challenges of using even the largest institutional corpora to study culture.

These limitations have given rise to a new frontier for corpus-based research: the large-scale "shadow libraries" that operate outside of formal institutions. Platforms like Library Genesis (LibGen), Sci-Hub, and Z-Library have emerged as a direct response to the access problem, aggregating tens of millions of books and articles in defiance of paywalls and copyright restrictions. These are not merely illicit collections; they are vast, user-driven archives whose very existence represents a form of crowd-sourced cultural curation, making them an essential new source of data for the computational social sciences and humanities (Karaganis). The recent development of Large Language Models (LLMs) has opened Pandora's box in this regard, moving the use of these datasets from a niche practice to a central component of modern AI. Major companies have publicly acknowledged using such sources in academic papers (Lu et al.) and court cases (*Kadrey et al. v. Meta Platforms, Inc.*; *Bartz et al. v. Anthropic PBC*), where their legal defence of "fair use" has already been partially upheld.

The rationale for our study derives from this context. While these new data sources have been processed as bulk, undifferentiated training data, their use has yet to permeate CSS research. The poor quality of shadow libraries' metadata hinders the precise identification and sampling of content. Such precision is essential for representing a period, a style, or literary culture as a whole, and for navigating between these abstractions and their precise instances, that is, clearly identified books.

To bridge this gap, this paper introduces MajinBook, an open catalogue designed to resolve this metadata challenge and unlock the full potential of these vast literary archives for cultural analytics. Our method leverages metadata from both shadow libraries and a social reading platform to construct a cohesive bibliographic scaffold. This scaffold binds available editions to their original work and first publication date, resulting in a high-precision corpus of over half-a-million books spanning three centuries, along with secondary datasets to foster future works.

This paper is organised as follows. First, we introduce our data sources and their initial processing, namely the shadow libraries LibGen and Z-Library, and the social network Goodreads. We then detail our linkage strategy, its evaluation, and its outcome. Finally, we conclude our paper with some legal considerations about our endeavour.

Shadow Libraries

Library Genesis, Z-Library, and Anna's Archive

The shadow libraries that form the basis of our corpus have distinct but overlapping histories. The oldest and most foundational is LibGen,[1] which was started around 2008 to consolidate and share academic texts. Its origins are rooted in the clandestine *samizdat* culture of the Soviet era and an ethos of providing free access to knowledge for academic communities facing economic hardship and institutional collapse. A pivotal moment in its history occurred in 2011, when LibGen absorbed the massive collection of the defunct shadow library Library.nu, transforming it from a primarily Russian-language archive into a global, multidisciplinary resource.

LibGen's core operating principle is to be radically open, distributing not just its content but also its catalogue and source code to allow for a resilient, mirrored ecosystem. Appearing around the same time, Z-Library launched in 2009 and grew into one of the largest shadow libraries, with a collection that partially overlaps with Lib-Gen's but is separately administered and arguably less open (Karaganis). Both platforms have faced significant legal challenges from publishers, culminating in Z-Library having several of its domains seized by the FBI in late 2022.

The most recent development in this ecosystem is Anna's Archive, which appeared in 2022 and aims to provide a comprehensive, searchable index and mirror of other shadow libraries, including LibGen and Z-Library. For this study, we used LibGen's own platform to access their data but relied on Anna's Archive to source Z-Library's content.

Combining the metadata of Z-Library, as of January 2025, and LibGen, as of March 2025, yields a list of 77, 567, 282 items, of which the vast majority are declared as PDF (77.9%) while 19.2% are referenced as EPUB.[2] The remaining 3.5% comprise various formats spanning from raw text files (.txt, .rtf) to word processing extensions (.doc,

[1] Throughout this paper, "LibGen" refers to the original libgen.rs project, and not any of the numerous forks established since its inception. We do not provide URLs for several platforms mentioned in this paper due to the ephemeral nature of their domain names.

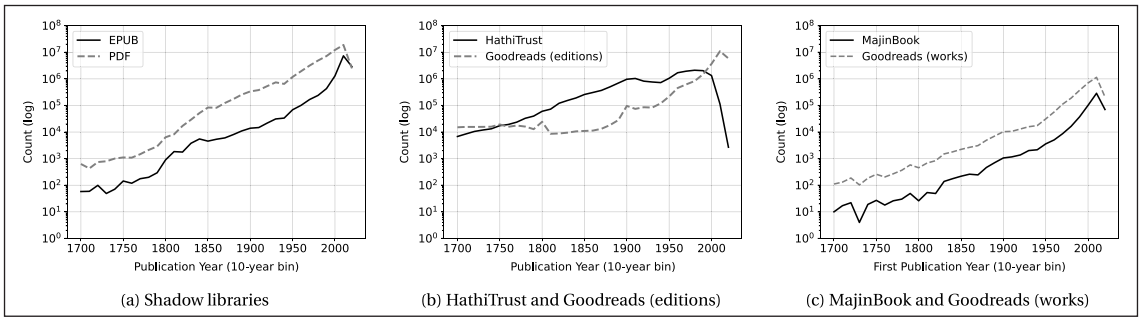
[2] We classify as EPUB any extension easily convertible to this open format without significant information loss, namely .mobi, .azw(3), and .fb2.

.odf), along with more exotic types given the context, such as .exe or .iso, all of which were discarded.

Discarding PDFs

The PDF format is highly varied, comprising everything from partial, amateur scans to well-indexed official publisher versions. Consequently, consistently parsing this content to extract clean, raw, integral text remains a significant technical challenge. Notable pitfalls include OCR discrepancies and the difficulty of preserving the original content's reading order. E-books, on the other hand, are natively digital structures, much like webpages, making their content and metadata entirely machine-readable. To include PDF content would be to place items of potentially dubious quality on a par with the very issues found in other scanned corpora, such as HathiTrust, thereby undermining our core objective of a high-quality, CSS-oriented, and natively digital catalogue.

For these reasons, we made the methodological decision to discard all PDFs. This decision introduces a significant temporal bias, favouring recent publications and older works deemed commercially viable enough to be reissued in a modern digital format. **Figure 1a** clearly illustrates this. On this semi-logarithmic plot, the PDF distribution's growth is relatively linear, indicating a stable exponential increase over time. In sharp contrast, the EPUB subset exhibits a convex shape, signalling a super-exponential growth rate that accelerates towards the present.



(a) Shadow libraries (b) HathiTrust and Goodreads (editions) (c) MajinBook and Goodreads (works)

Figure 1: Temporal distributions and biases of key corpora. The figures illustrate the distinct temporal biases of the key corpora, justifying our methodological focus on natively digital content. All three plots are semi-logarithmic (log y-axis), displaying item counts binned by publication decade. (a) Compares the EPUB and PDF subsets of shadow libraries. (b) Contrasts the scanned HathiTrust corpus with all Goodreads editions. (c) Compares our final MajinBook primary corpus (English) to its Goodreads works scaffold. The plots reveal a fundamental difference in corpus structure. The scanned corpora (PDF, HathiTrust) show relatively stable exponential growth (a linear shape), while the social and natively digital corpora (EPUB, Goodreads, MajinBook) exhibit super-exponential growth (a convex shape) accelerating towards the present. This validates our decision to discard PDFs and confirms that MajinBook (c) is a representative temporal sample of its source.

We argue, however, that this skew is not a simple limitation but a deliberate methodological filter. Rather than a bias against the full shadow library corpus, our choice acts as a productive sieve for a specific, n atively digital representation of culture. The trade-off is explicit: We sacrifice the historical completeness of scanned corpora for a dataset of cleaner, more structured, and machine-readable content. Moreover, this productive sieve defines our object of s tudy. We are not claiming to represent "world literature" in its historical entirety, but instead the digitally mediated canon: culture as it is curated, circulated, and consumed in the twenty-first c entury. This temporal bias towards natively digital and commercially viable reissues is not necessarily a flaw, but rather a defining characteristic of our corpus.

This contemporary constraint yields a significant and somewhat counterintuitive benefit. As **Table 1** shows, the EPUB subset is far more linguistically diversified (Herfindahl Index (HI) = 0.24) than its PDF counterpart (HI = 0.74). Notably, it is also less concentrated than the HathiTrust corpus (HI = 0.32), demonstrating that our sieve produces a dataset that is both high-quality and linguistically diverse. This resulting language fragmentation, which reduces the share of English from over 85% in the PDF set to only 48% in our corpus (**Table 1**), can be seen as a positive indicator of cultural breadth, enhancing the dataset's representativeness for cultural analytics.

	Shadow	**libraries**	**HathiTrust**	**Goodreads**
	EPUB	***PDF***		***Editions***
No. items *(in millions)*	15.2	50.5	18.9	24.2
Herfindahl Index *(normalized)*	0.24	0.74	0.32	0.57
English	47.93	**85.97**	54.95	75.44
French	**8.26**	1.09	7.04	4.63
German	5.94	3.96	**8.61**	3.66
Spanish	**9.45**	0.94	5.44	3.84
Russian	**4.31**	3.45	2.85	0.65
Chinese	**9.59**	2.35	3.26	0.58
Italian	**4.24**	0.51	2.27	2.17
Portuguese	**1.68**	0.33	1.23	1.16
Dutch	**2.06**	0.14	0.71	0.91
Japanese	0.75	0.09	**3.07**	0.85
Polish	**1.13**	0.08	0.64	0.61
Arabic	0.57	0.05	**1.05**	0.45

(Contd.)

	Shadow	libraries	HathiTrust	Goodreads
	EPUB	*PDF*		*Editions*
Czech	**0.39**	0.03	0.36	0.28
Swedish	0.25	0.01	**0.50**	0.38
Danish	0.26		**0.36**	0.27
Hungarian	**0.49**	0.10	0.26	0.17
Korean	0.29	0.02	**0.36**	0.09
Turkish	0.14	0.12	0.21	**0.52**
Bulgarian	**1.36**	0.11	0.14	0.17
Indonesian		0.10	**0.28**	0.18
Romanian	0.14	0.04	0.13	**0.29**
Ukrainian	0.06	0.11	**0.15**	0.09
Persian		0.02	0.17	**0.19**
Greek	0.01	0.07	0.12	**0.23**
Catalan	**0.15**		0.07	0.11
Serbian	0.02	0.01	0.16	**0.17**
Norwegian		0.01	**0.21**	0.14
Hebrew	0.07	0.02	**0.46**	0.08
Lithuanian	**0.10**	0.04	0.02	0.09
Bengali	0.05	0.06	**0.11**	0.07
Finnish	0.02		0.10	**0.29**
Croatian	0.01	0.01	**0.23**	0.11
Vietnamese	0.02	0.01	**0.11**	0.09
Slovak	0.04		0.06	**0.09**
Thai		0.01	**0.19**	0.09
Latin		0.02	**0.83**	0.07
Hindi	0.02	0.01	**0.25**	0.07
Afrikaans	**0.05**	0.01	0.03	**0.05**
Latvian	**0.05**	0.01	0.02	**0.05**
Slovenian			**0.07**	**0.07**

Table 1: Comparative analysis of corpus scale and linguistic diversity. This table provides the core quantitative justification for our methodological decision to focus on the EPUB subset. It compares the scale (in millions of items) and linguistic concentration (normalised Herfindahl Index) of the EPUB and PDF shadow library subsets against the HathiTrust and Goodreads corpora. The analysis reveals a stark trade-off: The PDF corpus, despite its size, is linguistically homogeneous (HI=0.74), with English comprising 85.97% of its content. In contrast, the EPUB subset (our chosen base) is the most linguistically diversified of all corpora (HI=0.24), significantly outperforming even the HathiTrust collection (HI=0.32). This demonstrates that our filtering process, while reducing the total item count, produced a smaller but far more balanced and representative dataset for cultural analytics.

Goodreads

Goodreads, founded by Otis Chandler and Elizabeth Khuri, is a prominent social reading platform, launched in 2006 and acquired by Amazon in 2013. It boasts over 150 million members who rate, review, catalogue, and discuss books, creating a vast digital archive of contemporary reader responses and amateur criticism. Academics utilise this extensive dataset to analyse reading activity at scale, compare modern literary reception with historical patterns, and understand how works gain or lose popularity (Walsh and Antoniak; Antoniak et al.; Bourrier; Kousha et al.; Hu et al.). Nevertheless, these studies acknowledge limitations, including demographic biases within the user base (predominantly white, female, and U.S.-centric), and the potential for review manipulation.

We first detail our data requirements and the rationale for our choice of source, before describing our crawling methodology and its outcome.

Data Requirements and Rationale

Bibliographic Scaffold

Homer's *Odyssey* can be understood as an abstract cultural artefact, a shared reference to ancient history, cunning, and homecoming. However, each physical copy of the text is a unique object that bears witness to the technologies of its creation, the history of its translations, and the pressures of censorship. The word "book" often carries both of these meanings. To be more precise, we use the term "work" to refer to the abstract text (the cultural artefact) and "edition" to refer to a specific published version. To some extent, the edition is the object and the work is the idea. To study the latter, one must necessarily go through the former, computationally or otherwise.

This *work-edition* scaffold partly relates to more advanced classification systems, for instance the *Functional Requirements for Bibliographic Records* (FRBR) (International Federation of Library Associations and Institutions), devised by the International Federation of Library Associations and Institutions (IFLA). Their *work* entity aligns closely with our eponymous category, both referring to a "distinct intellectual or artistic creation." With a degree of interpretative flexibility, and in the specific context of books, our notion of *edition* corresponds roughly to the grouping of FRBR's Expression and Manifestation entities. This approach also conceptually parallels the method used by the Online Computer Library Center (OCLC) to cluster bibliographic records into *work* entities within WorldCat.

A key advantage of this system is that it binds each edition not to its own publication date, but to the first publication date of its corresponding work. This

temporal grounding is crucial for diachronic analysis, as it anchors the work as a cultural representation to the period in which it was primarily written. This prioritises the history of literary production over the history of reception. While historians of the book might prefer to track the circulation of specific editions, our approach is designed to map, and therefore focus on, the cultural onset of the intellectual artefact. Simultaneously, the inclusion of Goodreads metadata offers a rare quantitative proxy for modern reception, capturing how these historical works are read and valued by users today.

In practice, we adopt the internal work-edition mappings provided by Goodreads. These mappings form the core structure of its platform and proved to have very few inconsistencies. In the dataset described below, 60.9% of the works feature a precise first publication date.

Genres and Popularity Metrics

While the first publication date is crucial for CSS researchers to select works from a specific period, other partitions in the data can prove highly relevant to narrow the corpus based on alternative criteria. First and foremost, in the context of books, "genres" are a common classification enabling thematic differentiation. The precise internal classification mechanism for genres on Goodreads remains opaque, but the genres section associated with most books is established at the work level, that is, all editions of a given work share the same genres. These categories appear to be the product of a top-down curation of the platform's crowd-sourced "shelves" that enable users to organise their books. Certain shelves, such as "fiction" or "romance," are promoted to the genres section, whereas others, such as "i-own" or "to-read," are excluded.

Popularity is another common corpus selection criterion. Indeed, the very inclusion of a work in any corpus represents a form of selection against obscurity, an escape from what Franco Moretti terms the "slaughterhouse of literature" (Moretti). Different corpora embody distinct curation models. For instance, the HathiTrust dataset, composed of contributions from partner libraries, represents a decentralised yet expert-driven curation model, while shadow libraries are largely crowd-sourced. Regardless of the base corpus, distinct popularity metrics can prove useful to differentiate between popular and obscure works. Goodreads, as a social network, offers typical features such as ratings and reviews. Like HathiTrust and shadow libraries, these metrics embody a specific survivorship bias that characterises the form of popularity they convey. In that regard, Goodreads ratings reflect approximately two decades of user engagement (2006–present) and should

therefore be understood as a snapshot of contemporary reading preferences and not a stable historical measure of popularity. Integrating these features into our catalogue offers significant analytical potential, allowing researchers both to study platform-specific popularity patterns and to select sub-corpora based on specific popularity criteria. This latter point is particularly relevant for computationally intensive AI research, where budget is a significant constraint. Quantitative popularity metrics allow a researcher to create an affordably sized sub-corpus by sampling the most visible works, while still preserving the catalogue's thematic and temporal representativeness, subject to the biases acknowledged above.

Goodreads versus Other Data Sources

Before selecting Goodreads, we explored several alternative data sources. We believe it is relevant for future research to make our rationale for this choice explicit. The most significant alternative was Open Library, which was founded by Aaron Swartz, Alexis Rossi, and others as a nonprofit initiative of the Internet Archive. It launched in 2006 with the ambitious goal of creating "one web page for every book ever published."

On the surface, Open Library is a compelling choice. Its primary strength is its open nature, providing free, bulk access to one of the largest structured bibliographic datasets in the world. It boasts tens of millions of records, covers a rich long-tail of lesser-represented languages, and utilises a formal work-edition data model. However, the project's greatest strength (its open, wiki-based contribution model) is also its most significant weakness. The metadata is unreliable, containing frequent duplicate records and incomplete entries. More critically for our study, we identified two fatal flaws: a low coverage rate for first publication date and pervasive inconsistencies in the work-edition linkages.

We also evaluated Wikipedia as a potential alternative or complementary source. Its appeal was twofold. First, it offers a distinct form of curation: A book's existence as a dedicated Wikipedia article signals a type of "encyclopaedic notability" that is different from, and complementary to, the social popularity metrics of Goodreads. Second, its underlying data model (via Wikidata) is often structured around works and first publication dates, aligning well with our bibliographic scaffold.

However, we encountered a significant technical challenge in isolating the correct entities. While many books are instances of the "written work" category, this exists within a deep and complex hierarchical system. It proved difficult to reliably discriminate actual books from other types of written works, such as articles or pamphlets. Given this difficulty in reliably scoping the corpus, we set this promising source aside for the current study.

We therefore accept this trade-off. In exchange for the high-quality, consistent metadata that Open Library lacks, and in avoidance of the entity-scoping ambiguity from Wikipedia we could not resolve, we build our scaffold upon Goodreads. We explicitly acknowledge this foundation is not neutral, but rather a structure that reflects a social and commercial curation of literature. Our dataset is therefore not a representation of "world literature" in its entirety, but a representation of twenty-first-century literary culture as it is mediated by a major commercial platform.

The Crawl of Goodreads

To harvest data from goodreads.com, we gathered 1,225,390 editions to serve as seeds, the initial sets of items from which our crawl expanded outward. These were drawn from the platform's various public book aggregations: 30 tags, 642 shelves, 1,419 genres, 8,194 awards, and 30,497 lists. Together, these seed editions corresponded to 990,010 distinct works by 927,707 authors. This initial set of works comprised a total of 11,319,891 editions.

We then recursively expanded this baseline by collecting works from the platform's recommendations (such as those featured in the "Readers also enjoyed" section), as well as other relevant works by the authors already gathered. To mitigate the long tail of obscure items, we applied a specific inclusion threshold for an author's additional works: They were required to have at least one rating and two editions. We refer to each successive round of this expansion as a depth, with the seed set at depth zero.

This iterative crawl continued until the author-led discovery of new items plateaued at depth four and ceased at depth five, coinciding with the complete exhaustion of the recommendation-led crawl. These simultaneous terminations suggest that our dataset is a comprehensive representation of Goodreads's discoverable content

This approach effectively neutralises the risk of algorithmic selection bias. By exhausting the network, we eliminate the path dependency of the recommender system, yielding a corpus defined by the platform's boundaries instead of its suggestions.

The final dataset comprises 4,778,124 works, 28,105,913 editions, and 2,150,522 authors (cf. **Figure 2**). In total, 15% of works were discovered via recommendations, 20% came from the seed set, and 65% were retrieved through author-based expansion.

To the best of our knowledge, and based on our review of the state of the art (Wan and McAuley; Wan et al.; Kaggle; BrightData; Hu et al.), this represents one of the largest publicly available crawls of Goodreads data published to date.

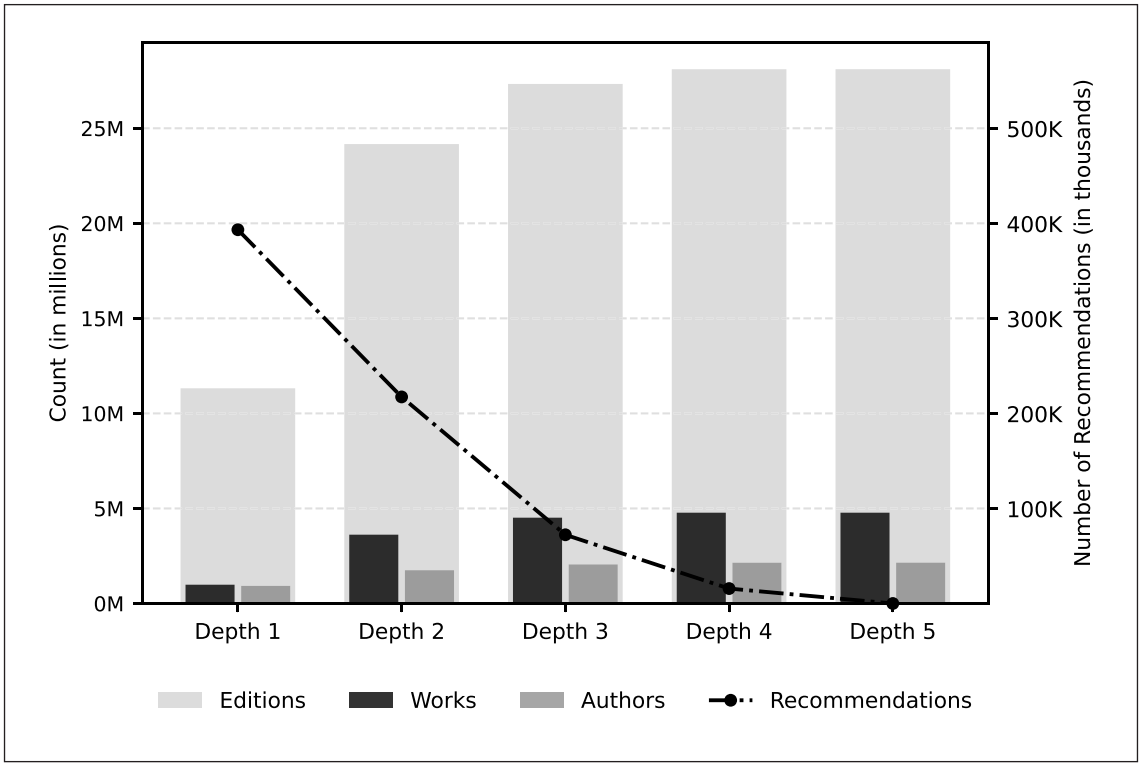


Figure 2: The crawl of Goodreads: Item acquisition and recommendation decay. The figure illustrates the efficiency of our crawl me thodology. The bars show the cumulative counts of editions, works, and authors (left axis, in millions) gathered at each stage. The line plot tracks the number of new recommendations (right axis, in thousands) discovered at each depth. The plot reveals a power-law distribution: The initial depths rapidly capture the most prominent items, while subsequent depths explore a long tail of less-connected content.

MajinBook

Preparing Data for Matching

Our matching strategy relied on book identifiers, primarily the International Standard Book Number (ISBN) and the Amazon Standard Identification Number (ASIN), extracted from both shadow library metadata and the EPUB files themselves. We observed that these identifiers were often unreliable for matching a file to its exact edition. However, we found that even an incorrect identifier was still a highly reliable pointer to the correct work. We therefore designed our strategy around this insight, linking files to a robust work-language pair rather than pursuing a more fragile edition-level match.

From the EPUB subset of the shadow libraries catalogue, we harvested all downloadable files. T hese w ere then standardised to a consistent EPUB format and converted into raw, un-marked-up text files; any error during this process resulted in the file being discarded. This initial filtering yielded our base dataset of 11,130,032 files.

We then applied a size filter, discarding 213,167 items deemed too small (< 10 KB, ca. 4 A4 pages) or too large (> 10 MB, ca. 4,000 A4 pages) to be actual books.

First, we computed a 128-element MinHash signature for every item. We then used Locality-Sensitive Hashing (LSH) to efficiently identify, for each item, all items with an approximate Jaccard similarity of 0.8 or higher.[3] This 0.8 threshold is a conservative standard for identifying near-duplicates, while still allowing for minor textual variations between different editions.

This enabled us to group items as potential duplicates or different editions from the same work in a given language, yielding a list of 1,954,010 unique clusters with more than one item, 77.6% of which had at least one element with at least one book identifier. This set of 1,516,332 clusters of identifiable shadow library items constitutes our base for matching with the Goodreads metadata.

To construct our Goodreads matching set, we filtered our 4.78 million works for those that included a first publication year, an essential field for our diachronic scaffold. This step yielded our base dataset of 2,904,994 works, about 60.9%. This filtered set is highly suitable for the matching task, as 99.7% of these works also feature at least one book identifier.

This significant reduction is a deliberate methodological choice, not a simple loss of data. We found strong evidence that the 39.1% of works we discarded are, on average, less central to the platform and of lower data quality. For instance, works without a first publication year are significantly less evaluated, with a median number of ratings of 9 (IQR = 2–41), compared to 17 (IQR = 4–99) for the works we retained. This suggests they are less engaged with by the platform's users. Furthermore, this metadata-completeness correlates with discoverability: A first publication year was present on over 80% of items in the initial crawl depths but dropped progressively to the final 60% average. This indicates that the most discoverable items on Goodreads are also the most likely to have complete metadata.

Matching and Evaluation

Our strategy for matching shadow library clusters with Goodreads works relied on identifier overlap. A given cluster was linked to a given work if the set of all identifiers found within that cluster's items had a non-empty intersection with the set of all identifiers aggregated from that work's editions. Any such match flagged the whole cluster as a potential candidate for that work, tagged with the cluster's language. This

[3] We used the Python library, *datasketch*, which automatically computed the optimal parameters of nine bands and thirteen rows to find pairs at this similarity level (see: github.com/ekzhu/datasketch).

process enabled us to link 770,840 Goodreads works to at least one cluster. In total, these retained clusters comprise 4,216,400 shadow library items.

Deeming identifier-based links to be a necessary but not sufficient condition, we implemented a second verification step using metadata. This step performs an exhaustive title comparison for each candidate. For a given cluster and a potential work, we compared every title within the shadow library cluster against every edition title for that work (in the cluster's language). For each of these pairs, we computed a partial ratio fuzzy match, yielding a score from 0 to 100. Finally, we calculated the mean of this entire comparison matrix to produce a single, robust title score for the candidate. Overall, the distribution of title scores is highly left-skewed, with a median score of 99.1 (IQR = 86.6–100.0). This indicates that the vast majority of identifier-based matches are confirmed by our title matching method.

We then conducted a human evaluation study to validate our matching method and determine an optimal title score threshold. For this, we sampled 200 English-language candidates, using a stratified approach that deliberately overrepresented items with lower, more ambiguous scores.[4] We built a simple web interface for this task, which, for each candidate, displayed the corresponding Goodreads title and authors alongside the first 100 paragraphs of a random item from the shadow library cluster.

Evaluators from our lab were then asked to assess each candidate and assign one of three labels: "Yes" (a correct match), "No" (an incorrect match), or "I don't know/I can't tell." We concluded the evaluation once every item in the sample had received at least one review. **Figure 3** shows the results for the 143 items for which evaluators could come to a conclusion and plots the resulting precision, recall, and dataset size as a function of the title score threshold. The figure clearly illustrates the classic trade-off: As the threshold increases, precision rises, while recall and dataset size fall. Based on our stated goal of a high-precision catalogue, and supported by this evaluation, we selected a final threshold of 80, which achieves a precision of nearly 1.0 (**Figure 3**).

The 57 items (28.5%) that evaluators labelled "I don't know/I can't tell" were not factored into the primary precision-recall computation. Our lab colleagues did not determine these items to be incorrect, but rather the items were not able to be evaluated. They typically lacked title pages and started in media res, offering few visual cues for confirmation. To determine whether a high title score could be trusted on these ambiguous files, we conducted a deeper, secondary analysis on the 24 ambiguous items with a title score above our operational threshold (80) and concluded that 22 were

[4] More precisely, five items with a title score between [0, 20], 15 in]20, 40], 10 in]40, 50], 15 in]50, 60], 25 in]60, 70], 30 in]70, 80], 50 in]80, 90], and 50 in]90, 100].

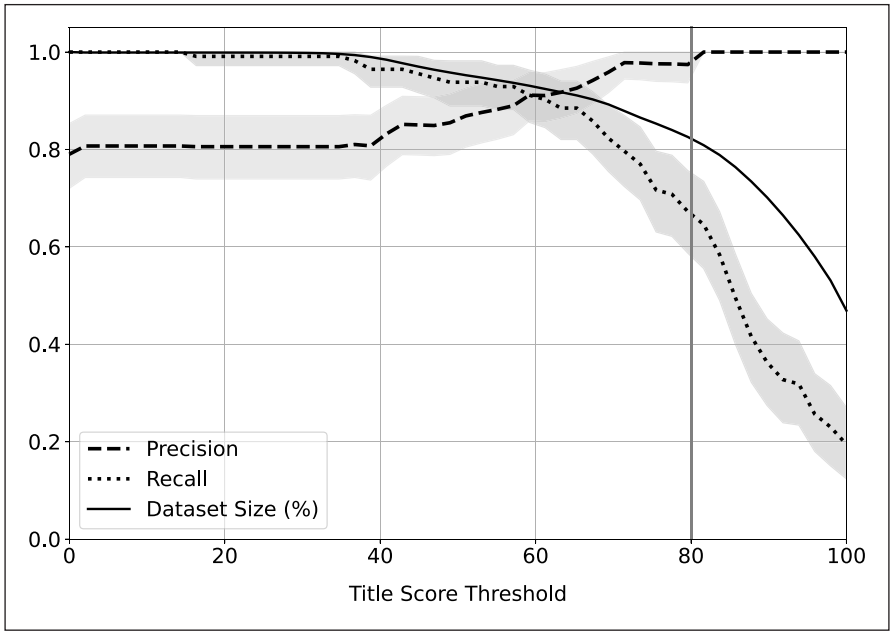


Figure 3: Precision and recall trade off against each other as the title score threshold varies. The plot shows the point estimates for precision (dashed line) and recall (dotted line), along with their 95% confidence intervals (shaded areas), derived from bootstrap resampling of 143 human evaluations. The solid black line indicates the percentage of the dataset retained at each threshold. A vertical line marks our chosen operational threshold of 80, which prioritises high precision for the final catalogue.

correct matches (91.7% precision). The two errors were not random: One was a single book matched to a three-book set, and the other was a different book by the right author. This demonstrates that the title score's predictive power is robust even for files that are visually ambiguous. It confirms that the formatting issue is statistically independent of the score's relevance and validates our decision to proceed with the threshold derived from the 143 conclusively-labelled items.

Release

Primary Dataset: The MajinBook Catalogue

We only evaluated our matching methodology on English titles, and 82.9% of our retained matches are for English-language content. This English subset, numbering 539,530 items, therefore constitutes our primary dataset, fulfilling the large-scale, high-quality ambition of our study.

Each entry in this final catalogue includes the following metadata fields:

- Goodreads work ID
- First publication year

- Authors' full names and Goodreads IDs
- Title
- Rating and number of ratings
- Shadow libraries IDs corresponding to this work in English

Additionally, 84.04% of the catalogue features a list of genres and 99.23% includes the number of reviews.

Secondary Datasets

Despite the overwhelming dominance of English, our methodology captured significant volumes of content in other languages with robust temporal distributions (**Figure 4**). Although we did not conduct human evaluation for these languages, their title score distributions are highly left-skewed, closely mirroring the distribution of our validated English set.

While this similarity is a promising heuristic, it is not a substitute for formal validation. Based on these two criteria (significant volume and a comparable title score distribution) we selected three of the largest non-English corpora for release: French (47,960 items), German (35,559), and Spanish (30,169). The sharp drop in volume relative to the source shadow libraries is explained by the fact that our dataset is constrained by the linguistic distribution of Goodreads (cf. **Table 1**).

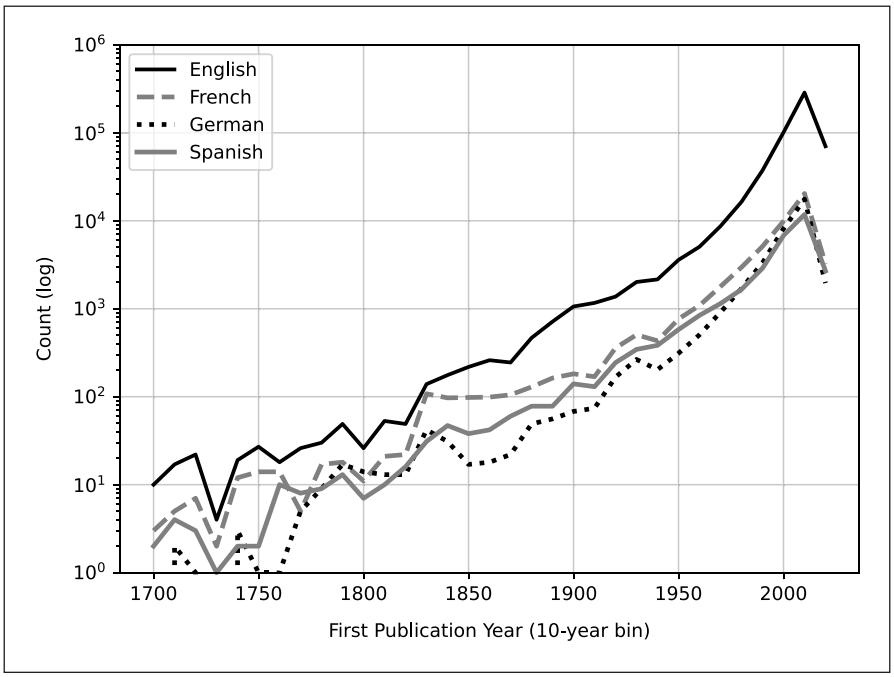


Figure 4: The primary (English) and secondary datasets differ in their temporal distributions.

The feature coverage for these datasets is identical to the primary English corpus, with the sole exception of genres, which falls to 77.4% for Spanish, 75.5% for German, and 70.1% for French. We hypothesise that this is not an artefact of our matching process but rather reflects the sparser metadata available for non-Anglophone works within Goodreads itself.

We must, however, stress that the precise quality of these matches remains unverified. We release these secondary catalogues as experimental datasets to foster future work, noting that they do not carry the same high-precision guarantee as our primary English corpus.

First Publication Year (10-year bin)

Underlying Datasets

Finally, in addition to the primary catalogues, we are releasing the underlying datasets that enabled this study, namely

- the formatted metadata for the retained elements from the shadow libraries EPUB subset;
- the formatted metadata extracted for every work, edition, and author from the Goodreads crawl;
- the MinHash signatures for every item considered in the shadow libraries.

Legal and Ethical Considerations

The legality of using shadow libraries for research remains a contested and evolving issue, as illustrated by recent lawsuits involving major AI companies (*Kadrey et al. v. Meta Platforms, Inc.*; *Bartz et al. v. Anthropic PBC*). At the time of writing this paper, legal frameworks in many countries are being adjusted to find a balance between copyright law and the AI use of copyrighted materials (Sag and Yu). Our project's legal standing rests on a critical distinction between our final product (a public, metadata-only catalogue) and our process (acquiring and analysing data from Goodreads and shadow libraries).

MajinBook's final output is a bibliographic index. The titles, author names, publication dates, and identifiers that populate our catalogue are factual data, not copyrightable expression (*Feist Publications, Inc. v. Rural Telephone Service Co.*). Moreover, our index reproduces no textual content from the underlying books, a posture considerably more restrained than the snippet display upheld as fair use in *Authors Guild v. Google* (2015), where the court permitted the copying of entire books for a transformative search index.

Goodreads data was acquired by scraping publicly accessible pages without circumventing any authentication mechanism or technical protection measure. Following *hiQ Labs v. LinkedIn* (2022), such access to publicly available data does not constitute "unauthorised access" in the sense of the Computer Fraud and Abuse Act (CFAA). From a European standpoint, where the data was acquired, this falls squarely within the Text and Data Mining (TDM) exception for research purposes, which provides a mandatory research exception (European Parliament and Council of the European Union) that cannot be overridden by contract (Art. 7). Even if Goodreads's terms of service were to prohibit automated access to their platform, this regulation grants academics acting in good faith and with no commercial purpose the right to do so.

Lastly, harvesting and computing data from shadow libraries as we did can be considered through the lens of several legal frameworks: In the EU, the above-mentioned "TDM exception," and in the U.S., the fair use doctrine incorporated into the Copyright Act of 1976 complemented by the recent Text and Data Mining exemption to the Digital Millennium Copyright Act (1998). The precise application of these frameworks is a complex and evolving debate, one that is beyond the scope of this paper and our expertise as nonlegal scholars. However, from our reading, the core of the discussion, both legal and ethical, appears to hinge on the intent and context of the use.

Our work is situated in a public research institution, with no commercial interest, and aims to foster public knowledge about language, literature, and cultural representations. Applying the factors commonly weighed by these frameworks to our endeavour: The purpose is noncommercial academic research conducted in good faith; the data was accessed through publicly available channels, though the legal status of the content hosted on shadow libraries remains contested; the amount of information extracted from each book is minimal (metadata, hash signatures), serving a purely bibliographic purpose; and the resulting catalogue neither substitutes for any book we may have accessed nor affects their potential market.

Data Availability

The datasets generated and/or analysed during the present study are available in the Zenodo public data repository.[5]

Acknowledgements

The authors thank Céline Castets-Renard and Benoît de Courson for their valuable input on this research. This work was funded in part thanks to the support of PRAIRIE-PSAI (Paris Artificial intelligence Research institute—Paris School of Artificial Intelligence, reference ANR-22-CMAS-0007).

Competing Interests

The authors have no competing interests to declare.

Works Cited

Andrea Bartz, Charles Graeber, and Kirk Wallace Johnson v. Anthropic PBC. 2025, No. C 24-05417 WHA (ND Cal).

Antoniak, Maria, David Mimno, Rosamond Thalken, Melanie Walsh, Matthew Wilkens, Gregory Yauney. "The Afterlives of Shakespeare and Company in Online Social Readership." *Journal of Cultural Analytics*, vol. 9, no. 2, 2024, arXiv:2401.07340. https://doi.org/10.22148/001c.116919.

Authors Guild v. Google, Inc. 2015, 804 F.3d 202 (2d Cir. 2015).

Bourrier, Karen. "The Social Lives of Books: Reading Victorian Literature on Goodreads." *Journal of Cultural Analytics*, vol. 5, no. 1, 2020. https://doi.org/10.22148/001c.12049.

BrightData. "Goodreads-Books." *Hugging Face*, 2024. https://huggingface.co/datasets/BrightData/Goodreads-Books.

Centivany, Alissa. "The Dark History of HathiTrust." *Proceedings of the 50th Hawaii International Conference on System Sciences*, 2017. https://doi.org/10.24251/HICSS.2017.285.

Christenson, Heather. "HathiTrust: A Research Library at Web Scale." *Library Resources & Technical Services*, vol. 55, no. 2, 2011, pp. 93–102. https://doi.org/10.5860/lrts.55n2.93.

Downie, J. Stephen, Sayan Bhattacharyya, Francesca Giannetti, Eleanor Dickson Koehl, and Peter Organisciak. "The HathiTrust Digital Library's Potential for Musicology Research." *International Journal on Digital Libraries*, vol. 21, no. 4, 2020, pp. 343–58. https://doi.org/10.1007/s00799-020-00283-7.

European Parliament and Council of the European Union. "Directive (EU) 2019/790 on Copyright and Related Rights in the Digital Single Market," 2019. eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32019L0790. Accessed 10 Mar. 2026.

Feist Publications, Inc. v. Rural Telephone Service Co. 499 U.S. 340. 1991.

hiQ Labs, Inc. v. LinkedIn Corp. 2022, 31 F.4th 1180 (9th Cir. 2022).

[5] Antoine Mazières, and Thierry Poibeau, "MajinBook: An Open Catalogue of Digitally Mediated World Literature." *Zenodo*, 14 Nov. 2025, doi.org/10.5281/zenodo.17609566.

Hu, Yuerong, Jana Diesner, Ted Underwood, Zoe LeBlanc, Glen Layne-Worthey, and John S. Downie. "Who Decides What Is Read on Goodreads? Uncovering Sponsorship and Its Implications for Scholarly Research." *Big Data & Society*, vol. 12, no. 3, 2025. https://doi.org/10.1177/20539517251359229.

IFLA Study Group, "Functional Requirements for Bibliographic Record." *International Federation of Library Associations and Institutions*, 1998. https://repository.ifla.org/handle/20.500.14598/830. Accessed 10 Mar. 2026.

Kadrey v. Meta Platforms, Inc., No. 23-cv-03417-VC. United States District Court, Northern District of California.

Karaganis, Joe, ed. *Shadow Libraries: Access to Knowledge in Global Higher Education*. MIT Press, 2018. https://doi.org/10.7551/mitpress/11339.001.0001.

Kousha, Kayvan, M.Thelwall, and M. Abdoli. "Goodreads Reviews to Assess the Wider Impacts of Books." *Journal of the Association for Information Science and Technology*, vol. 68, no. 8, 2017, pp. 2004–16. https://doi.org/10.1002/asi.23805.

Lu, Haoyu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun et al. "Deepseek-Vl: Towards Real-World Vision-Language Understanding," 2024. arXiv:2403.05525.

McEnery, Tony, and Andrew Hardie. "The History of Corpus Linguistics." *The Oxford Handbook of the History of Linguistics*, edited by Keith Allan, Oxford UP, 2013, pp. 727–45. https://doi.org/10.1093/oxfordhb/9780199585847.013.0034.

Michel, Jean-Baptiste, Yuan K. Shen, Aviva P. Aiden, Adrian Veres, Matthew K. Gray, Joseph P. Pickett et al. "Quantitative Analysis of Culture Using Millions of Digitized Books." *Science*, vol. 331, no. 6014, 2011, pp. 176–82. https://doi.org/10.1126/science.1199644.

Moretti, Franco. "The Slaughterhouse of Literature." *Modern Language Quarterly*, vol. 61, no. 1, 2000, pp. 207–27. https://muse.jhu.edu/article/22852.

Murdock, Jaimie, Colin Allen, Katy Börner, Robert Light, Simon McAlister, Andrew Ravenscroft, Robert Rose et al. "Multi-Level Computational Methods for Interdisciplinary Research in the HathiTrust Digital Library." *PloS One*, vol. 12, no. 9, 2017. https://doi.org/10.1371/journal.pone.0184188.

Sag, Matthew, and Peter K. Yu. "The Globalization of Copyright Exceptions for AI Training." *Emory Law Journal*, vol. 74, no. 5, 2025, p. 1163. https://scholarlycommons.law.emory.edu/elj/vol74/iss5/4.

Underwood, Ted. *Distant Horizons: Digital Evidence and Literary Change*. University of Chicago Press, 2019. https://doi.org/10.7208/chicago/9780226612973.001.0001.

Underwood, Ted, David Bamman, and Sabrina Lee. "The Transformation of Gender in English-Language Fiction." *Journal of Cultural Analytics*, vol. 3, no. 2, 2018. https://doi.org/10.22148/16.019.

Walsh, Melanie, and Maria Antoniak. "The Goodreads 'Classics': A Computational Study of Readers, Amazon, and Crowdsourced Amateur Criticism." *Journal of Cultural Analytics*, vol. 6, no. 2, 2021. pp. 243–87. https://doi.org/10.22148/001c.22221.

Wan, Mengting, and Julian J. McAuley. "Item Recommendation on Monotonic Behavior Chains." *Proceedings of the 12th ACM Conference on Recommender Systems*, edited by Sole Pera, Association for Computing Machinery, 2018, pp. 86–94. https://doi.org/10.1145/3240323.3240369.

Wan, Mengting, Rishabh Misra, Ndapa Nakashole, and Julian J. McAuley. “Fine-Grained Spoiler Detection from Large-Scale Review Corpora.” *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, edited by Aaron Halfaker, Loren Terveen, and Fabio Crestani, Association for Computational Linguistics, 2019, pp. 2605–10. https://doi.org/10.18653/v1/P19-1248.