MajinBook

An open catalogue of digital world literature with likes

Antoine Mazières* and Thierry Poibeau

Lattice (ENS-PSL, CNRS), Montrouge, France

November 21, 2025

Abstract

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 English-language books spanning three centuries, enriched with first publication dates, genres, and popularity metrics like ratings and reviews. Our methodology prioritizes natively digital EPUB files to ensure machine-readable quality, while addressing biases in traditional corpora like HathiTrust, and includes secondary datasets for French, German, and Spanish. We evaluate the linkage strategy for accuracy, release all underlying data openly, and discuss the project's legal permissibility under EU and US frameworks for text and data mining in research.

Keywords: Book corpus; Shadow Libraries; Goodreads; Computational Social Science.

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 [1], 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–utilising millions of books and social media data–and in tools–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. The scanning effort undertaken by Google, beginning in 2002, culminated in the Google Books project and yielded, among other resources, the dataset for a foundational 'quantitative analysis of culture' [2]. Shortly thereafter, the HathiTrust Digital Library [3] 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 making 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 [4], literary history [5], gender studies [6], and musicology [7].

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, non-affiliated 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' [8] 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 centre'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. These combined issues of representational bias, data quality, and the constraints of a closed research environment highlight the challenges of using even the largest institutional corpora to study culture.

*Corresponding author: mazr@ik.me

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 crowdsourced cultural curation, making them an essential new source of data for the computational social sciences and humanities [9]. The recent development of Large Language Models (LLMs) has opened a 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 [10] and court cases [11, 12], 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 (Sec. 1), and the social network Goodreads (Sec. 2). We then detail our linkage strategy, its evaluation and outcome (Sec. 3). Finally, we conclude our paper with some legal considerations about our endeavour (Sec. 4).

1 Shadow Libraries

1.1 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 Library Genesis (LibGen) [13], 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 operational 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 LibGen's but is separately administered and arguably less open [9]. 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 [14], 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.

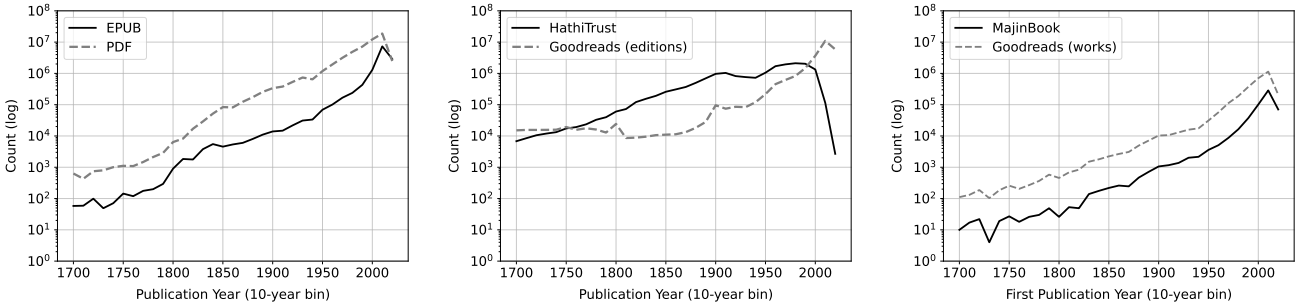
Combining the metadata of Z-Library as of January 2025 and Libgen as of March 2025 yields a list of 77,567,282 items, of which the vast majority are declared as PDF (77.9%) while 19.2% are referenced as EPUB[1]. The remaining 3.5% comprise various formats spanning from raw text files (`.txt`, `.rtf`) to word processing extensions (`.doc`, `.odf`), along with more exotic types given the context, such as `.exe` or `.iso`–all of which were discarded.

1.2 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 web pages–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 re-issued in a modern digital format. Fig. 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,

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



(a) Shadow Libraries (b) HathiTrust & Goodreads (editions) (c) MajinBook & Goodreads (works)

Figure 1: **Temporal distributions** and biases of key corpora.

The figure illustrates 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* PDF*s and confirms that MajinBook (c) is a representative temporal sample of its source.*

the EPUB subset exhibits a convex shape, signalling a super-exponential growth rate that accelerates towards the present.

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, natively *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 study. We are not claiming to represent 'world literature' in its historical entirety, but rather the digitally-mediated canon—culture as it is curated, circulated, and consumed in the 21st century. This temporal bias towards natively digital and commercially viable reissues is not necessarily a flaw, but a defining characteristic of our corpus.

This contemporary constraint yields a significant and somewhat counter-intuitive benefit. As Table 1 shows, the EPUB subset is far more linguistically diversified (Herfindhal 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 just 48% in our corpus (Table 1)–can be seen as a positive indicator of cultural breadth, enhancing the dataset's representativeness for cultural analytics.

2 Goodreads

Goodreads [15], a prominent social reading platform launched in 2006 and acquired by Amazon in 2013, 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 [16–20]. Nevertheless, these studies acknowledge limitations, including demographic biases within the user base (predominantly white, female, and US-centric) and the potential for review manipulation.

We first detail our data requirements and the rationale for our choice of source (Sec. 2.1), before describing our crawling methodology and its outcome (Sec. 2.2).

2.1 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) [21], 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

	Shadow Libraries		HathiTrust	Goodreads
	EPUB	*PDF*		*Editions*
N. 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
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 homogenous (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.*

(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.

In practice, we adopted the internal *work-edition* mappings provided by Goodreads. These mappings form the core structure of its platform and proved to have very few inconsistencies. On the dataset described below (Sec. 2.2), 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 other specific criteria. First and foremost, in the context of books, the *genres* are a common classification enabling thematic differentiation. The precise internal classification mechanism for genres on Goodreads is unclear, but the 'genres' section associated with most books is established at the work level, i.e., all editions of a given work share the same genres. These categories seem to be the product of a top-down curation effort applied to the platform's crowd-sourced 'shelves' that enable users to organise their books. Some shelves such as 'fiction' or 'romance' can appear in the 'genres' section of a book page, while some do not, such as 'i-own' or 'to-read'.

Popularity is another common corpus selection criterion. Indeed, the very inclusion of a work in any corpus *represents* some kind of selection against obscurity, an escape from what Moretti termed 'The Slaughterhouse of Literature' [22]. 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 more crowd-sourced. Regardless of the base corpus, having further popularity metrics can prove useful to differentiate between popular and less popular books. Goodreads as a social network offers typical features, such as ratings and reviews. Integrating these features into our catalogue offers significant analytical potential, allowing researchers both to study 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. Popularity metrics allow a researcher to create an affordably-sized sub-corpus by sampling the most popular works, while still preserving the catalogue's thematic and temporal representativeness.

Goodreads v. 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 [23], a non-profit initiative of the Internet Archive, 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 is 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 21st-century literary culture as it is mediated by a major commercial platform–that is, *with likes*.

2.2 The crawl of Goodreads

To harvest data from `goodreads.com`, we gathered 1,225,390 editions to serve as seeds. These were drawn from the platform's various public book aggregations namely 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

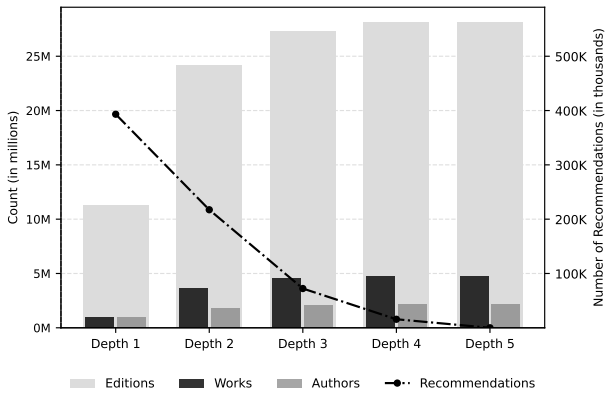


Figure 2: **The crawl of Goodreads**: Item acquisition and recommendation decay.

The figure illustrates the efficiency of our crawl methodology. 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.

those featured in the 'Readers also enjoyed' section), as well as other relevant works by the authors already gathered.[2]

This iterative crawl continued until the author-led discovery of new items began to plateau at depth four and ceased at depth five, along with the complete exhaustion of the recommendation-led crawl. Both concomitant terminations suggest that our dataset is a comprehensive representation of Goodreads' discoverable content, which resulted in our final dataset of 4,778,124 works, 28,105,913 editions, and 2,150,522 authors (cf. Fig. 2). To the best of our knowledge, and based on our review of the state of the art [20, 24–27], this represents one of the largest publicly available crawls of Goodreads data published to date.

3 MajinBook

3.1 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 (Sec. 1.2), we harvested all downloadable files. These were 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 (< 10KB, ca. 4 A4 pages) or too large (> 10MB, 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 (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.

3.2 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 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.

[2] To avoid the long tail of less relevant items, we applied a threshold for an author's other works to be considered; they must have at least one rating and two editions.

[3] We used the Python library *datasketch*, which automatically computed the optimal parameters of 9 bands and 13 rows to find pairs at this similarity level.

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 (Fig. 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. These items were not determined to be incorrect, but were rather *un-evaluatable* by our lab colleagues, as 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 correct matches (91.7% precision). The 2 errors were not random: one was a single book matched to a 3-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.

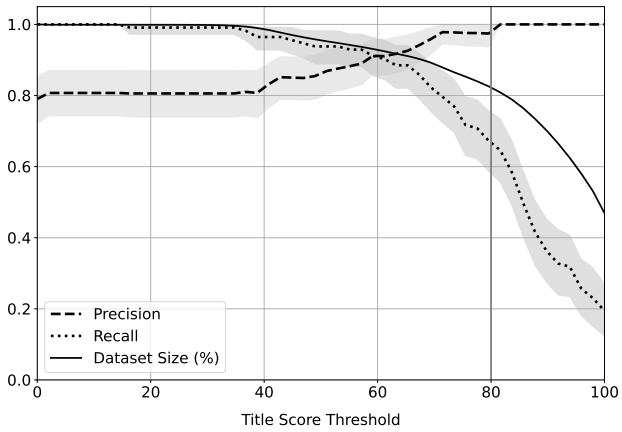


Figure 3: **Precision-recall trade-off for book matching** based on the title score threshold.

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.

3.3 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 retained significant volumes of content in other languages with fair temporal distributions (Fig. 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 (Sec. 3.2).

While this similar distribution is a promising heuristic, it is not a substitute for formal validation. Based on

[4] More precisely, 5 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].

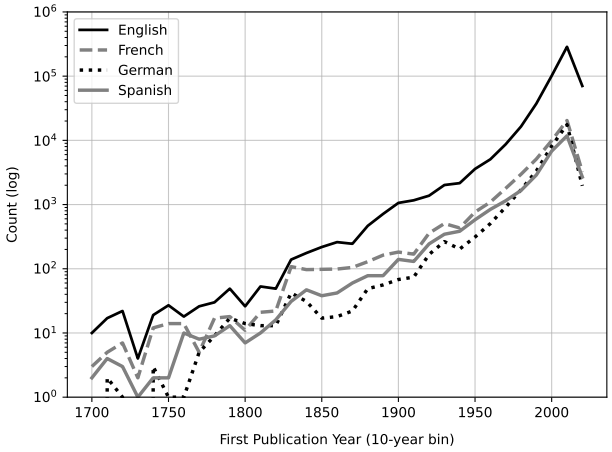


Figure 4: **Temporal distribution** of primary (English) v. secondary datasets.

these two criteria—significant volume and a promising title score distribution—we selected the three of the largest non-English corpora for release: namely French (47,960 items), German (35,559), and Spanish (30,169). 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.

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 speculate this is not an artefact of our matching but rather reflects a more sparse metadata structure for non-Anglophone works within Goodreads itself.

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.

4 Legal and ethical considerations

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 [28]. The legality of using shadow libraries is a central point of contention in recent lawsuits [11, 12] and remains a contested, evolving issue. Jurisprudence continues to refine the relevant criteria for balancing these interests: the presence of *social value* from contributions to public research, the absence of *market impact*, and the *fair use* of a transformative application, to name only a few.

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 shadow libraries). We argue that while the legal justification for the process requires a nuanced, research-specific interpretation of European and American law, the legal security of our final product is unambiguous.

MajinBook's final output–a bibliographic index–is legally secure under the controlling US Supreme Court ruling in *Feist Publications, Inc. v. Rural Telephone Service Co.* [29]. This case established that facts (such as the titles, authors, and publication dates in our catalogue) are not copyrightable. This same conclusion holds under the European Union's Database Directive (96/9/EC) [30].

MajinBook's process–harvesting and computing data from shadow libraries–can be considered through the lens of several legal frameworks: In the EU, the Article 3 of the Directive on Copyright in the Digital Single Market (2019/790) [31], and in the US, the *fair use* doctrine incorporated into the Copyright Act of 1976 (17 USC § 107) [32] complemented by the recent Text and Data Mining exemption to the Digital Millennium Copyright Act (DMCA) [33]. 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 non-legal 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 situates in a public research institution, acting in good faith with no commercial interest and with the sole aim to foster public knowledge about language, literature and cultural representations.

Data availability

The datasets generated and/or analysed during the present study are available in the Zenodo public data repository: doi.org/10.5281/zenodo.17609566.

The detailed description of the datasets' structure and metadata schemas is available on GitHub: github.com/mazieres/MajinBook.

Third-party integration

To facilitate the diachronic analysis of these datasets, the books referenced in our primary English and secondary language corpora were integrated into the Gallicagram platform[5] by its developers. Gallicagram is an n-gram viewer designed to explore term frequencies across a wide collection of French and international datasets [34]. The tool also enables programmatic access to its underlying n-gram data via an API.

[5] https://gallicagram-react.vercel.app/

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).

References

[1] T. McEnery and A. Hardie, "The history of corpus linguistics," in *The Oxford handbook of the history of linguistics* (K. Allan, ed.), Oxford University Press, 2013.

[2] J.-B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, G. B. Team, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, *et al.*, "Quantitative analysis of culture using millions of digitized books," *science*, vol. 331, no. 6014, pp. 176–182, 2011.

[3] H. Christenson, "Hathitrust. a research library at web scale," *Library Resources & Technical Services*, vol. 55, no. 2, pp. 93–102, 2011.

[4] J. Murdock, C. Allen, K. Börner, R. Light, S. McAlister, A. Ravenscroft, R. Rose, D. Rose, J. Otsuka, D. Bourget, *et al.*, "Multilevel computational methods for interdisciplinary research in the hathitrust digital library," *PloS one*, vol. 12, no. 9, p. e0184188, 2017.

[5] T. Underwood, *Distant horizons: digital evidence and literary change*. University of Chicago Press, 2019.

[6] T. Underwood, D. Bamman, and S. Lee, "The transformation of gender in english-language fiction," *Journal of Cultural Analytics*, vol. 3, no. 2, 2018.

[7] J. S. Downie, S. Bhattacharyya, F. Giannetti, E. D. Koehl, and P. Organisciak, "The hathitrust digital library's potential for musicology research," *International Journal on Digital Libraries*, vol. 21, no. 4, pp. 343–358, 2020.

[8] A. Centivany, "The dark history of hathitrust," in *Proceedings of the 50th Hawaii International Conference on System Sciences*, p. 1, 2017.

[9] J. Karaganis, *Shadow libraries: Access to knowledge in global higher education*. The MIT Press, 2018.

[10] H. Lu, W. Liu, B. Zhang, B. Wang, K. Dong, B. Liu, J. Sun, T. Ren, Z. Li, H. Yang, *et al.*, "Deepseek-vl: towards real-world vision-language understanding," *arXiv preprint arXiv:2403.05525*, 2024.

[11] Richard Kadrey, et al. v. Meta Platforms, Inc., 2025. Case No. 23-cv-03417-VC (ND Cal).

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

[13] "Library genesis." https://libgen.rs/, 2008. Accessed: 2025-06-06.

[14] "Anna's archive." https://annas-archive.org/, 2022. Accessed: 2025-06-06.

[15] O. Chandler and E. Khuri, "Goodreads." https://www.goodreads.com/, 2006. Accessed: 2025-06-06.

[16] M. Walsh and M. Antoniak, "The goodreads "classics": a computational study of readers, amazon, and crowdsourced amateur criticism," *Journal of Cultural Analytics*, vol. 6, no. 2, pp. 243–287, 2021.

[17] M. Antoniak, D. Mimno, R. Thalken, M. Walsh, M. Wilkens, and G. Yauney, "The afterlives of shakespeare and company in online social readership," *arXiv preprint arXiv:2401.07340*, 2024.

[18] K. Bourrier and M. Thelwall, "The social lives of books: Reading victorian literature on goodreads," *Journal of Cultural Analytics*, vol. 5, no. 1, 2020.

[19] K. Kousha, 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, pp. 2004–2016, 2017.

[20] Y. Hu, J. Diesner, T. Underwood, Z. LeBlanc, G. Layne-Worthey, and J. S. Downie, "Who decides what is read on goodreads? uncovering sponsorship and its implications for scholarly research," *Big Data & Society*, vol. 12, no. 3, p. 20539517251359229, 2025.

[21] I. F. of Library Associations and Institutions, "Functional requirements for bibliographic records." https://repository.ifla.org/handle/20.500.14598/830, 1998.

[22] F. Moretti, "The slaughterhouse of literature," *MLQ: Modern Language Quarterly*, vol. 61, no. 1, pp. 207–227, 2000.

[23] A. Swartz, B. Kahle, A. Rossi, A. Chitipothu, and R. Hargrave Malamud, "Openlibrary." https://openlibrary.org/, 2006. Accessed: 2025-06-06.

[24] M. Wan and J. J. McAuley, "Item recommendation on monotonic behavior chains," in *Proceedings of the 12th ACM Conference on Recommender Systems, RecSys 2018, Vancouver, BC, Canada, October 2-7, 2018* (S. Pera, M. D. Ekstrand, X. Amatriain, and J. O'Donovan, eds.), pp. 86–94, ACM, 2018.

[25] M. Wan, R. Misra, N. Nakashole, and J. J. McAuley, "Fine-grained spoiler detection from large-scale review corpora," in *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers* (A. Korhonen, D. R. Traum, and L. Màrquez, eds.), pp. 2605–2610, Association for Computational Linguistics, 2019.

[26] "Goodreads book datasets with user rating 2m." https://www.kaggle.com/datasets/bahramjannesarr/goodreads-book-datasets-10m, 2020. Accessed: 2025-07-10.

[27] "Goodreads books." https://huggingface.co/datasets/BrightData/Goodreads-Books, 2024. Accessed: 2025-07-10.

[28] M. Sag and P. K. Yu, "The globalization of copyright exceptions for ai training," *Emory LJ*, vol. 74, p. 1163, 2024.

[29] "Feist Publications, Inc. v. Rural Telephone Service Co.." 499 U.S. 340, 1991. United States Supreme Court.

[30] European Parliament and Council, "Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases." OJ L 77, p. 20, 1996.

[31] Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019, "on copyright and related rights in the digital single market and amending directives 96/9/ec and 2001/29/ec," 2019. Document 32019L0790.

[32] "Copyright act of 1976," 1976. 17 U.S.C. § 107 (Fair Use).

[33] "37 C.F.R. § 201.40 — "exemptions to prohibition against circumvention"." Electronic Code of Federal Regulations, 2023. Accessed: 2025-10-20.

[34] B. Azoulay and B. De Courson, "Gallicagram: un outil de lexicométrie pour la recherche," *SocArXiv, [sl]*, vol. 8, 2021.