Canons in the Shadows–A Critical Catalogue of AI's Unseen Reading List

Antoine Mazières[1], and Thierry Poibeau[1]

[1] Lattice, ENS-PSL, Montrouge, France

Abstract

This *data paper* details a work-in-progress to construct a large-scale, open book catalogue to address the critical need to understand the "shadow libraries" increasingly used to train large language models. We describe the aggregation and planned unification of tens of millions of bibliographic records from Library Genesis, Z-Library, OpenLibrary, and Goodreads. The paper outlines our methodology for tackling poor-quality metadata through a unified "work/edition/author" data model, cross-source validation, and heuristic-based enrichment. The resulting metadata-only catalogue will provide a novel resource for computational humanities research and enable a critical audit of opaque AI training corpora. We conclude with a legal analysis justifying our approach under EU and US law.

Keywords: shadow libraries, bibliographic metadata, critical data studies, computational humanities

1 Introduction

The increasing reliance on large-scale textual corpora for training and evaluating language models has brought renewed attention to the origins, composition, and ethical implications of the datasets involved. Among the most prominent–yet controversial–are the so-called "shadow libraries" such as Library Genesis (LibGen) and Z-Library [12]. They aggregate vast collections of texts that include both works in the public domain and copyrighted material distributed without authorization, a practice that raises serious ethical and legal concerns. Nevertheless, their immense scale and breadth have made them an appealing, if unofficial, resource for various actors, including major technology companies. While most avoid publicly acknowledging such use, the company DeepSeek[16] notably cites them as sources of training data, while recent court cases have revealed similar uses by Meta [17] and Anthropic [2].

This situation presents a dual imperative for the research community. Critically, it is essential to audit the contents of these shadow libraries to understand the biases and potential liabilities embedded in the models they help create. Pragmatically, these collections, which contain substantial public domain material, represent an invaluable resource for computational humanities research. To address both needs, this paper details the creation of a large-scale, open, metadata-only catalogue aggregating records from LibGen, Z-Library, OpenLibrary, and Goodreads. The core technical contribution lies in resolving heterogeneous and poor-quality metadata through a unified data model and robust record-linkage. The resulting corpus will provide a foundational resource for critically examining AI's unseen reading lists and for enabling new large-scale textual analyses.

The remainder of this article is organized as follows.

Antoine Mazières, and Thierry Poibeau. "Canons in the Shadows–A Critical Catalogue of AI's Unseen Reading List." In: *Computational Humanities Research 2025*, ed. by Taylor Arnold, Margherita Fantoli, and Ruben Ros. Vol. 3. Anthology of Computers and the Humanities. 2025, 1384–1393. https://doi.org/10.63744/jLwrgN3hnrVS.

Section 2 details our approach to data integrity and the validation of the core features of our final dataset. Then, in Section 3, we describe the gathering and processing of each source, commenting on their respective scopes and overall quality. Finally, Section 4 addresses the legal and ethical landscape surrounding the use of shadow libraries in academic research.

This article reports on a work in progress, with the main data compilation and cleaning expected to conclude by the end of 2025. As such, several key methodological steps are described here as ongoing or planned. Specifically, the final merging of the datasets and the comprehensive cross-referencing required to produce our final catalogue remain to be completed, and a detailed account of these stages is therefore outside the scope of the present document.

Ultimately, the catalogue aims to reposition questions of AI training data within the broader concerns of the humanities: authorship, canonicity, and the circulation of knowledge. By making the bibliographic skeleton of these corpora visible, it allows for a genuinely critical engagement with what machines read—and, by extension, with how they learn to speak.

Upon completion, the catalogue described here will be made available through a public academic data repository, along with code snippets intended to ensure reproducibility and encourage further improvement of our methods.

2 Data validation and integrity

	OpenLibrary	**Goodreads**	**LibGen** (epub)	**Z-Library** (epub)
Editions	53,576,739	28,105,913	3,032,730	8,097,488
Title	99.9%	99.9%	99.7%	99.9%
Pub. year	94,4%	88.7%	83.1%	80.1%
Language	85.1%	86.2%	99.8%	66.5%
ISBN	66.8%	62.3%	56.7%	28.9%
Works	39,105,972	4,778,124	—	—
Title	99.9%	—	—	—
First pub. year	11.33%	60.9%	—	—
Authors	14,139,728	2,150,522	—	—
Name	99.9%	100%	—	—

Table 1: Core metadata availability.
Values show the percentage of items (e.g., Editions, Works) in each dataset (column) possessing a given feature (row). Sub-headings indicate the item type and total count.

As stated in the introduction, a primary challenge in using shadow libraries is the poor quality of their metadata. This metadata is often the product of long-term, sedimented efforts, mixing manual input from users with automated processes that extract information from the content itself or import it from other sources [5]. To some extent, the same is true of platforms like Open Library, Goodreads, and Wikipedia that are–at least partially–crowd-sourced. Consequently, one cannot assume the integrity of a given field.

We therefore subjected all data gathered for this project to a rigorous *data validation* step, ensuring that the variables of interest are presented to the end user in a predictable and reliable manner.

Our data validation process prioritised four key objectives. Our first and most critical goal was to reliably map each book with its various editions and authors (Sec. 2.1). Secondly, we aimed to

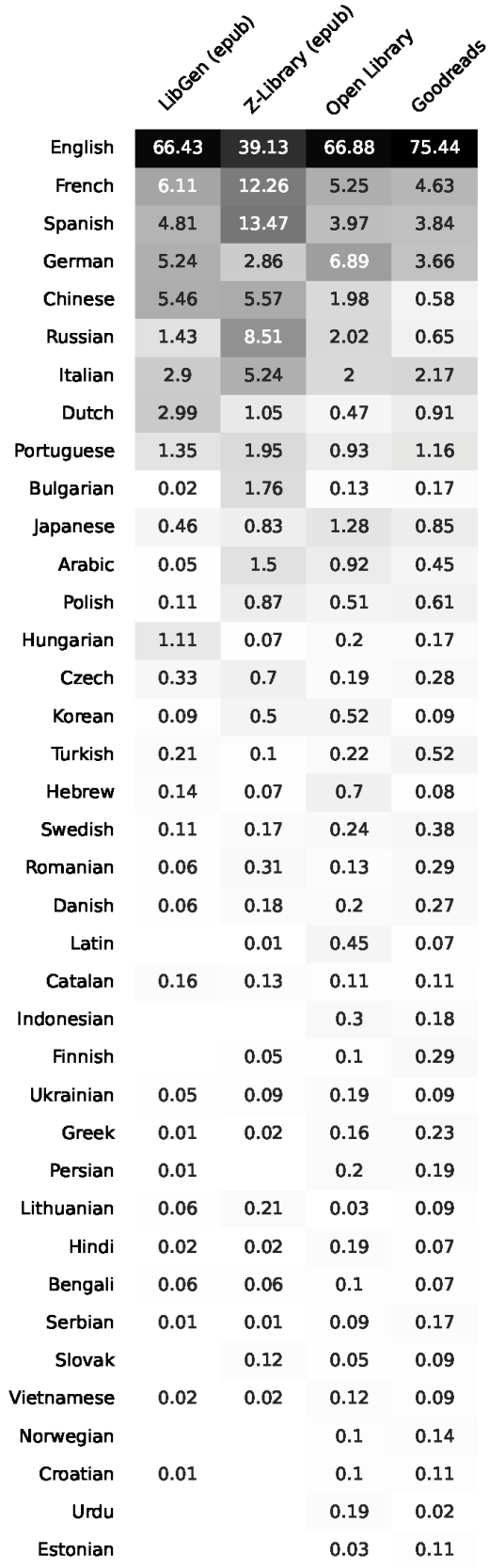

Figure 1: Language distribution across datasets.

The heatmap compares the language distributions of the four source corpora, with each cell showing the percentage of a given language within a dataset. The data reveals a profound bias towards English. Each source exhibits also distinct secondary biases; for instance, Z-Library (PDFs excluded) has significant concentration of Russian materials, while Open Library and Goodreads feature a broader "long tail" of many other languages.

determine a specific point of origin for each work by identifying its first publication date (Sec. 2.2). We also sought to apply a universal characterisation of the narrative mode for every book in the corpus, classifying each as either fiction or non-fiction (Sec. 2.3). Finally, this section will report on our efforts to preserve and integrate the remaining metadata into our final, consolidated catalogue (Sec. 2.4).

2.1 Work, authors, and editions

We use the term *work* to refer to the abstract text and *edition* to refer to a specific published version. We adopt this typology here, as it is used by data sources central to our study, such as Open Library and Goodreads, which serve as the main sources for building this scaffold. To some extent, the edition is the object and the work is the idea; and to study the latter, one must necessarily go through the former.

While we inherit this typology from these platforms, it is partly related to a more advanced classification system: the *Functional Requirements for Bibliographic Records* (FRBR) [14], 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 OCLC's clustering of bibliographic records into *Work* entities within WorldCat, although it operates exclusively on openly accessible data sources. In practice, we rely primarily on the internal *work-edition* mappings provided by Open Library and Goodreads, complemented by heuristic matching of titles, authors, and identifiers (ISBN, ASIN) to harmonise records across sources.

While the concept of an *author* may seem more straightforward, its application in bibliographic data can be complex. There can be as many variations of authorship as there are editions, since translators, illustrators, editors, or writers of prefaces can sometimes be listed as authors. To manage this complexity, we use *author* as a general term for any of these contributory roles. We will, however, use the specific term *main author* to refer to the original writer of a *work*.

2.2 First publication date

Establishing a work's first publication date is a primary objective of our cataloguing effort. This date serves as a crucial anchor for any diachronic analysis, allowing researchers to situate a work in its original historical and literary context. However, this information is often absent or inconsistent in the source metadata. Shadow library records, in particular, frequently provide only the publication date of a specific digital *edition*, which is often recent, rather than the first publication date of the original *work*.

Addressing this challenge will hinge on the cross-referencing effort previously outlined. We will use OpenLibrary, Goodreads, and Wikipedia as our principal sources to find and cross-validate a work's first publication date. Furthermore, by creating a high-confidence subset of works with validated dates, we can develop and test various heuristics to infer the first publication date for records where it is missing.

2.3 Narrative mode: Fiction and non-fiction

While the distinction between fiction and non-fiction is a historical construct, this classification is a fundamental feature for any contemporary book catalogue. In computational analysis, it enables a vast range of comparative studies, from tracking the evolution of literary genres to examining trends in scientific discourse.

We therefore sought to establish this classification as a universal feature across our entire dataset. The task is more tractable than determining the first publication date, as our various data sources allude to this classification through descriptive fields such as genres, keywords, and subject headings [19]. LibGen itself provides a preliminary fiction or non-fiction flag to all its data. Our method, therefore, relies on a straightforward cross-validation of these signals. By developing simple heuristics to resolve conflicting or ambiguous classifications we aim to assign a narrative mode to the entirety of works in our corpus.

2.4 Other metadata

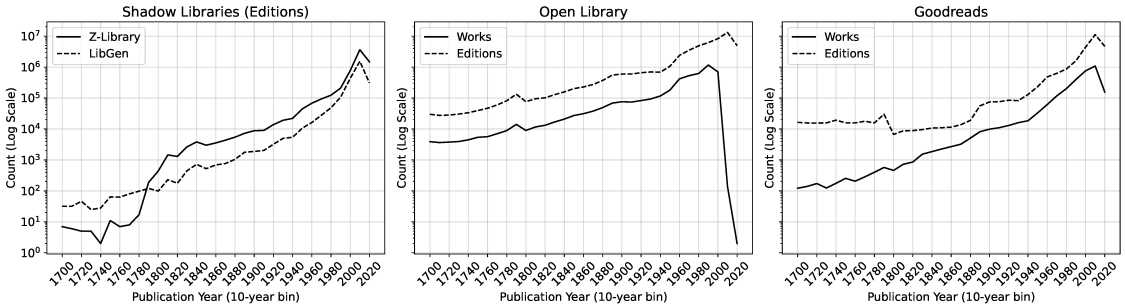

Figure 2: Temporal distribution of works and editions across data sources.

This figure highlights the different temporal characteristics of the datasets. The leftmost plot shows the number of editions *from LibGen and Z-Library (PDFs excluded), revealing a collection heavily skewed towards materials published after 1980. In contrast, the plots for Open Library (centre) and Goodreads (right) distinguish between the timeline of abstract* works *(solid line, based on first publication date) and their constituent* editions *(dashed line).*

Table 1 summarises the coverage of what we define as *core features*. These are the features we expect to successfully normalise across the vast majority of our final dataset and which–along with Title and Author–have been described in previous subsections. Beyond these, each source provides a rich set of additional features, from content descriptors like keywords and genres, to physical details like format and page count. These *secondary features* exhibit the full gamut of data quality one might expect from crowd-sourced efforts; as discussed previously, their lack of consistency makes them an unsuitable focus for our primary validation endeavour.

However, these secondary features could still serve an important purpose. Our approach is to include them in the final catalogue after our standard validation process, while clearly flagging them as user-generated or of lower confidence. This strategy empowers researchers to leverage this rich information while being mindful of its nature. For instance, while the millions-of-items-strong "contemporary romance" genre tag from Goodreads may prove useful for sampling recent works, a less consensual descriptor like "tortured hero"–with its approximately 15,000 items–may introduce significant bias when trying to identify a specific narrative pattern. Ultimately, we leave the final consideration of these features' relevancy to the individual researcher.

3 Data sources, data sieves

3.1 Shadow libraries

The shadow libraries that form the basis of our corpus have distinct but overlapping histories. The oldest and most foundational is Library Genesis (LibGen) [15], which was started around 2008 by Russian scientists rooted in the clandestine "samizdat" culture of the Soviet era.

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 [12].

The most recent development in this ecosystem is Anna's Archive [3], which appeared in 2022 and aims to provide a comprehensive, searchable index and mirror of other shadow libraries, including LibGen and Z-Library.

These extensive corpora of books come in two main formats: PDF and e-book (e.g. EPUB, AZW, MOBI). The PDF format is highly varied, comprising everything from partial amateur scans to well-indexed official publisher versions. However, consistently parsing their content to extract clean, raw text remains a significant technical challenge. To circumvent these manifold issues, we made the methodological decision to discard all PDF files, which represents an approximate cut of 80% of both shadow library datasets. We acknowledge that this decision introduces a significant bias, favouring more recent content and older books that were deemed commercially viable enough to be re-issued in a modern digital format. We are still in the process of analysing the precise impact of this decision on the corpus's temporal and language distributions to make this additional bias more explicit.

E-books, on the other hand, are natively digital structures–much like web pages–making their content and metadata more readily machine-readable. By choosing to focus solely on this 20% subset, we can ensure a much higher degree of quality for the underlying text. This process yielded a corpus of approximately 11 million books, which we standardised by converting to a uniform EPUB format. This standardisation allows for the reliable extraction of internal file metadata.

As shown in Figure 1, the language distributions of LibGen and Z-Library differ significantly. Both are strongly skewed towards English content, but Z-Library exhibits a unique profile compared to all other datasets processed here. While in LibGen, OpenLibrary, and Goodreads the proportion of English items is roughly ten times that of the next most frequent language, this ratio in Z-Library is only three-to-one.

In contrast to their linguistic profiles, the temporal distributions (Fig. 2) of the editions the two shadow libraries represent exhibit more similarity. As one might expect, both collections are heavily skewed towards recent publications. However, some distinctions are apparent: Z-Library's collection appears to decline more gradually into the 20th and 19th centuries, while LibGen, conversely, contains almost an order of magnitude more material from the less-represented 18th century.

3.2 OpenLibrary

Open Library [18] is 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".

As a source of metadata for our project, Open Library primary strength is its immense scale and open nature; it provides free, bulk access to one of the largest structured bibliographic datasets in the world, boasting tens of millions of records which cover a rich long tail of lesser represented languages (Fig. 1). This scale, combined with its formal work/edition data model, made it appear as an invaluable scaffold for structuring our own catalogue.

However, the project's greatest strength–its open, wiki-based contribution model–is also its most significant weakness. The metadata is notoriously inconsistent, often containing duplicate records, incomplete or nonsensical entries, and a low coverage rate for crucial analytical features such as a work's first publication date (Table 1). This mediocre data quality is starkly illustrated in Figure 2, where the near-identical temporal distributions for *works* and *editions* up to the 1990s strongly suggest that the "first publication year" field is often being conflated with edition publication dates, raising serious doubts about the reliability of this core feature without external validation. The upcoming cross-referencing steps will be essential in determining the extent to which

Open Library can be reliably used as a data source.

3.3 Goodreads

Goodreads [7], 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 [4; 6; 13; 19]. Nevertheless, these studies acknowledge limitations, including demographic biases within the user base (predominantly white, female, and US-centric) and the potential for review manipulation.

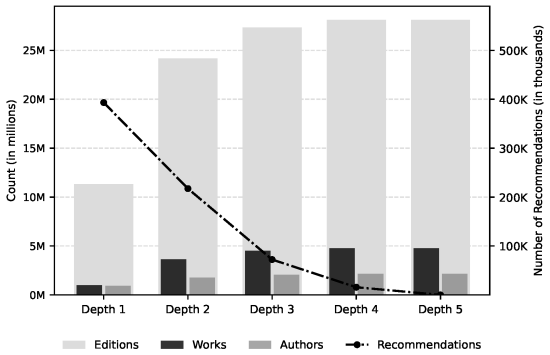

Figure 3: The crawl of Goodreads: Item acquisition and recommendation decay.
The figure illustrates the efficiency of our crawl methodology. The bars show the cumulative counts of Editions, Works, and Authors (left axis, in millions) gathered at each stage. The line plot tracks the number of new Recommendations (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.

To harvest data from `goodreads.com`, we gathered $1{,}225{,}390$ editions to serve as seeds. These were drawn from the platform's various book public aggregations–distinct, albeit partially overlapping–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.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.[1]

This iterative crawl continued until the author-led discovery of new items began to plateau at depth four and ceased at depth five, resulting in our final dataset of $4,778,124$ works, $28,105,913$ editions, and $2,150,522$ authors (cf. Figure 3). The complete exhaustion of the recommendation-led crawl suggests that our dataset is a comprehensive representation of Goodreads' discoverable content. To the best of our knowledge, and based on our review of the state of the art [10; 11; 20; 21], this is the most extensive crawl of Goodreads data ever published.

Overall, Goodreads' data quality is excellent, featuring a full work/editions/authors scaffold structuring almost 5 million works with very few inconsistencies and a 60% coverage for key features such as the first publication date (Table 1). However, its primary weakness is a profound linguistic bias. Figure 1 shows it has the most English-skewed content distribution of all our

[1] 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. We didn't collect reviews.

sources, which confirms a US-centric bias that has been acknowledged in previous research on the platform [19].

Focusing on the temporal distribution of its works (Fig. 2), the Goodreads data demonstrates significant historical depth. The collection maintains a volume of over $10,000$ works per decade until the beginning of the 20th century and then declines by approximately an order of magnitude per century, still retaining around 100 works per decade for the 1700s. Given that we can match this metadata with the editions in the shadow libraries, this historical range makes Goodreads an invaluable source for the diachronic analyses we aim to enable.

4 Legal consideration

Our project's legal standing rests on a distinction between our process (acquiring and analysing data from shadow libraries) and our product (a public, metadata-only catalogue). The analysis differs significantly between the European and American legal frameworks but yields, to our understanding, a favourable conclusion.

In the European Union, our project's methodology appears permissible under the Directive on Copyright in the Digital Single Market (2019/790) [8]. Article 3 of the Directive provides a mandatory copyright exception for Text and Data Mining (TDM) conducted by research organizations for scientific purposes. As a non-commercial, academic endeavour, our project clearly qualifies for this exception. The primary condition is that the works must be "lawfully accessed". The Directive's own guidance indicates that lawful access includes content that is "freely available online" without technical barriers like paywalls. Since the shadow libraries used as our data source are openly accessible on the internet, our process meets this requirement. Therefore, the computational analysis of these data for scientific research is a lawful activity under the EU's TDM framework.

In the United States, legality hinges on the fair-use doctrine, complemented by the recent Text and Data Mining exemption to the Digital Millennium Copyright Act (DMCA) [1]. Recent lawsuits against AI developers, notably *Kadrey v. Meta* [17] and *Bartz v. Anthropic* [2], provide crucial context. A key factor in the fair-use analysis is the "transformative" nature of the use; our project is highly transformative, as we do not use the books to be read, but rather as data from which to extract new information and research insights. The DMCA exemption explicitly recognises the legitimacy of text and data mining by researchers under controlled conditions. This provision reinforces the legality of our workflow, ensuring that our computational extraction of metadata from lawfully accessed sources remains compliant with U.S. copyright law. Moreover, our final output—the metadata-only catalogue—is legally secure: under the United States Supreme Court's ruling in *Feist v. Rural* [9], facts are not copyrightable. The bibliographic data in our catalogue (titles, authors, publication dates) are therefore unprotectable factual information.

Acknowledgements

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] "37 C.F.R. § 201.40 — ''Exemptions to prohibition against circumvention''". Electronic Code of Federal Regulations. Accessed: 2025-10-20. 2023. URL: `https://www.ecfr.gov/current/title-37/chapter-II/subchapter-A/part-201/section-201.40`.

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

[3] “Anna’s Archive”. https://annas-archive.org/. Accessed: 2025-06-06. 2022.

[4] Antoniak, Maria, Mimno, David, Thalken, Rosamond, Walsh, Melanie, Wilkens, Matthew, and Yauney, Gregory. “The afterlives of shakespeare and company in online social readership”. In: *arXiv preprint arXiv:2401.07340* (2024).

[5] Bodó, Balázs. “The Genesis of Library Genesis: The Birth of a Global Scholarly Shadow Library”. In: *Shadow Libraries: Access to Knowledge in Global Higher Education*, ed. by Joe Karaganis. Cambridge, MA: The MIT Press, 2018, pp. 25–51.

[6] Bourrier, Karen and Thelwall, Mike. “The social lives of books: Reading Victorian literature on Goodreads”. In: *Journal of Cultural Analytics 5*, no. 1 (2020).

[7] Chandler, Otis and Khuri, Elizabeth. “Goodreads”. https://www.goodreads.com/. Accessed: 2025-06-06. 2006.

[8] 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”. Document 32019L0790. 2019.

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

[10] “Goodreads Book Datasets With User Rating 2M”. https://www.kaggle.com/datasets/bahramjannesarr/goodreads-book-datasets-10m. Accessed: 2025-07-10. 2020.

[11] “Goodreads Books”. https://huggingface.co/datasets/BrightData/Goodreads-Books. Accessed: 2025-07-10. 2024.

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

[13] Kousha, Kayvan, Thelwall, Mike, and Abdoli, Mahshid. “Goodreads reviews to assess the wider impacts of books”. In: *Journal of the Association for Information Science and Technology 68*, no. 8 (2017), pp. 2004–2016.

[14] Library Associations, International Federation of and Institutions. “Functional Requirements for Bibliographic Records”. https://repository.ifla.org/handle/20.500.14598/830. 1998.

[15] “Library Genesis”. https://libgen.rs/. Accessed: 2025-06-06. 2008.

[16] Lu, Haoyu, Liu, Wen, Zhang, Bo, Wang, Bingxuan, Dong, Kai, Liu, Bo, Sun, Jingxiang, Ren, Tongzheng, Li, Zhuoshu, Yang, Hao, et al. “Deepseek-vl: towards real-world vision-language understanding”. In: *arXiv preprint arXiv:2403.05525* (2024).

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

[18] Swartz, Aaron, Kahle, Brewster, Rossi, Alexis, Chitipothu, Anand, and Hargrave Malamud, Rebecca. “OpenLibrary”. https://openlibrary.org/. Accessed: 2025-06-06. 2006.

[19] Walsh, Melanie and Antoniak, Maria. “The goodreads “classics”: a computational study of readers, Amazon, and crowdsourced amateur criticism”. In: *Journal of Cultural Analytics 6*, no. 2 (2021), pp. 243–287.

[20] Wan, Mengting and McAuley, Julian J. “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*, ed. by Sole Pera, Michael D. Ekstrand, Xavier Amatriain, and John O’Donovan. ACM, 2018, pp. 86–94. DOI: 10.1145/3240323.3240369.

[21] Wan, Mengting, Misra, Rishabh, Nakashole, Ndapa, and McAuley, Julian J. “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*, ed. by Anna Korhonen, David R. Traum, and Lluís Màrquez. Association for Computational Linguistics, 2019, pp. 2605–2610. DOI: 10.18653/V1/P19-1248.