Research with User-Generated Book Review Data: Legal and Ethical Pitfalls and Contextualized Mitigations

Yuerong Hu(✉), Glen Layne-Worthey, Alaine Martaus, J. Stephen Downie, and Jana Diesner

School of Information Sciences, University of Illinois Urbana-Champaign, Champaign, USA
{yuerong2,gworthey,martaus2,jdownie,jdiesner}@illinois.edu

Abstract. The growing quantity of user-generated book reviews has opened up unprecedented opportunities for empirical research on books, reading, and readership. While there is an abundance of literature addressing the legal and ethical use of user-generated and social media data in general, for user-generated book reviews, such discussions have been mostly absent. From a library and information sciences perspective, user-generated book reviews can pose novel challenges because each book reviewer may simultaneously be (1) a presumably anonymous and safe online user; and, (2) an identifiable reader who can suffer real harm, e.g., cyber doxing and personal attack. This user/reader duality can create conflicting recommendations regarding which legal or ethical guidelines to follow. According to our review, potential legal issues include copyright infringement and violations of terms of service/end-user license agreements and privacy rights, while ethical concerns are centered on users' expectations, informed consent, and institutional reviews. This paper reviews (1) potential legal and ethical pitfalls in leveraging user-generated book reviews; and, (2) professional and scholarly references that might serve as useful guidelines to avoid or manage these pitfalls.

Keywords: Book reviews · Digital humanities · Scholarly communication · User-generated content · Social computing · Responsible data science

1 Background and Introduction

Reading is one of the most ubiquitous activities in our daily lives. We have limited knowledge about historical everyday readers and their reading behavior due to a lack of records left by or collected from them [108]. In the last two decades, the increasing availability of user-generated book reviews from online sources[1]

[1] In the context of this paper, user-generated book reviews include not only actual book reviews but also numerical ratings, crowdsourced tags, user-curated book lists, virtual collections of books, graphic content, etc.

I. Sserwanga et al. (Eds.): iConference 2023, LNCS 13971, pp. 163–186, 2023.
https://doi.org/10.1007/978-3-031-28035-1_13

has opened up unprecedented opportunities for computational and empirical research on readerships and everyday reading behavior. Scholars from different fields, e.g., library science, digital humanities, communication studies, and natural language processing, have leveraged such data to examine a variety of topics, such as review classification, social network analyses of readers, impact assessment and sales prediction of books [20,64,69,80,85,99,150,152,163]. With the evolution of book review studies, challenges and limitations have also emerged, ranging from disciplinary divergences (such as reader-orientated theories vs. book-centric models [31,66]) to limitations of the scholarly usability of review corpora (such as review credibility and inclusiveness [66,67,70,73,115,118,168]). This paper asks another insufficiently discussed question that has yet to be fully explored in prior empirical and computational studies of user-generated book reviews: How to best use online book reviews for scholarly research from legal, ethical, and compliance perspectives?

This paper is motivated by two factors: First, while the ethical use of user-generated content and social media data for research purposes has been critically discussed [33,34,46,107,168], contextualized investigations of specific genres remain very much in need. As Crawford and Finn have pointed out, "*social and mobile datasets have limitations that, if not sufficiently understood and accounted for, can produce specific kinds of analytical and ethical oversights*" [29]. In their own research on crisis data, they demonstrated the necessity and potential of this research direction by critically examining (1) what crisis data actually represents; and, (2) how these data were used in crisis research [29]. Other studies that focused on specific datasets and use cases have also shed light on specific research challenges and responsibilities by examining ethical issues that stem from work in specific domains or research contexts [22,35,37,48,50,98].

Following these exemplary studies, we propose to scrutinize the challenge of legal, ethical, and compliant research conduct in the context of user-generated book reviews. We argue for a deeper engagement with these datasets because of the dual role that many book reviewers play as (1) social media producers and content consumers; and, (2) readers. When people voluntarily post book reviews, they also reveal aspects of their reading history, whether they are aware of that or not. As a result, user-generated book reviews, like most social media data, may contain directly or indirectly personally identifiable information[2] [10]. At the same time, similar to library patron data, user-generated book reviews record activities and thoughts that are protected as part of people's intellectual freedom and valuable contributions to the diversity of viewpoints in society [8]. For instance, online book reviewers might express opinions, values, and beliefs, which can be vehement, controversial, or even illegal (e.g., acquiring and reading banned books). Reviewers may also share personal experiences, information about their physical and mental health, and their socio-demographic identities. These types of sensitive information lead to concerns about the legitimacy and ethics of using such data in scholarly research.

[2] For example, book reviews may contain user names that overlap with real names, email addresses, identifying parts of addresses, or workplaces.

Second, we examine the usage of book reviews to further minimize potential risks for reviewers and researchers. In order to ensure library patrons' freedom to read, an unfettered exchange of ideas, and equal access to diverse materials and services, library professionals have long protested against policies that would harm the confidentiality of their patrons' data (e.g., search records, book loans, reference interviews) [6–9,77,78,134]. In practice, many libraries regularly remove circulation records and decline to keep certain patron data in order to protect their patron's privacy from "irresistible government requests" [43,90]. For similar reasons, book reviewers' reading records and opinions also need protection because reviews might be subject to censorship and could be used against those reviewers. However, online book reviews have not been protected or managed like library patron data, possibly because they have not been conceptualized in this way, but rather as reviews of consumer goods. This is problematic because censorship, trolling, scams, and harassment targeted at online book reviewers have increased [83,104]. Disliked online book reviews have led to cyber doxing and personal attacks on individual reviewers from book authors, translators, and the public, both online and offline, around the world [17,83,92,116,128]. For instance, in 2014, a teenage girl in the U.K. was tracked down and assaulted by an author because the girl had left a negative review about one of the author's books on Wattpad[3] [17,117]. Although this horrifying incident was an unexpected result of the review posting itself, without any research involved, researchers need to consider the potential for actual harm when designing their studies and reproducing (or even amplifying) potentially harmful content.

At the same time, researchers might be exposed to professional, institutional, and legal consequences of scraping and analyzing user-generated book reviews, such as copyright infringement, violations of policies and end-user license agreements (EULA)/terms of service (TOS),[4] and conflicts of interest with various stakeholders' policies. Most user-generated book reviews are considered copyrighted material and/or material governed by TOS/EULAs. Some platforms that make a profit with their user-generated book reviews have explicitly forbidden unauthorized third-party use of their data via TOS, which means researchers are expected to acknowledge the potential legal hazards that come with their accessing and using of reviews. Also, for research based on copyrighted data that is not subject to fair use, scholarly use of the data for non-commercial purposes or the public good does not serve as an exemption from the possibility of legal consequences. For example, the HathiTrust[5] was sued by The Authors

[3] Wattpad is a storytelling and social reading platform based in Canada [160].

[4] EULA is a contract between the licensor and the licensee, which establishes the licensee's right to use a proprietary product. TOS refers to a contract between a provider and a user which defines the rules that a user should follow in order to use a service. In our research contexts, we consider them interchangeable terms, as both of them specify the permissions and prohibitions for using the book review platforms' service, products, and/or data.

[5] The HathiTrust is a consortium of several hundred academic libraries that have collaborated (with scanning agencies like Google) to create a massive digital library [15,61].

Guild for copyright infringement because of the use of books scanned by Google [15], and the Internet Archive[6] was sued by major book publishers for "grossly" exceeding what libraries were permitted to do by providing "emergency" access to digital teaching materials during the COVID pandemic [57,146]. These cases are reminders that even for public institutions, it is difficult to manage the legal risks associated with their use of data. We conclude that researchers need to understand how they can access and use user-generated book reviews in ways that protect both their human research subjects and themselves from harm and risks.

Therefore, this paper examines the legitimacy and ethics of leveraging user-generated book reviews in scholarly research. We draw upon library standards and practices in addition to existing scholarly discussions to identify potential pitfalls and solutions. Specifically, we investigate (1) relevant laws; (2) platform policies; (3) user rights and expectations; and, (4) existing research on the ethical use of user-generated data at large. Here are the two primary questions we posit and how we analyze them:

1. **Question**: What does prior research say about compliance and ethical conduct of research that uses user-generated book reviews?
   Analysis: We review 100 research articles that feature empirical analyses of user-generated book review datasets and their creators/users. We collected these references as part of our empirical and computational research on book reviews [25,72,73,93,127].[7] The findings are presented in Sect. 2.
2. **Question**: What factors should researchers consider for assessing the appropriateness of their use of data while minimizing potential risks caused by their research?
   Analysis: We analyze a broader range of literature to understand the norms, regulations, and concerns for employing user-generated content (book reviews included) from the perspective of legislation, platform providers, users, and researchers. The analyses are presented in Sect. 3.

Then, in Sect. 4, we discuss the findings and limitations of our investigation. In Sect. 5, we summarize our research contributions and propose topics for future work. Due to variance in legislation, expectations, and norms for ethics and compliance across place, time, and disciplines, this paper does not provide a comprehensive review of prior research on user-generated book reviews, but is consciously situated primarily in a contemporary, U.S.-centric context. We invite readers to extend our approach to their own disciplinary and local contexts.

[6] The Internet Archive is a large digital library that preserves and provides digitized content to the public [154].

[7] Due to length constraints of this paper, we only discussed some of the articles that we reviewed for this paper. The full list of references is available at https://github.com/Yuerong2/iConference2023appendix/blob/main/iconference2023referencesAppendix.pdf. Our literature review is limited to empirical research on user-generated book reviews based on computational and/or qualitative methods. We did not consider theoretical work on user-generated book reviews without empirical data involved.

2 Literature Review of Computational and Empirical Studies that Use User-Generated Book Reviews

Existing research on user-generated book reviews has investigated a variety of datasets from different sources around the globe and in a variety of languages [115], such as reviews in Chinese [59,64,111,164,165], Dutch [19,86], and German [40,119]. Among these, book reviews in English obtained from Amazon, Goodreads, and LibraryThing [12,20,68,150,152,162,163][8] are most frequently used. Data leveraged include (1) actual review texts, crowdsourced tags, book ratings, rankings, and lists; (2) reviewers' public profiles and networks; (3) forum discussions and social media posts; and, (4) information about book sales and price [4,12,20,32,60,64,68,69,99,139,150,152,162,163].The scale and granularity of previously compiled and referenced datasets vary drastically, ranging from hundreds to millions of records [4,60,115,124,130,152]. For instance, Wan and colleagues scraped 1,378,033 English book reviews for spoiler detection [152], while Tan and He qualitatively compared 200 book reviews in Chinese and English as part of a multi-method analysis on cross-cultural reception [130].

These book review datasets have enabled computational and empirical research in various disciplines, including library and information sciences (LIS) [162,163], digital humanities and cultural analytics [20,85,150], computer supported cooperative work [12], social network analysis [99], computational linguistics [152], recommender systems and marketing [27,151], decision making [64,68,69], etc. In turn, each discipline has brought topics to the research. For instance, LIS scholars have studied reviews through the lenses of crowd cataloging and social tagging [16,24,97,139,149]; citation index and impact assessment [111,153,166,169]; and readers' social networks and activities [110,136–138,162]. Cultural historians and literary scholars have asked questions about the evolution of literary genres, the formation of literary canons, and reception of literary works [20,39,42,127,150]. Marketing, economics, and system scientists have examined the relationship between book reviews and book sales [27,64,99,129]. Natural language processing scholars and computational linguists have built models for review classification (e.g., fake, spoiler, and most helpful reviews) [50,68,141,152], sentiment analysis and opinion mining [69,96], and extracting narratives and relationships among characters [63,125]. Several taxonomies and conceptual frameworks have been proposed to map and synthesize prior work on user-generated book reviews [89,118].

Despite the differences between previously used datasets in terms of language, source, scale, and research topic, most datasets are collected via web scraping [26,75,150,152], using application programming interfaces (APIs) provided by the hosting platforms (for example, Goodreads used to provide an API, and

[8] Amazon (Amazon.com: Books) is currently the largest online bookseller worldwide. Goodreads is one of the dominant social reading and book review platforms based in the United States, with 90 million registered members as of 2019. LibraryThing is one of the most impactful social cataloging platforms based in the United States, with 2.6 million users as of 2021 [54,73,95,156–159].

Amazon web services (AWS) provides an API for individuals) [69,71,122,153,169], or a combination of the two [36,99].[9] Robots.txt files are a server-side solution for determining what data can be accessed and how, and can inform web scraping efforts. APIs implement the rules for data collection that providers define for their service, and are therefore a recommendable solution for data gathering. Not all platforms provide APIs, however, because enabling research may not be part of a provider's business model or might conflict with their user agreements. For instance, Goodreads shut down its API for accessing book review data in 2020 and made large-scale data scraping difficult by restricting its webpage content (e.g., sorting reviews with its proprietary algorithms) [36,150]. Given such implementations, data scraping is broadly adopted for data collection, although it might violate copyright and the EULA/TOS of a platform.

Legal risks and ethical concerns associated with book review scraping and related downstream tasks have been discussed before, but only in small numbers. One of the articles we reviewed mentioned copyright exemptions for research [114]. A few articles have discussed the acquisition of permissions for data collection [162,163] and attempts to request permissions [114] from the provider platforms. Considerations of human subjects research and institutional ethics review are also often absent.[10] Within publications of U.S.-based scholars, we only found two articles where consideration of and exemption from Institutional Review Board (IRB)[11] oversight was explicitly mentioned [12,102]. Relatedly, only a small number of articles explicitly discussed actions taken to protect the identities of the book reviewers, such as (1) removing user names and other user profile information that might reveal a reviewer's real-world identity (e.g., self-reported non-binary gender identities) [4,32,38,88,102,114,120,122,130], paraphrasing quoted reviews [20], and/or (2) not publishing the original data scraped, which might also violate copyright and EULAs [12,131,150]. In contrast, most research did not describe how researchers pre-processed potential personally identifiable information; such information might remain accessible in existing book review datasets [62,103].[12].

In conclusion, our literature review indicates a general absence of (1) informed consent from authors of book reviews; (2) permissions obtained from data sources; or (3) institutional ethics review in existing computational studies of

[9] In some publications, data collection methods are not explicitly specified, and general terms like "got", "collected", "downloaded" and "extracted" are used in lieu of providing more detailed collection method descriptions. [4,27,36,64,139].

[10] Such considerations might not apply to studies on user-generated data. We elaborate on this issue in Sect. 3.4.

[11] In the United States, an Institutional Review Board (IRB) is an administrative unit formally designated to review and monitor research activities using human research subjects. IRBs approve or disapprove research proposals prior to their initiation to ensure the rights and welfare of human research subjects [144].

[12] Due to copyright and perform restrictions, it is recommendable to share only unique key identifiers for collected data items instead of actual datasets such that other researchers can rehydrate the data, which bears the risk of collecting incomplete datasets [32,109].

user-generated book reviews. Discussions of legal and ethical risks associated with such practices were also largely absent. As discussed in the introduction, failure to consider these issues could pose risks to online users/readers, researchers, and academia alike. Therefore, we survey a broader range of literature and guidelines to fill this gap in legal and ethical considerations of the scholarly usage of user-generated book reviews.

3 Analysis and Findings

We analyze (1) relevant laws; (2) platform policies; (3) user rights and expectations; and, (4) researchers' discussion of ethical issues in user-generated data research. We combine our analysis with real-world and research cases, particularly studies on book reviews. Our findings are presented in the following four subsections. The four aspects we consider are not isolated; in practice, they intertwine with each other in complementary or sometimes conflicting ways (as exemplified in the following discussions). For example, some research aspects might be ethical but not legal, e.g., violating TOS to scrape publicly available book information, or legal but not ethical, for example, quoting snippets from identifying public information of vulnerable communities.

3.1 Legal Permissions and Risks

One primary legal risk associated with research based on user-generated book reviews comes from data scraping. Various data-scraping lawsuits have been initiated, claiming violations of TOS, copyright infringements, or unfair competition [15,57]. In this subsection, we consider cases in the U.S. as an example. Researchers from other jurisdictions should refer to the corresponding regulations that apply to their research scenarios. For U.S.-based studies, researchers should first consult the Copyright Law of the United States [143] and the fair use doctrine for risks associated with copyright infringement, and the Computer Fraud and Abuse Act (CFAA) [30] to minimize the risks of being sued. Fair use only conditionally permits unlicensed use of copyright-protected work under certain circumstances[13]. Scholarship and research activities are typically activities protected by the fair use doctrine [143], but a self-assessment of each use case and/or consultation with a copyright specialist can help to make responsible decisions.

For research based on large-scale scraped data [122,152], to reduce legal risks associated with copyrighted content, researchers may consider making

[13] The US copyright law demands consideration of four factors for determining whether fair use is applicable: purpose and character of the use; nature of the copyrighted work; amount and substantiality of the portion used; and the effect of the use upon the potential market for the copyrighted work. For research based on user-generated book reviews, the first two conditions of fair use may be less of a concern, but researchers should pay more attention to the third and fourth conditions.

transformative and non-consumptive use of the data[14], which has been increasingly adopted in computational studies of massive cultural data [79,113,123]. Furthermore, scholarly use of book review data might not fall under the concerns of the CFAA as the usage is non-commercial and for educational/research only [5]. However, it is essential for researchers to understand the CFAA and address other potential conflicts between their intended use of data and the provider platforms' policies (e.g., TOS/EULAs), which are discussed in the following subsections.

Second, researchers need to comply with laws that govern the use of personal data and privacy. In the U.S., applicable laws include privacy laws [3,82], state laws like the California Consumer Privacy Act (CCPA) [23], and state laws protecting the privacy of library records. Library records typically include online search records, circulation records, interlibrary loan records, personally identifiable uses of library materials and services, etc. Although no federal legislation or case law has been established to protect the privacy of library records, forty-eight states and the District of Columbia have established laws regarding the confidentiality of library records [7,90].[15] While accessing and presenting publicly accessible user-generated book reviews obtained from commercial websites is different from disclosing confidential user records held by libraries, both actions might expose individual reviewers' personal data to a third party or the public. Therefore, we advise researchers to check relevant laws on library records to understand legal requirements associated with library patron records and data alike.

Last but not least, researchers should note that user-generated content is often contributed by users from around the globe, regardless of where the platforms are based. For instance, while Goodreads is based in the U.S., its user base is global [122,140]. Therefore, researchers working on data collection from U.S.-based providers should examine international and regional regulations as well, such as The World Intellectual Property Organization (WIPO) Copyright Treaty [161], and Europe's Directive on Copyright in the Digital Single Market [135] and the General Data Protection Regulation (GDPR) [44], and China's Personal Information Protection Law (PIPL) [155]. This recommendation applies to research based in other areas of the world, too.

[14] "Transformative use" of the data alters original content to give it "new expression, meaning or message" [133]. "Non-consumptive use" refers to computer-assisted research, which has been found not to conflict with copyright holders' interests. For instance, in transformative and non-consumptive research, digital humanities scholars can conduct computational text analysis of millions of books (copyrighted books included) without actually reading or re-disseminating (i.e., without human "consumption" of) any expressive content of those books [113].

[15] It should be noted that "these state laws, however, are overridden or trumped by federal laws that allow federal agencies to seek library records" [21,90]. They vary by state, however, they reflect a consensus that library users' data are confidential and should only be disclosed under certain circumstances (e.g., with the user's informed consent, under a court order, etc.).

3.2 Policies and Guidelines Issued by Platforms Provider

Three types of documents from platform providers are most relevant for understanding the permitted use of book review data (any of them, or none, may be available): data access solutions provided by the platform (e.g., APIs, AWS), TOS/EULAs, and "robots.txt" files[16]. These files specify what and how data from these services can or cannot be used, among other things. For instance, the TOS of Goodreads [56][17] severely limit use of data to prevent inappropriate commercial competition, copyright infringement, and violations of user privacy rights. It states that the allowable use of Goodreads data does not include "*any use of data mining, robots, or similar data gathering and extraction tools*" [56] and restrict the data that people can access from their front page via review sorting algorithms and user-interface design [150]. In addition, Goodreads' robots.txt excludes a list of sitemaps and webpages from web scraping even though they are publicly accessible [55], and the site retired their API in 2020 [122]. Given these limitations, the next question for researchers might be: what are the consequences of scraping data from platforms that explicitly or implicitly prohibit scraping?

On one hand, researchers might argue for their use of data scraping or scraped data against the platforms' policies under certain conditions, e.g., when the research's "*benefits to society outweigh the harm of violating terms of service*" [145]. One important aspect in advocating for this position is to consider how "public" the scraped data are: while dominant social media platforms are likely to "*continue to push the boundaries on allowable methods to limit data scraping*", the Supreme Court's decisions on the case of hiQ Labs vs. LinkedIn signaled "*a shift in the way courts may be viewing attempts to restrict data scraping*" [53] in the U.S.[18] While heated debates on the implications of this verdict continue, a widely recognized takeaway for researchers is that scraping data that is publicly accessible without access control, such as passwords, paywalls, physical or technical barriers (e.g., software verification), is not necessarily unlawful, even if such scraping is prohibited by the platform's TOS/EULAs [14,49]. This verdict, to some extent, suggests that researchers are not doomed to be criminalized for scraping publicly accessible data without a platform's permission or against its policies. On the other hand, the legal and ethical consequences of violating TOS/EULAs in data collection for research purposes remain an open

[16] Robots.txt files are developed and used primarily to inform search engines and web scrapers whether data on a webpage is prohibited or permitted for harvesting. They are widely adopted by the websites to regulate scraping, although their prohibitions "*fall into a legal grey area*" [123].

[17] Accessed in August 2022.

[18] In this case, hiQ scraped publicly available user data from LinkedIn's website to supply its own business, in spite of LinkedIn's no-data-scraping policies, letters specifically addressed to hiQ, and technical measures enacted against hiQ. LinkedIn claimed that hiQ's scraping violated the CFAA, the Digital Millennium Copyright Act, and state trespass law, while hiQ denied these claims and asserted its right to scrape publicly accessible data [53].

question [46,145,148], as the feasibility and enforceability of platforms' TOS, particularly their prohibitions, are subject to further examination [5,28]. Existing research on the TOS of over a hundred global social media platforms found that "*though these provisions are very common, they are also ambiguous, inconsistent, and lack context*" and "*may reflect possibly conflicting values*" [3,46]. It is also important to note that platform policies might not align with the best interests of their users or researchers' ethical considerations [3,46].

In short, although there is no clear answer to "*whether researchers should be permitted to violate TOS when collecting data*" [46], a violation of TOS alone does not necessarily criminalize researchers' data scraping. In the U.S., current federal regulation does not enforce researchers to follow EULAs and does not criminalize scraping as a violation of the CFAA (although scraping might still violate copyright and privacy laws and regulations). Researchers whose plans for data scraping do not align with the platform policies are recommended to conduct a careful assessment of their use case. For instance, they should consider if the data to be scraped are publicly accessible, and they should avoid scraping from disallowed webpages/websites that are specified in robots.txt files/EULA/TOS. Finally, even if data collection procedures follow the requirements and guidelines of a platform, researchers also need to consider how to protect users, as EULAs/TOS do not necessarily align with the best interest of users [3,47].

3.3 User Rights and Expectations

Relevant laws and platform policies may fail to protect user rights or meet their privacy expectations: "*Users care about how their content can be used yet lack critical information*" [47]. Therefore researchers should assess how their planned work might conflict with the interests of users. To help with that, based on our literature review, we identified four potential pitfalls and approaches for avoiding them. First, a user's acceptance of TOS is not the same as their "informed consent" to any third-party use of their data. Prior surveys have shown that most users do not read the TOS they accept or consent to due to "*lack of choice, inaptitude, or habituation*" [18,105]. Meanwhile, without prior knowledge or additional information, it is beyond any individual user's capability to predict the third-party use of their data and potential hazards of that use. Therefore, responsible researchers should not assume that their use of user-generated data is within the expectations of the data creators simply due to their acceptance of a platform's TOS.

Second, researchers should not necessarily take publicly accessible data as "data open for use". This false assumption has led to various problems, such as re-identification of users in data shares and violations of user privacy [35,167,168]. There is a fundamental difference between (1) the data is public; and (2) the data has been consciously made public by users. The degree to which user (generated) data is public varies: some data are actively created and shared by users (e.g., book reviews that are set to be visible to all), while other data are passive traces automatically generated by algorithms based on user activities (e.g., location information based on IP addresses, time stamps associated with user

activities, etc.) [65]. For the first case, some platforms, such as LibraryThing, allow users to set and alter the level of visibility of their contributed content (e.g., write a review that is public to all or kept to oneself) [94]. If reviewers explicitly choose to make their data public, researchers can assume that users are aware of their choice, even though they might not anticipate use cases beyond the visibility of the given site.[19] Even in this case and moreover in general, users might not be aware that their data is part of passive digital traces or is available for third-party use.

Third, using public user data does not free the researchers from responsibility to avoid accidental or inappropriate use of private information, even though it might have been the users who disclosed their private information in the first place. As mentioned, user-generated book reviews may disclose personally identifiable and other personal information [62,103,122]. Additionally, online book reviews may disclose the identities of people other than the reviewers themselves [94], including vulnerable groups of people who have no knowledge of or control over the existence of a review. For instance, in online book reviews of children's books, ages, gender identities, grades, and first names of children are frequently shared by adult reviewers [106]. Such information, when cross-referenced with reviewer profiles, can put a child's real-world identity at risk. Responsible researchers are advised to remove any personally identifying information from their datasets.

Fourth, ethical research should respect and protect the book reviewers' intellectual freedom and freedom of speech, both of which are particularly pertinent to the missions and values of LIS [112]. Book reviews may contain controversial opinions that may not only frustrate or irritate other readers but also unsettle the public at large [104]. Taking library practices in the U.S. as an example, as long as a review does not break any laws or TOS, a reviewer is entitled to "*write what they think*" and "*dispute ideas and words without limitation*" [94], even though others may oppose them. Such principles are debated among online book reviewers. For example, a group of book reviewers on Goodreads repeatedly gave one-star reviews to LGBTQIA+ books, sometimes even before the release of advanced copies or as part of book campaigns [126]. Many users consider such behavior to be trans- and homophobic actions targeting LGBTQIA+ groups and marginalized authors, and demanded moderation from Goodreads to remove these reviews [126]. However, Goodreads did not remove the ratings as requested because one-star ratings themselves did not directly violate any platform regulations (while personal attacks and hate speech, for example, would violate their guidelines) [41]. In controversial cases, researchers from different disciplines and cultural backgrounds could potentially approach the data in different ways, which may or may not align with the interests and expectations of either the users or the platforms involved. We are not in a position to question anyone's research priorities or personal stances; we simply remind researchers that every reader is entitled to their intellectual freedom

[19] However, in practice, it is difficult for researchers to verify whether the reviewers are indeed aware of the public accessibility of their data. Researchers should not make assumptions about users' awareness.

and freedom of speech, and that library professionals adhere to these principles [84,91]. Responsible researchers should stay alert to any personal biases and feelings toward different groups of reviewers. All users/readers should be equally protected from unexpected and unwanted surveillance, tracking, blaming, and attacks in scholarly research.

3.4 Discussions and Concerns from the Research Community

There have been various case studies, guidelines, and statements for how to conduct compliant, responsible, and ethical research on user-generated data in general and for specific genres [1,2,11,46,52,58,81,87,121,148], as well as more specialized discussions on this topic from LIS perspectives [13,100,101]. Here we zoom in on three topics that have been heatedly discussed: (1) explicit informed consent from human research subjects; (2) institutional/administrative review and approval; and (3) platform restrictions.

As for informed consent and institutional/administrative review, while some researchers argue that such conventional research practices should be applied to research on user-generated data from online sources [46,51,147], others disagree [74,87,147]. The latter group argues that scholarly research of such data may be exempt from informed consent under certain conditions, e.g., when it is almost impossible to obtain "retrospective" informed consent for archival research [87][20]; and when research projects involve "*no more than minimal risk to the subjects*" and "*could not practicably be carried out without the waiver or alteration*" [74]. Other researchers claim that institutional/administrative review and approval, such as IRBs in the U.S., tend to apply "*overly restrictive guidelines developed for biomedical research to lower risk studies*", and sometimes lack "*the expertise to effectively evaluate technical proposals*" [147]. They also argue that tensions between conventional requirements (such as IRBs) and social computing research could actually "*increase risks to participants, delay data collection, or substantively change a research project*" [147]. Furthermore, researchers' attitudes toward platform restrictions also diverge. For example, some researchers insist that the legitimacy and enforceability of TOS are questionable [46,148], which raises concerns about the legal consequences and ethics of either following or violating the TOS. So far, no consensus has been reached on these three topics with regard to the unobtrusive analysis of user-generated content [147], although opinions are converging on other aspects of ethical social computing, such as ensuring participants' access to the research outcomes [148].

Nevertheless, there exists consensus on the holism, contextuality, and complexity of the ethical conduct of research [45,167]. It has been broadly acknowledged that weighing potential harms and intended benefits for all stakeholders (e.g., users, platforms, and society at large) and mitigating different considerations are

[20] Kosinski and colleagues argue that no consent is needed and user-generated online data can be conceptualized as archival data if (1) users consciously made their data public; (2) data collected is anonymized; (3) researchers do not interact with participants; and, (4) no identifiable user information is published. [87].

hard [46,147]. We have consistently found such dilemmas and trade-offs in existing book review studies. For example, some studies de-identified reviewers by removing their original usernames and partial user profiles (e.g., location, gender identities) [4,122]. This makes reviewers less likely to be tracked down, although risks of re-identification remain [122,168]. However, such de-identification deprives the book reviewers of credit for their intellectual contributions and copyrighted work, to which they are entitled as content creators [22]. To overcome this limitation, some researchers choose to seek informed consent from book reviewers they intend to quote in their research publications, particularly as to whether the reviewers want to be quoted verbatim under their scraped usernames [12,150]. However, getting permission from individual reviewers requires personal contact with human research subjects, which means their data collection is no longer unobtrusive. For U.S.-based studies, unless an IRB review is conducted, this strategy would be considered risky and inappropriate[21]. Similar trade-offs have emerged from data publication as well. Some researchers chose to selectively publish their scraped data, or not to publish any of their scraped data at all, in order to protect reviewers' data from inappropriate use [122,150]. However, this raises questions about research reproducibility and transparency [76,132].

4 Discussion and Limitations

When planning responsible research projects, different factors and considerations might not align or conflict with each other in actual practice, leaving researchers with a number of dilemmas to solve and difficult decisions to make. For instance, as book review platforms often neither provide APIs nor permit scraping, researchers need to evaluate the risks associated with violating platform policies or even laws. Researchers are furthermore expected to honor readers' rights and expectations, which are crucial concepts that are not always prioritized by platforms' policies. There are trade-offs and risks associated with many decisions that have to be made by researchers. While researchers might not always be able to resolve them, they should minimize potential harm and make situation-specific decisions to guarantee that the benefits of their research to society outweigh the risks of potential problems. Institutional review and oversight, such as IRBs, share this goal, but they might not apply to working with archival and/or online data, such that researchers need not only to understand these risks, but also have the knowledge and skills to mitigate them. Although our research emphasizes legal risks and ethical problems

[21] Different IRBs might make different decisions on requests for exemption based on specific research proposals. For instance, we learned from our own research experience that analysis of publicly available and de-identified book review data without any interaction with the reviewers is mostly likely to be considered "Not Human Subjects Research" (NHSR) by the IRB at our home institution [142]. In this case, researchers who believe their work does not require IRB review or oversight should submit a request to their institution's IRB for a designation as Not Human Subjects Research. They might also consider asking for an Exempt Status determination, in which case they are performing Human Subjects research but are exempt from regular oversight.

associated with research on user-generated book reviews, by no means do we intend to discourage research with this genre or type of data. We rather hope to critically engage with this research area by contributing LIS perspectives and facilitating future research by flagging potential pitfalls and suggesting potential solutions.

Our investigation is limited in several ways. As we are neither law practitioners nor policymakers, we are not in a position to give legal advice. Besides, given the broad multidisciplinary reach of user-generated data research, discussions about our research questions remain controversial, without a clear consensus or cross-disciplinary norms. Most importantly, scientific research often comes with risks and uncertainties, and decisions should be made based on the specific context of a research problem. As there is no panacea for minimizing research risks or guaranteeing ethical practice, instead of crafting "guidelines for everyone", we synthesized prior relevant literature, case studies, and library practices to understand (1) what researchers should look out for; and (2) what they should leverage to guide and assess their scholarly usage of user-generated book review data. Second, given the breadth and multidisciplinarity of book review research, our scope of analysis was unavoidably yet necessarily narrowed down. For instance, we took a U.S.-centric perspective, and some of that might not apply to other regions of the world. Nevertheless, the U.S. context serves to contextualize and exemplify the complexities of the legal and ethical issues in book review studies, and provides a regional research case. As an overview, our research outlines the primary legal and ethical concerns about scholarly usage of user-generated book reviews, which are not limited to research based in the U.S.

Finally, while we put legal and ethical considerations forward as an insufficiently discussed problem in research practice of user-generated book reviews, these considerations are by no means overlooked in research at large. Instead, as our discussion shows, there exist plenty of generally applicable and insightful papers and guidelines to refer to. Thus, this paper calls for more attention to both (1) the paucity of scholarly discussions about legal and ethical concerns in book review research; and, (2) how researchers can leverage existing resources to address this particular problem.

5 Conclusions and Future Work

This paper presents an overview of legal risks and ethical concerns associated with scholarly usage of user-generated book reviews. Our review was primarily motivated by (1) the lack of attention to this problem in prior computational and empirical studies of user-generated book reviews; and, (2) the dual role of the users and readers who are subject to potential harm caused by scholarly use of their data. We reviewed relevant laws, platform policies, user expectations, and prior research to inform future researchers of potential legal and ethical pitfalls, and offer some suggestions for how to avoid them through practical solutions. We also drew on library practices and guidelines to better understand why and how

researchers should protect data generated by users/readers. The pitfalls identified and discussed include copyright infringement, violations of TOS/EULAs, conflicts with user rights and expectations, and the role of informed consent and institutional reviews.

The intended contributions of this paper are threefold. First, given the dual role of online book reviewers as (1)content consumers and producers; and, (2) readers, we emphasized the significance of evaluating and reducing risks associated with scholarly usage of user-generated book reviews. Second, we analyzed legal and ethical concerns that have been under-investigated in the context of user-generated book reviews. We hope these insights help to inform future studies on how to reduce potential risks and better protect the users/readers. Third, under the overarching umbrella of responsible data-driven research, we demonstrated how to assess legal and ethical issues associated with the characteristics, stakeholders, and research contexts of book reviews.

For future work, there are more questions to scrutinize. First, there is a variety of data analyses on user-generated book reviews: some studies annotate individual book reviews word by word while others only map high-level patterns in corpora (e.g., average book ratings). Should different ethical expectations be applied to different use cases depending on the research scale, granularity, and "distance from the readers"? For instance, can researchers consider informed consent inapplicable for de-identified and paraphrased quotations or non-consumptive text mining of book reviews? To answer these questions, we need to examine more prior research to understand the needs and costs (e.g., time and administrative procedures) of different actions taken. There are also open questions from the perspective of libraries, such as the argument that libraries are losing competency as a result of their "hands off user data" practice, which sometimes limits their ability to serve their patrons [43,90]. Are user-generated book review datasets filling the gaps or taking advantage of libraries' "moral absence", and if so, where do researchers stand on this question? To explore this question, qualitative studies, such as interviews with researchers working with user-generated book reviews and/or questionnaires among online book reviewers, might be effective methods for gaining a nuanced understanding of different stakeholders' needs, expectations, and concerns. We also encourage collaborations among researchers from diverse communities and different cultures or regions to cross-examine and broaden our knowledge of this issue.

References

1. ACM Code 2018 Task Force: ACM code of ethics and professional conduct (2018). https://www.acm.org/code-of-ethics
2. ACM Technology Policy Council, ACM Europe Technology Policy Committee and ACM US Technology Policy Council: Statement on principles for responsible algorithmic systems (2022). https://www.acm.org/binaries/content/assets/public-policy/final-joint-ai-statement-update.pdf
3. Acquisti, A., Brandimarte, L., Loewenstein, G.: Privacy and human behavior in the age of information. Science **347**(6221), 509–514 (2015)

4. Albrechtslund, A.M.B.: Negotiating ownership and agency in social media: community reactions to amazon's acquisition of Goodreads. First Monday (2017)
5. American Civil Liberties Union: Federal court rules 'big data' discrimination studies do not violate federal anti-hacking law (2020). https://www.aclu.org/press-releases/federal-court-rules-big-data-discrimination-studies-do-not-violate-federal-anti
6. American Library Association: The USA patriot act (2009). https://www.ala.org/ala/washoff/WOissues/civilliberties/theusapatriotact/usapatriotact.htm
7. American Library Association: Intellectual freedom: issues and resources (2017). https://www.ala.org/advocacy/intfreedom
8. American Library Association: Ala statement on book censorship (2021). https://www.ala.org/advocacy/statement-regarding-censorship
9. American Library Association: State privacy laws regarding library records (2021). https://www.ala.org/advocacy/privacy/statelaws
10. American Library Association Council: Policy concerning confidentiality of personally identifiable information about library users (1991). https://www.ala.org/advocacy/intfreedom/statementspols/otherpolicies/policyconcerning
11. Annette Markham and Elizabeth Buchanan: Ethical decision-making and internet research: recommendations from the AoIR ethics working committee (version 2.0) (2012). https://aoir.org/reports/ethics2.pdf
12. Antoniak, M., Walsh, M., Mimno, D.: Tags, borders, and catalogs: social re-working of genre on librarything. Proc. ACM Hum.-Comput. Interact. **5**(CSCW1), 1–29 (2021)
13. Asher, A., et al.: Ethics in research use of library patron data: glossary and explainer (2018). https://doi.org/10.17605/OSF.IO/XFKZ6
14. Association for Computing Machinery: Scraping by: reconsidering law & technology for online data collection - 19 May 2022 (2022). https://www.acm.org/public-policy/ustpc/hottopics/online-data-collection
15. Band, J.: LCA comments on authors guild v. hathitrust decision (2012). https://www.arl.org/news/lca-comments-on-authors-guild-v-hathitrust-decision/
16. Bartley, P.: Book tagging on LibraryThing: how, why, and what are in the tags? Proc. Am. Soc. Inf. Sci. Technol. **46**(1), 1–22 (2009)
17. BBC News: Author Richard Brittain attacked reviewer with bottle (2015). https://www.bbc.com/news/uk-scotland-edinburgh-east-fife-34775814
18. Böhme, R., Köpsell, S.: Trained to accept? A field experiment on consent dialogs. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 2403–2406 (2010)
19. Boot, P., Koolen, M.: Captivating, splendid or instructive?: assessing the impact of reading in online book reviews. Sci. Study Lit. **10**(1), 35–63 (2020)
20. Bourrier, K., Thelwall, M.: The social lives of books: reading Victorian literature on goodreads. J. Cult. Anal. **1**(1), 12049 (2020)
21. Bowers, S.L.: Privacy and library records. J. Acad. Librariansh. **32**(4), 377–383 (2006)
22. Bruckman, A.: Studying the amateur artist: a perspective on disguising data collected in human subjects research on the internet. Ethics Inf. Technol. **4**(3), 217–231 (2002)
23. California Legislative Information: Title 1.81.5. California consumer privacy act of 2018 (2018). https://leginfo.legislature.ca.gov/faces/codes_displayText.xhtml?division=3.&part=4.&lawCode=CIV&title=1.81.5

24. Carman, N.: LibraryThing tags and Library of Congress Subject Headings: A comparison of science fiction and fantasy works. School of Information Management at Victoria University of Wellington (2009)
25. Chang, K., et al.: Book reviews and the consolidation of genre. In: DH2020 (ADHO) Proceedings (2020). http://dx.doi.org/10.17613/02q2-1v27
26. Chen, P.Y., Dhanasobhon, S., Smith, M.D.: All reviews are not created equal: the disaggregate impact of reviews and reviewers at amazon.com (2008)
27. Chevalier, J.A., Mayzlin, D.: The effect of word of mouth on sales: online book reviews. J. Mark. Res. **43**(3), 345–354 (2006)
28. Court of Appeal, Second District, Division 3, California.: Long v. Provide Commerce Inc (2016). https://caselaw.findlaw.com/ca-court-of-appeal/1729412.html
29. Crawford, K., Finn, M.: The limits of crisis data: analytical and ethical challenges of using social and mobile data to understand disasters. GeoJournal **80**(4), 491–502 (2015)
30. Computer Crime and Intellectual Property Section Criminal Division: Prosecuting computer crimes manual (2010). https://www.justice.gov/criminal/file/442156/download
31. Dai, L.: From the history of the book to the history of reading: theories and methods for historical studies of reading. Xinxing (2017)
32. De Greve, L., Martens, G.: # bookstagram and beyond: the presence and depiction of the Bachmann literary prize on social media (2007–2017). Digit. Humanit. Benelux J. **3**, 81–102 (2021)
33. Diesner, J., Chin, C.: Seeing the forest for the trees: considering applicable types of regulation for the responsible collection and analysis of human centered data. In: Human-Centered Data Science (HCDS) Workshop at 19th ACM Conference on Computer-Supported Cooperative Work and Social Computing (2016)
34. Diesner, J., Chin, C.L.: Usable ethics: practical considerations for responsibly conducting research with social trace data. In: Proceedings of Beyond IRBs: Ethical Review Processes for Big Data Research (2015)
35. Diesner, J., Chin, C.L.: Gratis, libre, or something else? Regulations and misassumptions related to working with publicly available text data. In: Actes du Workshop on Ethics In Corpus Collection, Annotation & Application (ETHICA2), LREC, Portoroz, Slovénie (2016)
36. Dimitrov, S., Zamal, F., Piper, A., Ruths, D.: Goodreads versus amazon: the effect of decoupling book reviewing and book selling. In: Ninth International AAAI Conference on Web and Social Media (2015)
37. Drew, C.: Data science ethics in government. Philos. Trans. R. Soc. A Math. Phys. Eng. Sci. **374**(2083), 20160119 (2016)
38. Driscoll, B., Rehberg Sedo, D.: Faraway, so close: seeing the intimacy in goodreads reviews. Qual. Inq. **25**(3), 248–259 (2019)
39. Driscoll, B., Rehberg Sedo, D.: The transnational reception of bestselling books between Canada and Australia. Global Media Commun. **16**(2), 243–258 (2020)
40. Ehrmann, T., Schmale, H.: The hitchhiker's guide to the long tail: the influence of online-reviews and product recommendations on book sales-evidence from German online retailing. In: ICIS 2008 Proceedings, p. 157 (2008)
41. Ellis, D.: What charles and anti-charles reveal about goodreads homophobia (2020). https://bookriot.com/goodreads-homophobia/
42. English, J., Ungar, L., Dhakecha, R.H., Scott, E.: Mining goodreads (literary reception studies at scale) (2018). https://pricelab.sas.upenn.edu/projects/goodreads-project

43. Estabrook, L.S.: Sacred trust or competitive opportunity: using patron records. Libr. J. **121**(2), 48–49 (1996)
44. European Union (EU): Complete guide to GDPR (general data protection regulation) compliance (2016). https://gdpr.eu/
45. Fiesler, C.: Ethical considerations for research involving (speculative) public data. Proc. ACM Hum.-Comput. Interact. **3**(GROUP), 1–13 (2019)
46. Fiesler, C., Beard, N., Keegan, B.C.: No robots, spiders, or scrapers: legal and ethical regulation of data collection methods in social media terms of service. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 14, pp. 187–196 (2020)
47. Fiesler, C., Lampe, C., Bruckman, A.S.: Reality and perception of copyright terms of service for online content creation. In: Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing, pp. 1450–1461 (2016)
48. Fiesler, C., Proferes, N.: "Participant" perceptions of twitter research ethics. Soc. Media+ Soc. **4**(1), 2056305118763366 (2018)
49. Fiesler, C.: Law & ethics of scraping: what HiQ v Linkedin could mean for researchers violating TOS (2017). https://cfiesler.medium.com/law-ethics-of-scraping-what-hiq-v-linkedin-could-mean-for-researchers-violating-tos-787bd3322540
50. Fornaciari, T., Poesio, M.: Identifying fake amazon reviews as learning from crowds. In: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pp. 279–287. Association for Computational Linguistics (2014)
51. Franzke, A.S., Bechmann, A., Zimmer, M., Ess, C.: Internet research ethics guidelines (IRE 3.0 6.1) (2019). https://aoir.org/reports/ethics3.pdf
52. Gilbert, E., Karahalios, K.: Understanding deja reviewers. In: Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work, pp. 225–228 (2010)
53. Goldfein, S., Keyte, J.: Big data, web 'scraping' and competition law: the debate continues. New York Law J. **258**(49), 1–3 (2017)
54. Goodreads: About goodreads (2022). https://www.goodreads.com/about/us
55. Goodreads: Goodreads robots.txt file (2022). https://www.goodreads.com/robots.txt
56. Goodreads: Terms of use (2022). https://www.goodreads.com/about/terms
57. Gray, J., Foong, C.: Publishers vs the internet archive: why the world's biggest online library is in court over digital book lending (2022). https://theconversation.com/publishers-vs-the-internet-archive-why-the-worlds-biggest-online-library-is-in-court-over-digital-book-lending-187166
58. Greene, D., Hoffmann, A.L., Stark, L.: Better, nicer, clearer, fairer: a critical assessment of the movement for ethical artificial intelligence and machine learning. In: Proceedings of the Annual Hawaii International Conference on System Sciences, pp. 2122–2131 (2019)
59. Guan, X., Li, Y., Gong, H., Sun, H., Zhou, C.: An improved SVM for book review sentiment polarity analysis. In: 2018 International Conference on Transportation Logistics, Information Communication, Smart City (TLICSC 2018). Atlantis Press (2018)
60. Hajibayova, L.: Investigation of goodreads' reviews: kakutanied, deceived or simply honest? J. Doc. **75**(3), 612–626 (2019)
61. HathiTrust Digital Library: Our digital library (2022). https://www.hathitrust.org/digital_library

62. He, R., McAuley, J.: Ups and downs: modeling the visual evolution of fashion trends with one-class collaborative filtering. In: Proceedings of the 25th International Conference on World Wide Web, pp. 507–517 (2016)
63. Holur, P., Shahsavari, S., Ebrahimzadeh, E., Tangherlini, T.R., Roychowdhury, V.: Modelling social readers: novel tools for addressing reception from online book reviews. Roy. Soc. Open Sci. **8**(12), 210797 (2021)
64. Hong, H., Xu, D., Xu, D., Wang, G.A., Fan, W.: An empirical study on the impact of online word-of-mouth sources on retail sales. Inf. Discov. Deliv. **45**(1), 30–35 (2017)
65. Howison, J., Wiggins, A., Crowston, K.: Validity issues in the use of social network analysis with digital trace data. J. Assoc. Inf. Syst. **12**(12), 2 (2011)
66. Howsam, L.: Old Books and New Histories: An Orientation to Studies in Book and Print Culture. University of Toronto Press, Toronto (2006)
67. Hu, N., Bose, I., Gao, Y., Liu, L.: Manipulation in digital word-of-mouth: a reality check for book reviews. Decis. Support Syst. **50**(3), 627–635 (2011)
68. Hu, N., Bose, I., Koh, N.S., Liu, L.: Manipulation of online reviews: an analysis of ratings, readability, and sentiments. Decis. Support Syst. **52**(3), 674–684 (2012)
69. Hu, N., Koh, N.S., Reddy, S.K.: Ratings lead you to the product, reviews help you clinch it? The mediating role of online review sentiments on product sales. Decis. Support Syst. **57**, 42–53 (2014)
70. Hu, N., Liu, L., Sambamurthy, V.: Fraud detection in online consumer reviews. Decis. Support Syst. **50**(3), 614–626 (2011)
71. Hu, N., Liu, L., Zhang, J.J.: Do online reviews affect product sales? The role of reviewer characteristics and temporal effects. Inf. Technol. Manag. **9**(3), 201–214 (2008)
72. Hu, Y.: Synthesizing digital libraries and digital humanities perspectives for illuminating under-investigated complexities associated with user-generated book reviews. In: Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries, pp. 1–2 (2022)
73. Hu, Y., LeBlanc, Z., Diesner, J., Underwood, T., Layne-Worthey, G., Downie, J.S.: Complexities associated with user-generated book reviews in digital libraries: temporal, cultural, and political case studies. In: Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries, pp. 1–12 (2022)
74. Hudson, J.M., Bruckman, A.: "Go away": participant objections to being studied and the ethics of chatroom research. Inf. Soc. **20**(2), 127–139 (2004)
75. Hui, N.: Content-specific ranking prediction for online reviews-case of douban book reviews. Manag. Rev. **33**(2), 176 (2021)
76. Hutton, L., Henderson, T.: Making social media research reproducible. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 9, pp. 2–7 (2015)
77. International Federation of Library Associations and Institutions: IFLA code of ethics for librarians and other information workers (full version) (2012). https://www.ifla.org/publications/ifla-code-of-ethics-for-librarians-and-other-information-workers-full-version/
78. International Federation of Library Associations and Institutions: IFLA statement on privacy in the library environment (2015). https://www.ifla.org/publications/ifla-statement-on-privacy-in-the-library-environment/
79. Jett, J., Cole, T., Maden, C., Downie, J.: The hathitrust research center workset ontology: a descriptive framework for non-consumptive research collections. J. Open Humanit. Data **2** (2016)

80. Jiang, M., Diesner, J.: Issue-focused documentaries versus other films: rating and type prediction based on user-authored reviews. In: Proceedings of the 27th ACM Conference on Hypertext and Social Media, pp. 225–230 (2016)
81. Jiang, M., Diesner, J.: Says who. . .? Identification of expert versus layman critics' reviews of documentary films. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pp. 2122–2132 (2016)
82. Kaminski, M.: A recent renaissance in privacy law. Commun. ACM **63**(9), 24–27 (2020)
83. Kayla: Book chat: Authors being negative towards reviewers (2017). https://gracelingaccountantblog.wordpress.com/2017/12/06/book-chat-authors-being-negative-towards-reviewers/
84. Klinefelter, A.: Reader privacy in digital library collaborations: signs of commitment, opportunities for improvement. ISJLP **13**, 199 (2016)
85. Koolen, M., Neugarten, J., Boot, P.: 'This book makes me happy and sad and i love it'. a rule-based model for extracting reading impact from English book reviews. J. Comput. Literary Stud. **1**(1) (2022)
86. Koolena, M., Bootb, P., van Zundertb, J.J.: Online book reviews and the computational modelling of reading impact. In: Proceedings of Workshop on Computational Humanities Research (CHR), vol. 1613, p. 0073 (2020)
87. Kosinski, M., Matz, S.C., Gosling, S.D., Popov, V., Stillwell, D.: Facebook as a research tool for the social sciences: opportunities, challenges, ethical considerations, and practical guidelines. Am. Psychol. **70**(6), 543 (2015)
88. Kuijpers, M.M.: Bodily involvement in readers' online book reviews: applying text world theory to examine absorption in unprompted reader response. J. Lit. Semant. **51**(2), 111–129 (2022)
89. Kutzner, K., Petzold, K., Knackstedt, R.: Characterising social reading platforms-a taxonomy-based approach to structure the field. In: Proceedings of the 14th International Conference on Wirtschaftsinformatik (2019)
90. Lambert, A.D., Parker, M., Bashir, M.: Library patron privacy in jeopardy an analysis of the privacy policies of digital content vendors. Proc. Assoc. Inf. Sci. Technol. **52**(1), 1–9 (2015)
91. Lamdan, S.S.: Why library cards offer more privacy rights than proof of citizenship: librarian ethics and freedom of information act requestor policies. Gov. Inf. Q. **30**(2), 131–140 (2013)
92. Lanjinger: One-star reviewing bombing started from the truce (the diary of martín santomé) (orginally in Chinese) (2021). https://k.sina.com.cn/article_5617041192_14ecd3f280200135ul.html
93. Lavin, M.J., et al.: Cultural analytics and the book review: models, methods, and corpora. In: DH2020(ADHO) Proceedings (2020). https://dh2020.adho.org/wp-content/uploads/2020/07/516_CulturalAnalyticsandtheBookReviewModelsMethodsandCorpora.html
94. LibraryThing: Privacy policy, community rules, and terms of service (2020). https://www.librarything.com/privacy
95. LibraryThing: About librarything (2022). https://www.librarything.com/about
96. Lin, E., Fang, S., Wang, J.: Mining online book reviews for sentimental clustering. In: 2013 27th International Conference on Advanced Information Networking and Applications Workshops, pp. 179–184. IEEE (2013)
97. Lu, C., Park, J.R., Hu, X.: User tags versus expert-assigned subject terms: a comparison of librarything tags and library of congress subject headings. J. Inf. Sci. **36**(6), 763–779 (2010)

98. Lunnay, B., Borlagdan, J., McNaughton, D., Ward, P.: Ethical use of social media to facilitate qualitative research. Qual. Health Res. **25**(1), 99–109 (2015)
99. Maity, S.K., Panigrahi, A., Mukherjee, A.: Book reading behavior on goodreads can predict the amazon best sellers. In: Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017, pp. 451–454 (2017)
100. Mannheimer, S., Pienta, A., Kirilova, D., Elman, C., Wutich, A.: Qualitative data sharing: data repositories and academic libraries as key partners in addressing challenges. Am. Behav. Sci. **63**(5), 643–664 (2019)
101. Mannheimer, S., Young, S.W., Rossmann, D.: On the ethics of social network research in libraries. J. Inf. Commun. Ethics Soc. (2016)
102. Martens, M., Balling, G., Higgason, K.A.: # booktokmademereadit: young adult reading communities across an international, sociotechnical landscape. Inf. Learn. Sci. (ahead-of-print) (2022)
103. McAuley, J., Targett, C., Shi, Q., Van Den Hengel, A.: Image-based recommendations on styles and substitutes. In: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 43–52 (2015)
104. McCluskey, M.: Goodreads' problem with extortion scams and review bombing (2021). https://time.com/6078993/goodreads-review-bombing/
105. McDonald, A.M., Cranor, L.F.: The cost of reading privacy policies. ISJLP **4**, 543 (2008)
106. Mengting, W.: UCSD book graph: Goodreads datasets (2019). https://sites.google.com/eng.ucsd.edu/ucsdbookgraph/home
107. Metcalf, J., Crawford, K.: Where are human subjects in big data research? The emerging ethics divide. Big Data Soc. **3**(1), 2053951716650211 (2016)
108. Milligan, I.: The problem of history in the age of abundance (2016). http://hdl.handle.net/10012/11817
109. Mishra, S., Saini, A., Makki, R., Mehta, S., Haghighi, A., Mollahosseini, A.: Tweetnerd-end to end entity linking benchmark for tweets. arXiv preprint arXiv:2210.08129 (2022)
110. Nakamura, L.: "Words with friends": socially networked reading on goodreads. PMLA/Publ. Mod. Lang. Assoc. Am. **128**(1), 238–243 (2013)
111. Nan, X., Li, M., Shi, J.: Using altmetrics for assessing impact of highly-cited books in Chinese book citation index. Scientometrics **122**(3), 1651–1669 (2020)
112. Oltmann, S.M.: Intellectual freedom and freedom of speech: three theoretical perspectives. Libr. Q. **86**(2), 153–171 (2016)
113. Organisciak, P., Downie, J.S.: Research access to in-copyright texts in the humanities. In: Information and Knowledge Organisation in Digital Humanities, pp. 157–177. Routledge (2021)
114. Pianzola, F., Rebora, S., Lauer, G.: Wattpad as a resource for literary studies. quantitative and qualitative examples of the importance of digital social reading and readers' comments in the margins. PLoS ONE **15**(1), e0226708 (2020)
115. Pianzola, F., et al.: Books' impact in digital social reading: towards a conceptual and methodological framework. In: Digital Humanities 2022 Conference Abstracts, pp. 94–98 (2022). https://dh2022.dhii.asia/dh2022bookofabsts.pdf
116. Pinch, T.: Book reviewing for amazon.com: how socio-technical systems struggle to make less from more. In: Managing Overflow in Affluent Societies, pp. 80–99. Routledge (2012)
117. Reads with Rachel: Author attacks book reviewer |Richard Brittain | authors behaving badly (2022). https://www.youtube.com/watch?v=4Z5iIP8c5qs

118. Rebora, S., et al.: Digital humanities and digital social reading. Digit. Scholarsh. Humanit. **36**(Supplement_2), ii230–ii250 (2021)
119. Rebora, S., Messerli, T., Herrmann, J.B.: Towards a computational study of German book reviews. A comparison between emotion dictionaries and transfer learning in sentiment analysis. 8. Jahrestagung «Digital Humanities im deutschsprachigen Raum»(DhD), Potsdam, D. (2022)
120. Rebora, S., Pianzola, F.: A new research programme for reading research: analysing comments in the margins on wattpad. DigitCult-Sci. J. Digit. Cult. **3**(2), 19–36 (2018)
121. Rezapour, R., Diesner, J.: Classification and detection of micro-level impact of issue-focused documentary films based on reviews. In: Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, pp. 1419–1431 (2017)
122. Sabri, N., Weber, I.: A global book reading dataset. Data **6**(8), 83 (2021)
123. Samberg, R.G., Hennesy, C.: Law and literacy in non-consumptive text mining: guiding researchers through the landscape of computational text analysis (2019)
124. Sen, S., Lerman, D.: Why are you telling me this? an examination into negative consumer reviews on the web. J. Interact. Mark. **21**(4), 76–94 (2007)
125. Shahsavari, S., et al.: An automated pipeline for character and relationship extraction from readers literary book reviews on goodreads.com. In: 12th ACM Conference on Web Science, pp. 277–286 (2020)
126. Sharma, R.: Black and LGBTQ+ authors say they're being harassed on goodreads and trolled with one-star book reviews (2021). https://inews.co.uk/culture/books/goodreadsbookreviewsblacklgbtq-authorsharrassedtrolled949179
127. Sharmaa, A., Hu, Y., Wu, P., Shang, W., Singhal, S., Underwood, T.: The rise and fall of genre differentiation in English-language fiction. In: DH2020 (ADHO) Proceedings, vol. 1613, p. 0073 (2020)
128. Sheila (Book Journey): When authors attack... (2011). https://bookjourney.net/2011/12/04/when-authors-attack/
129. Shen, X., Zhang, K.Z., Zhao, S.J.: Understanding information adoption in online review communities: the role of herd factors. In: 2014 47th Hawaii International Conference on System Sciences, pp. 604–613. IEEE (2014)
130. Shenglan, T., Haiqing, H., JIANG, L., Xu, Z., SELMAN, R.L.: Chinese and English reviews of a story about teenagers' struggles: a multi-method analysis of cultural differences in narrative interpretation. Beijing Int. Rev. Educ. **2**(3), 365–387 (2020)
131. Sourati Hassan Zadeh, Z., Sabri, N., Chamani, H., Bahrak, B.: Quantitative analysis of fanfictions' popularity. Soc. Netw. Anal. Mining **12**(1), 1–11 (2022)
132. Srivastava, A.K., Mishra, R.: Analyzing social media research: a data quality and research reproducibility perspective. IIM Kozhikode Soc. Manag. Rev. **12**(1), 39–49 (2021)
133. Supreme Court: Campbell v. acuff-rose music (92-1292), 510 U.S. 569 (1994). https://www.law.cornell.edu/supct/html/92-1292.ZS.html
134. Szkolar, D.: The USA patriot act: should your library have an official policy? (2013). https://ischool.syr.edu/the-usa-patriot-act-should-your-library-have-an-official-policy/
135. The European Parliament and the Council of the European Union: 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). https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32019L0790&from=EN

136. Thelwall, M.: Book genre and author gender: romance > paranormal-romance to autobiography > memoir. J. Assoc. Inf. Sci. Technol. **68**(5), 1212–1223 (2017)
137. Thelwall, M.: Reader and author gender and genre in goodreads. J. Librariansh. Inf. Sci. **51**(2), 403–430 (2019)
138. Thelwall, M., Kousha, K.: Goodreads: a social network site for book readers. J. Am. Soc. Inf. Sci. **68**(4), 972–983 (2017)
139. Thomas, M., Caudle, D.M., Schmitz, C.: Trashy tags: problematic tags in librarything. New Library World (2010)
140. Slee, T.J.: Who is the average goodreads user? You'll be surprised! (2017). https://www.goodreads.com/author_blog_posts/14538341-who-is-the-average-goodreads-user-you-ll-be-surprised
141. Tsur, O., Rappoport, A.: Revrank: a fully unsupervised algorithm for selecting the most helpful book reviews. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 3 (2009)
142. University of Illinois Office for the Protection of Research Subjects: Decision trees (2022). https://oprs.research.illinois.edu/review-processes-checklists/decision-trees
143. US Copyright Office: Copyright law of the united states (title 17) (2021). https://www.copyright.gov/title17/
144. U.S. Food and Drug Administration: Institutional review boards frequently asked questions (1998). https://www.fda.gov/regulatory-information/search-fda-guidance-documents/institutional-review-boards-frequently-asked-questions
145. Vaccaro, K., Karahalios, K., Sandvig, C., Hamilton, K., Langbort, C.: Agree or cancel? Research and terms of service compliance. In: ACM CSCW Ethics Workshop: Ethics for Studying Sociotechnical Systems in a Big Data World (2015)
146. Verma, P.: The fight between authors and librarians tearing book lovers apart (2022). https://www.washingtonpost.com/technology/2022/07/25/internet-archive-digital-lending-lawsuit/
147. Vitak, J., Proferes, N., Shilton, K., Ashktorab, Z.: Ethics regulation in social computing research: examining the role of institutional review boards. J. Empir. Res. Hum. Res. Ethics **12**(5), 372–382 (2017)
148. Vitak, J., Shilton, K., Ashktorab, Z.: Beyond the belmont principles: ethical challenges, practices, and beliefs in the online data research community. In: Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing, pp. 941–953 (2016)
149. Voorbij, H.: The value of librarything tags for academic libraries. Online Inf. Rev. **36**(2), 196–217 (2012)
150. Walsh, M., Antoniak, M.: The goodreads 'classics': a computational study of readers, amazon, and crowdsourced amateur criticism. J. Cult. Anal. **4**, 243–287 (2021)
151. Wan, M., McAuley, J.J.: Item recommendation on monotonic behavior chains. In: Pera, S., Ekstrand, M.D., Amatriain, X., O'Donovan, J. (eds.) Proceedings of the 12th ACM Conference on Recommender Systems, RecSys 2018, Vancouver, BC, Canada, 2–7 October 2018, pp. 86–94. ACM (2018). https://doi.org/10.1145/3240323.3240369
152. Wan, M., Misra, R., Nakashole, N., McAuley, J.J.: Fine-grained spoiler detection from large-scale review corpora. In: Korhonen, A., Traum, D.R., Màrquez, L. (eds.) Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, 28 July–2 August 2019, Volume 1: Long Papers, pp. 2605–2610. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/p19-1248

153. Wang, K., Liu, X., Han, Y.: Exploring goodreads reviews for book impact assessment. J. Informet. **13**(3), 874–886 (2019)
154. Wikipedia contributors: Internet archive. Wikipedia (2022). https://en.wikipedia.org/wiki/Internet_Archive
155. Wikipedia contributors: Personal information protection law of the people' s republic of china (2021). https://en.wikipedia.org/wiki/Personal_Information_Protection_Law_of_the_People%27s_Republic_of_China
156. Wikipedia contributors: Amazon books (2022). https://en.wikipedia.org/wiki/Amazon_Books
157. Wikipedia contributors: Amazon (company) (2022). https://en.wikipedia.org/wiki/Amazon_(company)
158. Wikipedia contributors: Goodreads (2022). https://en.wikipedia.org/wiki/Goodreads
159. Wikipedia contributors: Librarything (2022). https://en.wikipedia.org/wiki/LibraryThing
160. Wikipedia contributors: Wattpad (2022). https://en.wikipedia.org/wiki/Wattpad
161. World Intellectual Property Organization (WIPO): Wipo copyright treaty (1996). https://wipolex.wipo.int/en/text/295166
162. Worrall, A.: "like a real friendship": translation, coherence, and convergence of information values in librarything and goodreads. In: iConference 2015 Proceedings (2015)
163. Worrall, A.: "connections above and beyond": information, translation, and community boundaries in librarything and goodreads. J. Assoc. Inf. Sci. Technol. **70**(7), 742–753 (2019)
164. Zhang, C., Tong, T., Bu, Y.: Examining differences among book reviews from various online platforms. Online Inf. Rev. **43**(7), 1169–1187 (2019)
165. Zhou, Q., Zhang, C.: Relationship between scores and tags for Chinese books-in the case of douban book. J. Data Inf. Sci. **6**(4), 40 (2013)
166. Zhou, Q., Zhang, C., Zhao, S.X., Chen, B.: Measuring book impact based on the multi-granularity online review mining. Scientometrics **107**(3), 1435–1455 (2016). https://doi.org/10.1007/s11192-016-1930-5
167. Zimmer, M.: Addressing conceptual gaps in big data research ethics: an application of contextual integrity. Soc. Media+ Soc. **4**(2), 2056305118768300 (2018)
168. Zimmer, M.: "But the data is already public": on the ethics of research in Facebook. In: The Ethics of Information Technologies, pp. 229–241. Routledge (2020)
169. Zuccala, A.A., Verleysen, F.T., Cornacchia, R., Engels, T.C.: Altmetrics for the humanities: comparing goodreads reader ratings with citations to history books. Aslib J. Inf. Manag. (2015)