AUTHORS ALLIANCE

Text and Data Mining under U.S. Copyright Law: Landscape, Flaws & Recommendations

Last Revised October 2024

This report was primarily drafted by Christian Howard-Sukhil, the TDM Legal Fellow for Authors Alliance (Fall 2023), with editorial contributions from Authors Alliance staff.

I. Introduction **3**
II. Glossary **5**
III. Different Materials for a TDM Research Corpus **7**
A. The challenge of creating a TDM corpus 7
B. Out-of-copyright works 8
C. Digitization and OCR 9
D. Licensed datasets 10
E. HathiTrust 13
F. Breaking TPMs on Digital Copies 16
IV. Towards a Better TDM Exemption **17**
A. University affiliation required by the TDM exemption 18
B. Limitations on sharing a corpus of dataset with TPMs removed 18
C. Secure storage requirements 20
D. Licensing terms overriding the TDM exemption 22
V. A Better TDM Exemption Can Help Alleviate Other Problems **24**
A. A lack of guidance and legal certainty 24
B. Low risk-tolerance 25
C. Concerns about future regulations for Artificial Intelligence 26
VI. Future Research Questions **28**

I. Introduction

Authors Alliance[1] was founded in 2014 with the support of authors who recognize the transformative role that technology has in enabling new forms of research and creative expression. Text and Data Mining research, in particular, is capable of groundbreaking discoveries by uncovering new patterns from analyzing large datasets. Authors Alliance was the lead petitioner to obtain exemptions from the Library of Congress to allow circumventing technological protection measures (TPMs) on DVDs and ebooks for Text and Data Mining (TDM) research, and we are visiting university campuses nationwide to help researchers better understand the copyright laws surrounding TDM. This report aims to collect and document how researchers work within the current TDM legal framework in the United States[2].

For this report, Authors Alliance interviewed approximately 40 academic scholars and library support staff. Interviewees had a range of backgrounds that included graduate students, adjunct faculty, tenured and tenure-track faculty, digital scholarship librarians, copyright librarians, and library acquisition specialists. Additionally, interviewees represented a variety of disciplines and institution types. Interviewees' affiliations included the humanities, social sciences, computer and data sciences, and computational research departments. Institution types ranged from small liberal arts institutions, to mid-sized universities, and large R1 universities. The information collected in this report seeks to encompass the experiences of a variety of

[1] Authors Alliance is a 501(c)(3) nonprofit that exists to advance the interests of authors who want to serve the public good by sharing their creations broadly. Authors Alliance focuses on copyright and other information law issues, working both to educate authors so they can better understand their rights and to promote policies that make knowledge and culture more widely available and discoverable. Virtually all creative work builds upon the creativity of others, and Authors Alliance believes that it is critically important that researchers understand the legal rights they have to use cultural and historical materials in ways that allow them to study, learn, and share their own discoveries with the world.

[2] This report is made possible by the generous support from the Mellon Foundation's through its grant to Authors Alliance for its Text and Data Mining: Demonstrating Fair Use project.

stakeholders as they conduct or support TDM research.[3] The direct quotes excerpted from the interviews are italicized throughout the report for better readability.

The interviews that make up the foundation of this report took place between September and November 2023. At the time of the interviews, TDM research was subject to the TDM exemptions granted in 2021, which provided an exception to the anti-circumvention rules detailed in the Digital Millennium Copyright Act (DMCA). In October 2024, the Library of Congress renewed and expanded the TDM exemption, prompted by the petition submitted by Authors Alliance, the American Association of University Professors, and the Library Copyright Alliance.[4] Even though the new exemptions offer meaningful improvement (such as allowing access to existing corpora for new TDM projects) to the legal landscape, many same challenges facing TDM researchers continue to exist, and the takeaway from our 2023 interviews remain painfully relevant.

The report is organized so that readers can read it in its entirety or focus on specific sections of interest. In Part II, we define keywords and concepts that help establish a shared understanding of the field. In Part III, we delineate the most common sources scholars obtain materials to form

---

[3] We are particularly grateful to the following, non-exhaustive list of scholars and librarians: Rafael Alvarado, University of Virginia; Mark Algee-Hewitt, Stanford University; David Bamman, University of California, Berkeley; John Bell, Dartmouth College; Joel Burges, University of Rochester; Iliana Burgos, Cornell University; Allison Cooper, Bowdoin College; Kyle Courtney, Harvard University; Quinn Dombrowski, Association for Computers and the Humanities; Gabriel Egan, De Montfort University; Heather Froehlich, University of Arizona; Martin Gliserman, Rutgers University; Cody Hennesy, University of Minnesota; John Hunter, Bucknell University; Brandon Hurst, University of Connecticut; Jennifer Isasi, Pennsylvania State University; John Ladd, Washington and Jefferson College; Glen Layne-Worthey, HathiTrust Research Center; Hoyt Long, University of Chicago; Zack Marshall, University of Calgary; Peter McCracken, Cornell University; Dez Miller, Emory University; Kristin Moo, University of Rochester; Paige Morgan, University of Delaware; Edwin Roland, University of California, Santa Barbara; Anna Sackmann, University of California, Berkeley; Xanda Schofield, Harvey Mudd College; Emily Sherwood, University of Rochester; Stuart Shulman, Texifter; Dan Sinykin, Emory University; Todd Suomela, Bucknell University; Sarah Swanz, Washington University in St. Louis; Janet Swatscheno, HathiTrust Research Center; Laure Thompson, Princeton University; Lauren Tilton, University of Richmond; Ted Underwood, University of Illinois, Urbana-Champaign; Melanie Walsh, University of Washington; and Alex Wermer-Colan, Temple University. Please note that all opinions expressed are those of the individuals alone and do not necessarily represent the views of the institutions or organizations with which these individuals are affiliated. To protect the anonymity of our interviewees, all quotes in the report are unattributed, and pronouns may have been changed.

[4] *Authors Alliance and Allies Petition to Renew and Expand Text Data Mining Exemption*. Authors Alliance blog, (September 6, 2023), https://www.authorsalliance.org/2023/09/06/authors-alliance-and-allies-petition-to-renew-and-expand-text-data-mining-exemption/.

a corpus and the particular challenges associated with each source. In Part IV, we focus particularly on the limitations of the TDM exemptions granted by the Library of Congress pursuant to the DMCA Section 1201. In Part V, we cover some non-legal challenges facing TDM researchers that could be addressed in part by improving the current legal landscape.

II. Glossary

Copyright refers to the set of rules that apply to a work pursuant to the Copyright Act. These rules touch on when a user can use a work freely, such as fair use, and when a user must use a work with authorization from the rightsholder. When the set of copyright rules restrict the use of a work, we call it an **In-Copyright** work; when the rules no longer apply, we say the work is **Out-of-Copyright**.

Digital Millennium Copyright Act is a law passed in 1998 that contains civil and criminal penalties for the circumvention of TPMs and the sharing of tools that make circumvention possible. Liability under this law is independent of the underlying copyright infringement claim.

DMCA Section 1201 prohibits circumvention of technical protection measures ("TPMs"). It is generally interpreted to mean that circumvention of TPMs on copyrighted works is penalized, even when people circumvent TPMs to engage in non-infringing activities, such as fair use.[5] To address this overbroad suppression of First Amendment rights, Section 1201 also establishes a triennial rulemaking process whereby interested stakeholders can petition for new exemptions (or renewal of existing exemptions) to the prohibition on bypassing TPMs.[6]

Fair Use is a limitation to rightsholders' exclusive control over in-copyright works. Fair use is essential to copyright because it safeguards people's First Amendment right to free expression and allows for uses that support the goals of the U.S. Constitution to "promote the progress of science." When the use of an in-copyright work is fair, no permission is needed from the rightsholder. TDM research is often considered a fair use, because it is a transformative, non-consumptive use that does not affect the market of the original work.[7]

---

[5] There is some conflicting caselaw on this point. Cf. *Chamberlain Group v. Skylink Tech Inc.,* 381 F. 3d 1178, 1202 (Fed. Cir. 2004) ("17 U.S.C. § 1201 prohibits only forms of access that bear a reasonable relationship to the protections that the Copyright Act otherwise affords copyright owners.") *with* Universal City Studios, Inc. v. Corely, 273 F 3d. 429, 458-59 (2d Cir. 2001) (rejecting fair use as a defense in DMCA claim).

[6] 17 U.S.C. § 1201.

[7] Exemption to Prohibition on Circumvention of Copyright Protection Systems for Access Control Technologies, 86 Fed. Reg. 59,633, (October 28, 2021), (granting exemption as "likely … noninfringing" fair use).

OCR stands for Optical Character Recognition. It is a technology that allows conversion of scanned documents into machine-encoded text. OCR relies on the recognition of characters as well as the prediction of surrounding characters based on machine learning.

Text and Data Mining (TDM) encompasses a broad swath of research practices and can span numerous disciplines. In general, TDM involves computer-assisted analysis of large amounts of digital texts or data, such as articles, books, images, or film. A TDM project can require hundreds, if not thousands, of texts, images, or films for meaningful analysis. Researchers aim to extract useful information and insights from analyzing large volumes of text and data.

TDM Exemptions were last granted in October 2024 pursuant to the DMCA triennial rulemaking. With the TDM exemptions, academic scholars can circumvent TPMs in ebooks and films and obtain the raw text and data needed to conduct their TDM research. The 2024 TDM exemption for films, similar to the 2021 exemption, permits researchers affiliated with nonprofit institutions of higher education to bypass technical protection measures on motion pictures "on a DVD protected by Content Scramble, a Blu-ray disc protected by the Advanced Access Content System, or made available for digital download where … the copy … is lawfully acquired and owned by the institution, or licensed to the institution without a time limitation on access."[8] The TDM exemption for texts permits bypassing TPMs on literary works distributed electronically, "excluding computer programs and compilations that were compiled specifically for text and data mining purposes."[9]

TPM stands for a technical protection measure, such as encryption or DRM. TPMs are used by rightsholders to control access and deter copying. Because TDM research necessitates accessing and copying digital copies of works protected by TPMs, it is crucial that exemptions are made to DMCA section 1201 to allow scholars to circumvent TPMs. When works are encumbered with TPMs, it can also obstruct preservation and accessibility; other exemptions address this issue, such as the exemptions for software and video game preservation.

Triennial Rulemaking is the process by which the Library of Congress adopts exemptions to ameliorate the overbroadness of DMCA section 1201. At the conclusion of a triennial rulemaking process, the Librarian of Congress grants or denies petitions for exemptions, generally based on recommendations from the Register of Copyright. In order to obtain TDM exemptions during the ninth cycle, for exemptions effective during 2024–2027, a collective of non-profit organizations—Authors Alliance, the Library Copyright Alliance, and the American Association of University Professors—petitioned the U.S. Copyright Office in 2023.

[8] 37 C.F.R. § 201.40(b)(4)(i).
[9] 37 C.F.R. § 201.40(b)(5)(i).

III. Different Materials for a TDM Research Corpus

The first step of TDM research is usually to collect and clean large volumes of texts and structured data. TDM research can lead to transformative new discoveries in fields ranging from material science, engineering, literary scholarship, and the social sciences, when scholars have built usable corpora where they can extract and identify patterns among a wide variety of texts, images, and video.

The process of collecting a corpus of materials that is suitable for a proposed TDM research project can be grueling as well as costly. In this section, we will take a look at what this process is like in practice, the different ways scholars obtain materials for assembling their TDM corpus, and the unique challenges associated with each source of materials.

A. The challenge of creating a TDM corpus

A corpus for TDM research is usually assembled by:

- scanning and OCRing physical copies, such as books;
- copying digital works, such as ebooks or DVDs, which often requires breaking TPMs; and/or
- licensing and downloading existing datasets, such as those sold by publishers or other vendors.

The time and effort that it takes to collect a new corpus is considerable, especially because a corpus can span thousands of works. For example, one scholar we spoke with created a corpus by purchasing tens of thousands of ebooks, breaking TPMs on each ebook, and then cleaning the ebooks by removing the title, publication information, and other extra-textual information, such that only the literary text remained to make up the corpus. This scholar estimated that a thousand hours were spent to assemble this corpus before they could even begin any analysis let alone publish any discoveries.

Additional steps are often necessary to transform a corpus into something useful for TDM research. After collecting the pertinent volumes for assembling a corpus, scholars often must then mark up each volume according to a carefully determined schema[10]. Such a schema may identify chapter breaks, paragraph breaks, and even sentence boundaries. Additionally, it may include marking up, or "tagging," for outdated spelling, line indentations, obsolete verbiage, and alternate wordings, especially when scholars are working with older texts. One scholar estimated that it took two full months for a team of two to three research assistants (generally comprised of undergraduate and/or graduate students), each working five to ten hours per week, to adequately mark up a single volume in their corpus. This substantial time estimate was

[10] There was ambiguity on whether marking a corpora, such as by annotation, was allowed under the 2021 exemption, but the Register of Copyright clarified in the 2024 rulemaking that such use was indeed allowed. *Section 1201 Rulemaking: Ninth Triennial Proceeding to Determine Exemptions to the Prohibition on Circumvention. Recommendation of the Register of Copyrights.* October 2024.

for an already-trained group of research assistants. Training people who are new takes up additional time and funds. Even the hiring of undergraduate research assistants can easily add up to thousands of dollars per semester.

Beyond these general obstacles, different kinds of materials pose special challenges, as discussed below.

B. Out-of-copyright works

One may imagine that out-of-copyright works are low-hanging fruit when collecting materials for assembling a corpus because out-of-copyright works are not encumbered with any legal constraints. Yet other challenges abound.

One basic problem is physical access. TDM scholars experience difficulties—sometimes, insurmountable obstacles—when accessing and digitizing out-of-copyright works. One scholar explains that although some materials may be out-of-copyright, the physical copies are often rare or constitute personal records held in archives or special collections. Often, these records are not digitized. Even if they are digitized, those digitizations can be "*really dirty*," such that it is impossible to convert into machine-encoded text for TDM research.

Additional constraint on copying is imposed by the institution that owns the physical copy. A scholar points to what they describe as the "*most important*" manuscript archive for their area of study, which refuses any request to digitize files for TDM research. The only way to access these manuscripts is to go to the hosting archive to view the physical copies; scholars are further prohibited from making any digital copies of the unique archival materials. The archive refuses digitization on the grounds that they "*owned*" the texts, even though these are out-of-copyright materials that should contribute to everyone's collective understanding of knowledge and culture. The scholar expresses their frustration: "*This is anti-intellectual; [these manuscripts] belong to the public!*"

Some other restrictions come from the original donor. One librarian mentions that there can be additional issues related to access based on restrictions donors have placed on the donations. This librarian says: "*Sometimes the [donating] family will only give permission if [the scholar] pays a fee or compensates them in some way. And sometimes the family will also impose restrictions on how the materials can be used*."

One successful example of collecting and sharing an out-of-copyright corpus is the *EarlyPrint* project.[11] *EarlyPrint* is a huge online archive that seeks "to transform the early English print record, from 1473 to the early 1700s, into a linguistically annotated and deeply searchable text corpus."[12] Not only have the texts in *EarlyPrint* been digitized, a team of scholars has created an online schema to identify disparate spellings, line and paragraph breaks, and even part-of-speech information about various words. This allows the *EarlyPrint* Lab, an offshoot of

[11] EarlyPrint, https://earlyprint.org/ (last visited September 4, 2024).
[12] *Id.*

the *EarlyPrint* project, to visualize for TDM scholars certain aspects of the textual materials contained in the *EarlyPrint* holdings.[13]

The difficulty of accessing out-of-copyright materials is well documented. Interested parties should consult Kenneth D. Crews's article, "Museum Policies and Art Images: Conflicting Objectives and Copyright Overreaching."[14] To the extent hosting institutions assert copyright over their scans, at least in the U.S. it is clear that those claims are ineffective.[15] But, there is nothing to prevent institutions from restricting access, or placing paywall or contractual restrictions on access to their collections.

C. Digitization and OCR

Even when a request to access and digitize materials is granted, whether for in-copyright or out-of-copyright works, the physical materials still need to be processed and encoded to be usable for TDM research. Obtaining access and converting physical files into any digital format, such as by scanning or photographing, is not enough. Digital files also need to accurately convey the underlying contents (i.e., properly OCRed) for TDM research to be possible.

Cost can be a major obstacle to good-quality OCR. OCR software can be costly to license and is usually priced according to the volume of texts it processes. It is also costly to provide the OCR software with good-quality materials—this can include purchasing a physical copy of the book, cutting off the book's spine, scanning the pages of the book, and then manually correcting any errors that occur during the OCR process. For TDM scholars who are working with thousands or tens of thousands of books, this process is far too labor-intensive and fails to present itself as a viable option. Furthermore, when a corpus cannot be shared, each interested researcher must bear this cost individually, which not everyone can afford. As a result, many researchers are left without access to adequately OCRed corpora.

This is especially true for non-English texts. One scholar says: "*The novels that I study are not well-digitized in the first place, and those that are are not cleaned*." Indeed, most OCR technology in the United States has been developed in accordance with the English language syntax. About this process, one scholar explains: "*The reason that OCR works as well as it does is because it is using an underlying language model to predict the next word. [This prediction allows the OCR to] deal with the fact that sometimes there's a random ink mark or [space between letters]*." For non-English texts, however, this prediction model does not always work as intended. As a result, OCR can introduce more errors and inaccuracies, producing "dirty" machine-readable versions of a given non-English language text. A scholar laments: "*It is impossible to do the kind of analysis on [non-English] novels that [my peers perform on English-language novels] because I don't have access to clean texts!*"

---

[13] EarlyPrint Lab, https://earlyprint.org/lab/ (last visited September 4, 2024).
[14] Kenneth D. Crews, *Museum Policies and Art Images: Conflicting Objectives and Copyright Overreaching*, Keys for architectural history research in the digital era: Handbook (Juliette Hueber and Antonio Mendes da Silva, eds., 2014), http://books.openedition.org/inha/4924.
[15] *Bridgeman Art Libr., Ltd. v. Corel Corp.*, 36 F. Supp. 2d 191 (S.D.N.Y. 1999).

Given the high cost of OCR, many scholars rely on grant funding. But, it can be difficult to secure funding to support digitization and OCR when the materials in question cannot be shared freely. For example, some grant-funding organizations unfamiliar with Digital Humanities work didn't appreciate the importance of digitizing in-copyright texts for ingestion by HathiTrust when access to these texts are not made public. One scholar postulates that "*the obstacle of copyright was a huge factor in why [grant] reviewers refused to give money to digitize something that other scholars can't fully access online, especially when other projects [that use materials in the public domain] are able to upload all their materials online*."

D. Licensed datasets

In order to overcome the difficulties with access, digitization, and OCR, many TDM scholars choose to rely heavily on licensed datasets sold by vendors. Academic libraries pay large sums of money to vendors for licensed access to academic journals, newspaper databases, and similar sources.

In many cases, academic libraries have negotiated successfully for broad rights for their users, including clauses that protect users' fair use rights with respect to the licensed materials. And many publishers explicitly support TDM on any of their materials that a researcher has lawful access to, without insisting on a separate TDM license. However, increasingly, libraries are seeing vendors insist on clauses that either limit TDM uses in practice (e.g., clauses that forbid automated scraping) or specifically prohibit TDM uses unless the library purchases an additional TDM package.

Vendors of these sources have essentially created a two-tiered system whereby a fee is charged for accessing the materials, and an additional fee is charged for scholars to use the materials for TDM research (a.k.a., the "add-on TDM package"). As one librarian explains, in practice*, "you're … paying a second subscription for a second service to get the same information in a different format. It's absurd, but there's no other way around it*."

It should be noted that several librarians—while expressing exasperation at the TDM services provided by vendors—acknowledged that not everything about these services was bad. In some cases, vendors of these services will transform content into a more useful format, as one librarian explains how the vendor will transform contents *"into XML so you can work with them in Python, for instance."* As another librarian explains, vendor services can often make the "*storage and exportation [of data] easier from a copyright standpoint.*"

One drawback of relying on vendors for accessing TDM corpora is that licensed access may not be stable or permanent when compared to having a local corpus. For example, one researcher criticizes services such as these, in part because of this researcher's prior experiences with like-services. The researcher states: "*We've been burned in the past [by relying on services like these]. . . . Some of these databases break, and [when they break] there is no indication []. Relying on some mysterious black box is not a great way of proceeding*."

Beyond functionality, the double-dipping cost of these "add-on TDM packages" can be exorbitant. On average, academic libraries are shrinking in relative terms.[16] Depending on the size of the university and the nature of the vendor, librarians and scholars reported that these add-on TDM packages typically cost between $18,000 to $30,000 per vendor per year. "*Think about it like paying for a new car every year—and that is for each add-on package*," one librarian observes. Another laments: "*We're playing the game because we have to play the game, but it's frustrating. And the publishers know that they can make money doing this*."

Librarians at several institutions—including at some larger research universities—stated that they simply cannot afford the add-on TDM packages. In the words of one librarian: "*This is way beyond what we can budget for, so we will tell scholars to use grant money for [these services], if they have it*." This problem, according to another librarian, will continue to widen the gap between those universities that can afford these services and those that cannot afford access for their affiliated TDM scholars. In the words of one librarian: "*Our budgets are flat, so we can't add a new subscription without taking something away*." In fact, one librarian states that when considering whether to subscribe to a TDM add-on package, it is vital that multiple scholars at their institution express the need for such a service. They describe the situation: "*These [TDM add-on services] are expensive, so we consider how many people want access to it. . . . There need to be enough requests around this from different departments [and from] more than a single researcher*." Such criteria for assessing when to subscribe to a given TDM add-on package can severely disadvantage TDM scholars who are the only scholar at their institution conducting TDM research on a given topic.

TDM scholars and librarians are bound by the vendors' licensing terms because the consequences of ignoring the restrictive terms—both for general database subscriptions and for any add-on TDM packages—can hinder research activities. Some scholars try to avoid these TDM licensing costs by downloading the PDFs of journal articles or newspapers directly from the vendor's main database (i.e., without going through the add-on TDM package) and then converting those PDFs to a useful format for their research. It is not uncommon for vendors to restrict downloads of their PDFs, for example by restricting downloads to 200 PDFs in a 24-hour period.[17] Should a researcher exceed this restriction, the vendor may cut off service to the entire university. One librarian says that one of their vendors has a provision stating that people

[16] Joshua Kim, *Three Questions on Academic Library Spending for the Scholar Who Wrote the Book on University Budgets*, Inside Higher Ed. (Jan. 26, 2023), https://www.insidehighered.com/blogs/learning-innovation/3-questions-academic-library-budgets-assessment-and-planning-librarian ("Library budgets have not kept pace with the escalation of database prices. For-profit scholarly publishing has one of the highest profit margins of any industry. Publishers are able to raise prices almost without limit because the demand for information is inelastic. … For most academic libraries, materials and personnel are already bare bones, so there is little to cut in a crisis if the library is to continue to function.")

[17] Multiple researchers discussed needing thousands of articles to conduct meaningful TDM analysis, so a 200 article cap in a 24-hour period can easily slow down and frustrate their research aims.

at their institution can only download as many articles as is "humanly possible." The institution has been unsuccessful at pushing back against this vague language, and as such, "*we're being shut down over and over by the vendor for bulk downloads.*" Another librarian states: "*We can't risk losing our institutional license for something because one of our faculty members decides to do a massive download [for TDM].*" Multiple librarians from both small and large institutions cite the ability of vendors to simply cut off services to their entire campus as a major issue. Worse still, is one example of researchers who had been required to retract their study because their TDM research was not conducted on a properly licensed database.[18]

Additionally, because each of these vendors have their own contract, librarians[19] often have to wade through dozens of vendor contracts to determine what a researcher can or cannot do for a given research project. In the words of one librarian: "*It's really frustrating. . . . We have to try to navigate whatever restrictions and explicit prohibitions we receive from providers*." On top of this, vendors have increasingly started using their own proprietary TDM platforms which are almost always incompatible with one another. On this issue, another librarian states: "*The new platforms that allow TDM research often allow researchers to bring in data from other sources, but you can't share the data you get from the platforms with other platforms.*" As such, it is extraordinarily difficult—if not impossible—to analyze data sold by different vendors at the same time for a given research project.

The upshot of all of these licensing deal related challenges is that librarians often have to work with researchers to make changes to—and in many cases, to curtail—their research projects to fit the resources available at the institution. One librarian says: "*I tend to direct people to resources that we already have legal access to.*" Another discusses having to "*walk back*" the researcher from their proposed topic. This librarian states, "*I've had to tell researchers, 'you*

---

[18] Shantanu Dutta, Ashok Kumar, Moumita Dutta, and Caolan Walsh published an article titled, *Tracking COVID-19 vaccine hesitancy and logistical challenges: A machine learning approach* in PLoS One (July 2021). Post-publication, it came to light that the authors had not received appropriate permissions to mine certain data that they used. As a result, they had to retract their article: "The authors obtained news articles for this study on Factiva. While the authors represented to PLOS that they had legitimate permissions to access the articles, concerns were noted post-publication that the authors' data mining of news articles on Factiva did not comply with the terms of the University of Ottawa's license with Factiva. Therefore, the authors retract this article." *Retraction: Tracking COVID-19 vaccine hesitancy and logistical challenges: A machine learning approach*, National Library of Medicine, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8297835/ (July 22, 2021); see also Adam Marcus, *'A Very Unfortunate Event': Paper on COVID-19 Vaccine Hesitancy Retracted*, Retraction Watch (July 30, 2021), https://retractionwatch.com/2021/07/30/a-very-unfortunate-event-paper-on-covid-19-vaccine-hesitancy-retracted/ (describing retraction process).

[19] In some cases, researchers could independently sift through these contracts, but from our interviews, it seems that most researchers go to their librarians with questions about accessing these resources. As such, this problem more often falls on the librarians' shoulders.

need to change your topic because what you want isn't available unless you have endless funds to be able to purchase additional access from these aggregators and publishers.'"

There are several ongoing efforts to assist librarians and other stakeholders navigate these licensing issues. "EResource Licensing Explained" is a "a comprehensive guide for librarians and library professionals who license electronic resources (eResources)."[20] Another useful source is Peter McCracken and Emma Raub's article, "Licensing Challenges Associated With Text and Data Mining: How Do We Get Our Patrons What They Need?"[21] Additionally, stakeholders should consider getting involved in the eBook Study Group, which is "a new emerging coalition of state legislators, librarians, and library stakeholders in numerous states [who] are recommending the adoption of state law based on consumer protection, contract law, and contract preemption to regulate library ebook contracts with publishers."[22]

E. HathiTrust

Beyond a build-it-yourself corpus or a licensed database, scholars have access to one unique resource, HathiTrust (HT) and its affiliated HathiTrust Research Center (HTRC), that is worth describing separately here. Many of the scholars whom we interviewed had either directly worked with HTRC or were familiar with it. One scholar says: "*Working with [the HTRC] is the only way I can get access to the works I need*." Another states that the HTRC "*is the main way to legally do [TDM] research*." And in the words of a third: " *I'm really glad that [the HTRC] is there; it allows me to do a lot.*"

Since the conclusion of the interviews, we have learned that the HTRC will discontinue its services in 2026.[23] We are keeping this section untouched both because for now HTRC remains a valuable tool for TDM scholars, and also because the information included here could shed light on the development of future tools that would replace HTRC.

HTRC was specifically devised to assist text-based TDM research by drawing on HT's "collection of millions of titles digitized from libraries around the world."[24] The collaboration among HT member libraries resulted in over 18 million digitized volumes that span topics from philosophy

---

[20] *eResource Licensing Explained* (R. Samberg, K. Zimmerman, S. Teremi, and S. Enimil, eds., 2023) https://librarylicensingguide.pubpub.org/ (last visited September 4, 2024).
[21] Peter McCracken and Emma Raub, *Licensing Challenges Associated With Text and Data Mining: How Do We Get Our Patrons What They Need?*, Journal of Librarianship and Scholarly Communication Volume 11 Issue 1, (February 3, 2023), https://doi.org/10.31274/jlsc.15530.
[22] eBook Study Group, https://www.ebookstudygroup.org/ (last visited September 4, 2024).
[23] *Plans for the HathiTrust Research Center, October 10, 2024*, HathiTrust, https://www.hathitrust.org/press-post/plans-for-hathitrust-research-center/ (last visited October 31, 2024).
[24] *About HathiTrust Research Center*, HathiTrust, https://www.hathitrust.org/about/research-center/ (last visited September 4, 2024).

and world history to technology and literature.[25]

When scholars wish to conduct TDM research on the massive in-copyright collections in HT, they are able to do so based on the HTRC Non-Consumptive Use Policy. The Policy in many ways tracks the reasoning and holding of two important fair use cases, *Authors Guild v. HathiTrust* and *Authors Guild v. Google*.[26] The Policy states that the HTRC aims to "facilitat[e] the widest possible variety of non-consumptive research and educational use with the HT collection while remaining clearly within the bounds of the fair use rights courts have recognized as applying to this type of activity. More generally, the policy aims to achieve the same goals as copyright itself: to promote progress in the discovery and spread of knowledge, without harming the commercial interests of authors, publishers, and other stakeholders."[27]

Because of HTRC's requirement of "non-consumptive use" when working with in-copyright texts, researchers often rely on the existing TDM tools provided by HTRC. Scholars are allowed to use a given corpus in three ways, which HTRC describes broadly as: limited access, transformed access, and capsule access.

Limited access employs "web-accessible data analysis and visualization tools" on a given corpus assembled by a researcher from HT's digital holdings.[28] These tools "allow researchers to assemble collections (worksets) of volumes, and analyze them using the HTRC supported off-the-shelf algorithms and visualization interfaces."[29] Throughout this process, the researcher is unable to view any substantial portion of text from any volume. In fact, one individual from the HTRC described this as "*even less than snippet view*."

Transformed access largely takes the form of derived datasets. One of the most popular tools available via transformed access is called the "HTRC Extracted Features dataset." This dataset is "derived from bibliographic and paratextual metadata and includes part-of-speech-tagged unigram counts."[30] Describing this process, one individual at HTRC says: "*Every volume [of the researcher's chosen corpus] is represented by a JSON file, and every page [of every volume] … is depicted as a bag of words.*" Each "*bag of words*" contains the complete text of the given page, yet the words on that page are rendered in a randomized order such that no one could reconstruct the actual text. Thus, while the researcher is able to download these derived

[25] *About the Collection*, HathiTrust, https://www.hathitrust.org/the-collection/ (last visited September 4, 2024).
[26] *Authors Guild, Inc. v. HathiTrust*, 755 F.3d 87 (2d Cir. 2014); and *Authors Guild v. Google, Inc.*, 804 F.3d 202 (2d Cir. 2015).
[27] *Non-Consumptive Use Policy*, HathiTrust, https://www.hathitrust.org/the-collection/terms-conditions/non-consumptive-use-policy/ (last visited September 4, 2024).
[28] *Id.*
[29] *Id.*
[30] *Id.*

datasets, the researcher is not able to view or access any substantial portion of the text in sequence.

Capsule access occurs through HTRC "data capsules." Essentially, these data capsules are a "system that grants a user access to a virtual machine which is a dedicated, secure desktop environment (called a "Capsule") that exists within the HTRC's secure computer environment … through which a user can carry out non-consumptive research on HT collection using the HTRC-provided or their own data analysis and visualization tools."[31] In other words, within the secure computing environment—the capsule—a researcher will have full access to the texts within their corpus. A researcher can then use tools provided by HTRC or else bring their own Python or other code to the capsule in order to analyze the texts. Once the researcher has analyzed the texts, he or she will be able to export a set of files out of the capsule. These files are manually reviewed by HTRC staff to ensure that the researcher is not "exporting substantial portions of text." In short, only transformed data is able to be exported out of the data capsules.[32]

HTRC Data Capsules can be tricky to learn and use. As one scholar explains, "*HathiTrust built the right model, but it's just not a good software.*" An individual at the HTRC agrees that "*almost nothing is quite as user-friendly as it ought to be. This [problem] is compounded by the fact that we didn't have any UX designers when the Data Capsule environment was created*." In fact, use of the data capsules often requires scholars to "use a command-line tool to get data," which presents additional burdens on researchers—particularly those who do not have a computer programming background or extensive tech support. The HTRC has been unable to address this issue because of a lack of funding to update how their system operates: "*We don't have funding to do a revamp and make things more streamlined.*" Additionally, individuals at the HTRC recognize that, "*because of the secure way we do the data capsules, there's a ton of lag.*" Nonetheless, the HTRC enables scholars to mount the data capsules onto a supercomputer, which can address some of these computing issues. About this option, one individual at the HTRC explains: "*We do this for big projects; we're happy to do this for researchers if they need it, but we don't offer this as a regular service*." Currently, this service is restricted to scholars who are part of HT member libraries.

HTRC can be a limited solution in another way: resources and tools available through the HTRC are only as good as the data contained in the HT digital library. Because the HT digital library is drawn from works held in academic libraries, the digital library naturally "mirrors an academic

[31] *Id.*

[32] There are additional security measures in place to ensure that data capsules are not read for consumptive purposes and that "improper outputs (e.g., leaks) are prevented." See *Non-Consumptive Use Policy*, supra note 26. Additionally, to use a data capsule, the researcher must be affiliated with a HT-member library; the researcher must also have his or her own .edu email address.

library collection."[33] Additionally, not all books by all HT member libraries are digitized or sent to the HT digital library; rather, the digital library only contains those books that member libraries chose to digitize and send to HT. As such, the digital library may contain duplicates of numerous popular books (such as Jane Austen's Pride and Prejudice) even as there are gaps in other areas. For instance, the HTRC has acknowledged that it has a "dearth of romance novels" and its holdings contain "[g]aps in speculative fiction, for example, we've noticed that some very prominent Black female authors' books are missing from the collections."[34]

Many modern works in general are missing from HT collections. One scholar states: "*Not everything is in HathiTrust. Trying to make claims around twenty-first century literature is pretty difficult because the library is limited, especially with current fiction.*" Another scholar summarizes the issue: "*There's a lot of twentieth and twenty-first century literature that is not in HathiTrust because libraries don't want to digitize in-copyright works. . . . Libraries [should] team up and scan every twentieth century and twenty-first century book to be ingested into HathiTrust. But libraries aren't doing that. And so there's a lot of twentieth and twenty-first century literature that's not in HathiTrust, and there's not really a place to go find it.*"

While Hathitrust offers an important platform for conducting TDM research, it is important to recognize that we need continued robust support both within HTRC and in the broader TDM legal framework. One scholar comments: "*[The HTRC has the] infrastructure and people to support [scholars], but not enough of either*." An individual at the HTRC similarly acknowledges: "*[Although I believe] the HTRC is a success story, . . . I don't want to imply that we don't need [the help of advocacy organizations] to make [copyright and the legal framework better] because I think so many of the things that we do and the ways we do them are workarounds*."

F. Breaking TPMs on Digital Copies

Researchers may also use existing digital copies to build their TDM corpus. Sometimes, digital copies are the only viable options, such as when film scholars conduct TDM on DVDs. Sometimes, using digital copies has the advantage of avoiding the time, expenses and potential for error introduced by scanning and OCRing texts. Using digital copies can also avoid some of the challenges associated with licensed TDM access from vendors, and the limitations inherent in the HathiTrust corpus. However, these existing digital copies, such as ebooks and DVDs, are often restricted by TPMs that prevent copying.

The TDM exemption granted in 2021 allows researchers to bypass TPMs on both films and electronic texts such as ebooks. Scholars working with each of these media indicated unique issues that they faced. While the 2024 TDM exemption made minor adjustments that addressed a few of these issues, we expect many to continue.

---

[33] Janet Swatscheno, *Introduction to HathiTrust and HTRC Tools for Text Data Mining,* University of Toronto TDM in Libraries Colloquium (2023), https://mdlutoronto.github.io/TDM2023/schedule/7.%20Janet%20Swatscheno.pdf.
[34] *Id.*

One of the biggest challenges for scholars working with text-based media relates to the problems inherent in licensing ebooks and other textual media, particularly given the possibility of contractual restrictions on TDM research. Both popular, consumer ebooks and those related more specifically to academic literature and journals have contractual language that seeks to override the TDM exemption and fair use rights researchers are granted by law. Film scholars working with online streaming media experience similar problems, where licensing terms of the streaming sites restrict TDM research. We will discuss the problem with licensing terms overriding the TDM exemption in more detail in Section IV. D.

Film scholars encounter a different problem when working with physical copies. The prohibitively large size of the media assets and associated artifacts presents a recurring challenge. It is essential for film scholars to store high-fidelity large-sized copies when creating a corpus, because TDM often focuses on details that would be otherwise lost in a smaller-sized file. A scholar emphasizes that for their work, the "*precision of the images is important [because I'm] trying to analyze background. So the size and crispness of the images really matters*."

Multiple film scholars state that, while their university provides free online storage solutions for professors, these storage solutions can be limited. In fact, one scholar says that their university imposes "*limits on how large a given file can be*" when using the university storage system. As a result, this scholar often has to "*rip the DVD at a lower [fidelity] level.*" A librarian similarly encountered these limits, stating: "*Our storage solutions are pretty woeful: We have new storage caps on Box and Drive. For a while, we had unlimited caps on these, but now, people are reduced to having 150 GB total*." Scholars who have data that exceeds the free storage solutions offered by their universities generally have to pay for their own storage, which can be "*very pricey*." One film scholar says: "*Film and photography take up huge amounts of data—terabytes and terabytes. The costs can be exponential.*" Another scholar says that within their university eco-system, they can purchase extra storage; but even so, the storage costs can be prohibitively expensive, costing around $500 per month for 12 terabytes of storage.[35]

IV. Towards a Better TDM Exemption

Numerous scholars laud the TDM exemptions as "*game-changing*" in allowing them to conduct their research on TPM-restricted digital materials. Interviewees working on film studies, who had no legally authorized way to gain access to films for TDM research prior to the exemption universally expressed their gratefulness that the exemption now exists. One scholar comments, "*We now have a pretty good exemption for those using moving images*," though this scholar clarified that some confusions and frustration points continue to exist with the current regulatory framework.

[35] Textual TDM researchers also complained of storage costs for their materials, though the scale of such costs is often much greater for film scholars. Several text-based scholars who were working on large corpora cited storage costs as a "*big issue,*" and one explicitly stated: "*I've had to negotiate for startup money for cloud computing storage for the data*."

A. University affiliation required by the TDM exemption

While the TDM exemptions granted in 2021 and 2024 were a win for the academic community, the exemptions came with stringent requirements. One condition of the TDM exemptions requires university affiliation for researchers working on TDM projects. Independent scholars who are not affiliated with a university cannot legally break TPMs under DMCA section 1201; they are effectively cut out from conducting TDM research on many in-copyright works. Because the current DMCA legal framework in the US does not enable non-university-affiliated TDM research, our interviews do not cover the perspective of these independent scholars. We suspect it is challenging to even find independent TDM scholars in the U.S., given the impracticalities arising out of the current restrictions imposed by the DMCA. This is an area worth further investigation.

B. Limitations on sharing a corpus of dataset with TPMs removed

The 2021 TDM exemption states that researchers may share their corpus with other "researchers affiliated with other institutions of higher education solely for purposes of collaboration or replication of the research."[36] As clarified in the Recommendation of the Register of Copyright during the 9th triennial rulemaking, "collaboration" in the 2021 exemptions does not permit corpora to be used for independent research projects, and the Office's interpretation of "access" only permits a corpus to reside in the institution that compiled it while outside researchers cannot receive a copy of the corpus even if their own institutions have bought and/or licensed the same materials.[37]

The new 2024 TDM exemptions offer the possibility of improved collaboration. Although it remains true that corpora cannot be shared among institutions, researchers can now access a TDM corpus hosted in another institution even when their own institutions do not have legal access to the copyrighted works included in the corpus. Furthermore, when accessing these corpora hosted in other institutions, researchers are now allowed to work on independent new projects, instead of being limited to collaborating on the same projects for which the corpora were compiled.

The 2024 exemptions could greatly facilitate TDM research by removing the need to duplicate existing corpora. During the 2023 interviews, we found out that scholars had trouble finding grants that would fund TDM projects that work with in-copyright materials that have been digitized before. Interviewees are skeptical about the patience or continued support of grant institutions to fund similar projects over and over, when those projects would call for the (re)digitization of a given book or corpus. In the words of one scholar: "*How many people spend money to digitize the same book? Who wants to fund the same thing over and over*?" A librarian similarly states: "*It's just really interesting to me that we're spending a ton of money and a ton of time building out these datasets, but then we can't make those datasets public for other*

[36] 37 C.F.R. § 201.40(b)(4)(i)(D) (2021).

[37] *Section 1201 Rulemaking: Ninth Triennial Proceeding to Determine Exemptions to the Prohibition on Circumvention. Recommendation of the Register of Copyrights,* p. 82-85. October 2024.

people to use. . . . We're just putting a ton of money into one corpus that one person at one institution can use once." In sum, scholars and librarians alike say it is simply "*not financially sustainable if everyone has to rebuild the corpus themselves.*" The 2024 exemptions open up many new opportunities for scholars to utilize existing corpora.

However, many of the same practical difficulties will remain despite the new expansion, particularly because actual copies of the corpus cannot be duplicated and transferred for use at other institutions. As mentioned before, some grant-funding organizations are unwilling to digitize in-copyright texts when broader access to these texts is not made available to the public.

Furthermore, even under the 2024 exemptions, outside researchers must rely on the goodwill and willingness of host institutions to grant access to their systems. Several interviewed scholars and librarians lament the difficulties of accessing corpus for legitimate and legally-sanctioned research purposes because of the complex security protocols surrounding the storage of a corpus. For those working in ultra-secure environments (such as institutions with needs to store patient information), their collaborators at other institutions are simply not allowed access to corpus stored in these environments. As such, it is all but impossible for researchers who are compelled to work in these ultra-secure environments to collaborate with scholars outside of their institution. One scholar states: "*This [ultra-secure environment] is not conducive to research, and [the data we work with in such environments] is not shareable*."

For instance, one librarian says that they are often approached by scholars seeking to collaborate with scholars at other institutions, yet in order to share the relevant corpus, the librarian would need to create a special access account to a service provided by the university. Researchers regularly ask this librarian to simply provide such an account with necessary access privileges to their collaborators, to which this librarian replies: "*Unless their institution has signed a license [to the given platform], we can't allow them to have their own account*." Similarly, another librarian says that it is "*all but impossible to get someone who is not institutionally affiliated to have access through a Shibboleth login*." This librarian explains that in cases of cross-institutional collaboration, the corpus "*almost always has to live outside the institutional infrastructure in things like Google Drive or Dropbox.*" These practical challenges can complicate compliance with the secure data storage requirement in the TDM exemptions, often making cross-institutional collaborations impractical.

As one scholar plainly states: "*I'd like to be able to give this dataset to other scholars, and possibly to put it up on some central repository*." The free sharing of corpora remains crucial for constructing a better ecosystem for TDM research. A new proposal in the next triennial rulemaking with "logistical details" on who "would be responsible for ensuring that recipient institutions and researchers use effective security measures to safeguard [the sharing of

corpora]"[38] may be able to achieve this goal.[39]

C. Secure storage requirements

The TDM exemptions contain vague language around the security requirements on data storage. The exemptions require that the TDM researcher's institution employ: "effective security measures," which is defined as "security measures that have been agreed to by interested copyright owners of [motion pictures and literary works] and institutions of higher education; or, in the absence of such measures, those measures that the institution uses to keep its own highly confidential information secure."[40] At the time of this report, copyright owners and institutions of higher education are not known to have developed or agreed upon terms that constitute "effective security measures." As such, each scholar employs the measures that its own institution uses to secure its own "highly confidential information."

There are two major challenges with this formulation of the security requirement. First, universities employ a wide range of measures to protect their highly confidential information. These measures are commensurate with the nature of the "highly confidential information" stored by each university. Smaller universities, for instance, count student data, including students' names, ID numbers, addresses, and other personal information, among the most sensitive data that they possess, but may nevertheless allow liberal access for university affiliates to information classified as sensitive. Large research institutions that are affiliated with a hospital system or that may have a research group with a government contract, however,

---

[38] *Section 1201 Rulemaking: Ninth Triennial Proceeding to Determine Exemptions to the Prohibition on Circumvention. Recommendation of the Register of Copyrights,* p. 90. October 2024.

[39] Researchers who wrote letters in support of the petition described a multitude of exciting projects, and have built "a rich set of corpora to study, such as a collection of fiction written by African American writers, a collection of books banned in the United States, and a curated corpus of movies and television with an 'emphasis on racial, ethnic, sexual, and gender diversity.'" Many of those who wrote letters in support of our petitions recounted requests they have received from other researchers to use their corpora, and who were frustrated that the exemption's prohibition on non-collaborative sharing and their limited capacity for collaboration prevented them from sharing these corpora. *Authors Alliance Submits Long-Form Comment to Copyright Office in Support of Petition to Expand Existing Text and Data Mining Exemption*, Authors Alliance Blog, (January 29, 2024), https://www.authorsalliance.org/2024/01/29/authors-alliance-submits-long-form-comment-to-copyright-office-in-support-of-petition-to-expand-existing-text-and-data-mining-exemption/#:~:text=To%20recap%3A%20our%20expansion%20petitions,of%20researchers%20qualifies%20under%20the

[40] 37 C.F.R. § 201.40(b)(4)(ii)(B). *Exemption to Prohibition on Circumvention of Copyright Protection Systems for Access Control Technologies*, 86 Fed. Reg. 206, §§ 4(ii)(B) & 5(ii)(B) (Oct. 28, 2021).

impose way more stringent security protocols for "highly confidential information." This could include, for example, requiring researchers to treat digital copies of DVDs used for TDM research with the same strict standards as highly confidential medical information or even classified government information. The security requirements for TDM scholars vary based upon the activities of the university with which they are affiliated, and TDM scholars report divergent experiences working with the security requirements of their respective institutions.

Researchers at larger institutions with heightened security requirements felt exasperated by the burden of the security requirement. One researcher, for instance, states that their university is "*way more locked down with anything that has to do with medical texts and data*." As such, when the same strict security measures used for medical data are used as the security measures for TDM corpora, the "[*required*] *data security guarantees can be difficult to work with*." Another scholar explains: "*Our TDM data is stored in [the university's ultra-secure platform] with extra layers of security. Yet due to the processing within the ecosystem of [this ultra-secure platform], we don't have access to the same kind of computing resources that we ordinarily would be able to use to process the data.*" A third scholar says: "*The net time that is spent trying to on-board people [to work in these ultra-secure environments] is huge; it can take up to a semester to get approval to work on the data. And once [you do receive approval to work on] the data, there are huge technical issues. For instance, when working on the data in that environment, you can't install something that you may find important, and you can't share results with people. Research suffers from this environment*."

The industry trade associations are staunchly denying the reality that the security storage requirements and the impracticalities associated with sharing corpora are significant barriers preventing many researchers from conducting lawful TDM research.[41] For example, in its comment to repeal the current TDM exemption, the Association of American Publishers (representing industry giants such as Elsevier) states that TDM researchers are currently free to collaborate with each other without any realistic challenges.[42] Worse still, the Motion Picture Association—a trade association with members such as Netflix, Paramount, Sony, Universal, Disney, etc.—advances a similar line of argument, using the lack of lawsuits against TDM researchers not as evidence showing rigorous self-policing within the TDM research community,

[41] MPA, N/MA, and RIAA Opposition Comment in response to Section 1201 Exemptions to Prohibition Against Circumvention of Technological Measures Protecting Copyrighted Works, https://www.copyright.gov/1201/2024/comments/opposition/Class%203%28a%29%20and%203%28b%29%20-%20Opp%27n%20-%20Joint%20Creators.pdf (last visited September 4, 2024). AAP Opposition Comment in response to Section 1201 Exemptions to Prohibition Against Circumvention of Technological Measures Protecting Copyrighted Works, https://www.copyright.gov/1201/2024/comments/opposition/Class%203%28b%29%20-%20Opp%27n%20-%20Association%20of%20American%20Publishers.pdf (last visited September 4, 2024).

[42] *AAP Opposition Comment*, supra note 39 at P.3.

but to argue that TDM researchers currently enjoy unbridled freedom when it comes to collecting and sharing corpora.[43]

Beyond dismissing the difficulties facing TDM scholars within the current legal framework, the industry associations also sought to exploit the TDM exemptions' disclosure requirement to terrorize TDM scholars and their universities. As the renewal and expansion of TDM exemptions was being proposed by Authors Alliance and others, AAP sent TDM researchers letters demanding the written disclosure of all details related to security protocols related to all TDM projects, including the training of personnels, relevant contracts, description of alarm systems upon intrusion, and more, at the same time providing a stringent two weeks' window for a response.[44] The letters showed little understanding of how research management works at large universities, and specifically referred to the researchers' lawful involvement in the rulemaking process as the reason the demands were being made. Though it may be speculation that the trade associations intended this, one predictable result of sending such letters was to deepen TDM researchers' fear that TDM is legally risky and not worthwhile. Thankfully, the 2021 TDM exemptions require that institutions disclose their security protocols when requested by "a copyright owner whose work is contained in the corpus,"[45] and TDM researchers had no legal obligations to respond to the demands of the AAP and MPA when they failed to point to any in-copyright works of theirs that are included in a corpus.

The 2024 exemptions now clarify that trade associations may request the disclosure of security protocols, but only may do so on behalf of an "author"[46] and only based on reasonable belief that their members' in-copyright materials are included in a corpus. It remains to be seen if such a lax standard (that only requires the belief of a rightsholder) will be as detrimental to fair use and free expression as the notice-and-takedown process has been for the past two decades.

D. Licensing terms overriding the TDM exemption

Many popular consumer ebook vendors have language in their licensing contracts that prevent anyone—scholars included—from breaking TPMs, even when scholars intend to use the ebook texts for TDM research at a qualified institution allowable under the TDM exemption. Prominent examples of problematic licensing contracts include Amazon Kindle—which is one of the best sources for ebooks, as well as Apple iBooks and Barnes & Noble Nook.

Most humanities TDM scholars working with literary texts need access to the digital works published by these consumer ebook vendors. As one librarian puts it: "*Most of the things that people want to text-mine are ebooks, [which are] not in services like ProQuest [—a popular*

[43] *MPA, N/MA, and RIAA Opposition Comment*, supra note 39 at P.4-5.
[44] For an example of the letter sent by trade associations, see *AAP Opposition Comment*, supra note 39 at P.21.
[45] 37 CFR § 201.40(b)(5)(ii)(B).
[46] We are unsure of why the regulations allow for such requests from trade associations who represent authors, rather than rightsholders. This may be a point to seek clarification from the Copyright Office.

academic textual resource] that would allow users to pay an additional licensing fee for TDM research." To the best of our knowledge, the vast majority of consumer ebook vendors do not offer separate licensing agreements with academic libraries or institutions; rather, the public licensing terms—with language that prohibit TDM and/or breaking TPMs—are crafted to restrict all kinds of uses. At least one librarian (at a large research institution) confirms that their institution believes "*none of the popular [ebook] titles allow any kind of mining in any form at all*."

A risk-free way for overcoming this contractual override issue is to find ebooks through a vendor that does allow scholars to break TPMs. Yet partly because of the consolidation of ebook companies,[47] and partly due to the exclusive deals authors have with publishers, it is difficult for scholars to find alternative ebook offerings of the titles that they need. In fact, we estimated that Amazon Kindle has captured approximately 81% of the market share for popular consumer ebooks.[48] Other than some small indie publishers, only one major ebook vendor arguably has licensing terms that allow academic scholars to circumvent TPMs. For one scholar, this particular ebook vendor has presented a viable option for obtaining digital copies of books, but for many others, the vendor simply does not have the ebook offerings that scholars need for their research.

Almost every scholar working on text-based projects bemoaned the contractual override issue, citing it as a major obstacle to their TDM research. This is because, as one scholar says: "*sources like Amazon specifically say you can't break [TPMs] to do [TDM] analysis*." Another scholar laments: "*I'd love to make use of this exemption, but effectively being able to do so is dead in the water because of Amazon's Terms of Service*." A third expresses similar sentiments: "*I'm so thrilled that we got the exemption, and yet I don't actually see the exemption as anything stronger than just symbolic because of licensing culture*." Many scholars have "*not used the [TDM exemption for ebooks] at all.*"

---

[47] *See, e.g.*, Frederic Iardinois, *Consolidation in the EBook Market: Amazon Acquires Stanza*, Readwrite (April 27, 2009), https://readwrite.com/consolidation_in_the_ebook_market_amazon_acquires/; Katelyn Mirabelli, *The Consolidation of Book Publishing in the US: A Network Graph Study*, School of Information, Pratt Institute (May 11, 2021), https://studentwork.prattsi.org/infovis/visualization/the-consolidation-of-book-publishing-in-the-us-a-network-graph-study/; and *The Consolidation of Publishing Houses, Past and Present*, Authors Alliance blog (Dec. 8, 2021), https://www.authorsalliance.org/2021/12/08/the-consolidation-of-publishing-houses-past-and-present/.

[48] Dave Hansen and Rachel Brooke, *The scale and scope of contractual override of fair use in ebooks and streaming movies*, Protecting User Rights From Contractual Override Symposium, American University & Association of Research Libraries, May 18, 2023, https://docs.google.com/presentation/d/1cX6WhHuE2OAo3ASzybV-qM5XlzS9wMUSxbfaHoCMfgE/edit#slide=id.p.

The TDM exemption on its own does not guarantee researchers' ability to conduct TDM research on fictional and literary works. One scholar aptly articulates the shared frustration: "*It should not be possible for a private company to stop me from my research when [my research] does not imperil their sales.*"

In general, contractual limitations placed on TDM research are prevalent and robust. Worse yet, the restrictive impact is particularly acute in areas where licensing terms abound yet copyright laws afford TDM scholars no safeguard for conducting TDM on in-copyright materials. For example, the 2021 as well as the 2024 exemption for films only applies to films that are "on a DVD, … on a Blu-ray disc, … or made available for digital download."[49] It does not extend to streaming content, a limitation which is becoming increasingly problematic for film scholars. One film scholar explains that "*Netflix has never done a physical medium release of anything it has produced in-house. And HBO Max and a few other streamers have pulled things off of their services. In these cases, this content will never be available to anyone*." Another film scholar says that "*companies aren't necessarily holding onto all the [films and TV shows] that they've made*." A third scholar explains: "*We're in a situation where our only access to the history—certainly of electronic and digital media—is online. We need to think about—and scholars need to be able to work on—our cultural heritage. But we can't do this if we don't have access to those materials.*" Film scholars interviewed universally urged policymakers to allow scholars to access these materials for TDM purposes.

V. A Better TDM Exemption Can Help Alleviate Other Problems

In conducting the interviews, we discovered the many non-legal problems facing TDM scholars. Many of these problems cannot be directly fixed by a change in law or policy, however, we believe that by addressing the issues raised in Section III, the non-legal difficulties discussed below could be ameliorated.

A. A lack of guidance and legal certainty

A good law should be predictable and easily understandable for those affected by it. In the previous section, we discussed the major flaws of the TDM exemptions, revealing how the exemptions fall short in safeguarding TDM research activities. These negative effects caused by the flaws of the exemptions are further compounded by the misunderstandings of TDM researchers. Many TDM researchers, not unreasonably based on their anecdotal experiences, perceive copyright laws regulating TDM are stringent and unrelenting.

Based on the interviews we conducted, we notice that there is a general lack of understanding for the current TDM exemptions. Many scholars struggled to articulate the contours of the exemptions. Of particular note were misunderstandings about the necessity to destroy data after a given TDM research project has concluded; there is, in fact, no such requirement in the language of the exemptions, though multiple scholars believed this language to exist.[50] Another common misconception was that the exemptions did not apply to teaching, but only to TDM

[49] 37 C.F.R. § 201.40(b)(4)(i).

[50] Previous rulemaking record

scholars actively conducting research. The exemptions, however, apply broadly to "scholarly research and teaching."[51] When the legal framework for TDM is complex, ambiguous, and changing every three years, and when TDM as a research methodology is challenging to learn and master, it becomes indispensable to have specialized support staff to assist TDM scholars.

At many institutions—particularly smaller institutions—there is simply no one available with the requisite copyright background to provide legal support to scholars who are engaging in TDM research. A librarian at one such small institution says that at their university, "*there's no one who has a background in the legal area of copyright. Even the university general counsel is just that: a general counsel who covers a lot of things. We don't have a dedicated copyright attorney who can offer guidance on [TDM research or copyright questions more broadly].*" Another librarian says: "*Many of the people who want to do TDM have very little copyright training and education, and they have almost zero grasp of the [potential legal] issues. This positions them to be passive and reactive rather than proactive.*" For their part, librarians—even if trained in copyright law—are not able to offer legal advice on the best course of action for a TDM researcher based on his specific project. In the words of one librarian: "*I don't give legal advice, so I simply tell [scholars] that there are definite legal risks associated with [various approaches to conducting TDM research].*"

There is also a general lack of support staff to meet the needs of scholars conducting TDM research. For instance, one librarian states: "*There are really only two of us who have technical skills to help out [with TDM projects]. . . . Our ability to help on some of these major projects is limited because of lack of personnel.*" Another librarian says: "*There aren't enough librarians at institutions to help support this work in the first place, [and] I don't think a lot of the library leadership understands the [need for TDM support].*"

B. Low risk-tolerance

In general, many digital humanities scholars (and to some extent, their institutions) have a low tolerance for risk, which negatively impacts TDM research practices. One scholar states that in their experience "*universities are incredibly risk-averse [when it comes to potential] legal issues*." Multiple other scholars and librarians expressed similar sentiments during the interviews.

One of the most serious consequences of this perception of low risk-tolerance is the sheer amount of scholars who abandon valid and valuable research projects over TDM-related concerns. For instance, one researcher states that they know many scholars who "*just cut their losses and do research on things that are less risky.*" A librarian estimates that of the scholars who approach them with an idea for a TDM project, ninety percent abandon their projects because of the amount of work that goes into creating a representative dataset and the uncertainty regarding copyright issues.

[51] 37 C.F.R. § 201.40(b)(4)(i)(A) and37 C.F.R. § 201.40(b)(5)(i)(A).

The view is echoed by many that the current legal framework only allows scholars to choose two out of these three desired things: (1) legal certainty, (2) TDM as methodology, and (3) quality materials to study. The problem is most pronounced for graduate students because they are particularly restrained on money and time. Graduate students "*are steered toward corpora that are [already] available. This is sad because a lot of students are interested in contemporary materials [not already compiled, cleaned, and shared in a corpus], but they are steered elsewhere*." Scholars often have to choose between studying a given time period (such as 20th or 21st century literature) or adopting TDM as their methodology. One such interviewed scholar had shifted their focus to the 19th century in order to pursue their chosen methodology. This scholar says: "*I'm mostly a post-[19]45 scholar; that was always what I really wanted to do. But partly what made me go to the nineteenth century was [access to materials provided by] the public domain.*"

Scholars who do choose to continue with their TDM projects despite the current TDM legal framework generally express one of two attitudes. Either they have taken a legally conservative approach limiting their research to licensed materials from vendors, or they prioritize the quality of their dataset and work under legal uncertainties. A scholar in the former group says that they simply "*use under-representative datasets*" because that is all they can legally obtain. Another admits: "*My entire research program has been constrained by the fact that I don't want to be the test case in a court battle.*" A scholar in the latter group, on the other hand, describes the "*price*" they pay: "*I don't know that I feel constrained in the research questions that I ask, but the price of that is an intense paranoia about my legal exposure*."

C. Concerns about future regulations for Artificial Intelligence

Because of a lack of robust long-term legal protection for conducting TDM research, many interviewees are concerned that their ability to do TDM work will entirely depend on how the public's sentiment sways in the ongoing debate about Artificial Intelligence (AI). More specifically, researchers are concerned that efforts—whether legal or otherwise—to constrain AI's ingestion of in-copyright works will affect their ability to use in-copyright works for TDM research. There is, in fact, evidence that in the public's eyes, AI training and TDM research are already conflated, to the detriment of TDM scholars. Multiple scholars point to the case of ProseCraft, a literary analytics tool that performed sentiment analysis—a fairly common kind of TDM analysis—on in-copyright works.[52] Because of pushback from authors,[53] the creator of

[52] See Benji Smith, *Taking Down Prosecraft.io*, Medium (Aug. 7, 2023), https://blog.shaxpir.com/taking-down-prosecraft-io-37e189797121; and *Prosecraft, Text and Data Mining, and the Law*, Authors Alliance (Aug. 14, 2023), https://www.authorsalliance.org/2023/08/14/prosecraft-text-and-data-mining-and-the-law/.

[53] Kate Knibbs, *Why the Great AI Backlash Came for a Tiny Startup You've Probably Never Heard Of*, Wired (Aug. 14, 2023), https://www.wired.com/story/prosecraft-backlash-writers-ai/ (quoting a tweet from author Hari Kunzru: "This company Prosecraft appears to have stolen a lot of books, trained an AI, and are now offering a service based on that data. …I did not consent to this use of my work.").

Prosecraft removed the tool and issued an apology.[54] About this incident, one scholar comments: "*There has been and is still so much confusion, both on the part of creators and text and data miners, around the division between AI and TDM work.*"

Granted, there is some overlap between AI training and TDM research, but there are also important distinctions. Rachael Samberg, the Scholarly Communication Officer at the University of California Berkeley Library, explains: "Not all TDM research methodologies necessitate the usage of AI systems. . . . In other cases, though, scholars must employ machine learning techniques to train AI models before the models can make a variety of [TDM] assessments."[55] While it is believed by many copyright experts that the use of in-copyright works to train AI models is a fair use, at the time of this report, no court has definitely made this determination; so it remains unclear what kinds of TDM is considered legal and what illegal.

Should courts issue over-broad opinions restricting fair use, or Congress or State Legislatures issue sweeping laws or regulations prohibiting the ingestion of in-copyright works for AI models,[56] scholars may be forestalled from conducting TDM research. Samberg has advocated that it should be considered fair use to utilize AI technology to conduct TDM research: "For the same reasons that the TDM process is fair use of copyrighted works, the training of AI tools to do that TDM should also be fair use, in large part because training does not reproduce or communicate the underlying copyrighted works to the public."[57]

Fortunately, some of the initial confusion and worries are ameliorated, as policy makers become more informed on machine learning technologies. During the ninth triennial rulemaking, the Register of Copyright unequivocally stated that noncommercial TDM research is a fair use, no matter how the AI debate may shake out.[58]

---

[54] See Smith, *Taking Down Prosecraft.io*, supra note 49. Though, part of the issue with Prosecraft was that the creator had illegally obtained most of the books he analyzed from book-pirating websites. See Knibbs, *Why the Great AI Backlash Came for a Tiny Startup You've Probably Never Heard Of*, supra note 50. This incident suggests that authors (and perhaps non-academics more broadly) confuse TDM and TDM-like analyses with the rise of generative AI systems.

[55] Rachael Samberg, *UC Berkeley Library to Copyright Office: Protect fair uses in AI training for research and education*, Berkeley Library Update (Oct. 24, 2023), https://update.lib.berkeley.edu/2023/10/24/uc-berkeley-library-to-copyright-office-protect-fair-uses-in-ai-training-for-research-and-education/.

[56] See *Artificial Intelligence and Copyright*, 88 Fed. Reg. 59942 (Aug. 30, 2023), https://www.copyright.gov/ai/docs/Federal-Register-Document-Artificial-Intelligence-and-Copyright-NOI.pdf.

[57] Samberg, *UC Berkeley Library to Copyright Office: Protect fair uses in AI training for research and education*, supra note 52.

[58] *Section 1201 Rulemaking: Ninth Triennial Proceeding Recommendation of the Register of Copyrights*, at p. 75.

VI. Future Research Questions

This report documents the major challenges facing TDM researchers today and provides a general analysis on how copyright law interacts with TDM research. We realize there are many unanswered questions still. We would like to flag some of the areas that could benefit from future research.

Should the TDM exemptions be permanently available?

At least one librarian we interviewed was particularly concerned about the three-year renewal process built into the TDM exemption. More specifically, they were reluctant to teach scholars and graduate students how to circumvent TPMs because they viewed the three-year "limit" on the exemption with precarity. They argue: "*The exemption is short-term; what happens if the exemption isn't renewed?*" More in-depth analysis is needed for the practical and legal implications for making the TDM exemptions permanent. Right to repair may be a good comparison here, where a growing number of states are legislating to establish a right to repair, though it is not clear how they are meaningfully enforced.

Should scholars with no university affiliation be included in the TDM exemptions?

Independent scholars who are not affiliated with a university cannot legally break TPMs under current TDM exemptions. They are effectively cut out from conducting TDM research on many born-digital in-copyright works. More research could be done to investigate what kind of research needs these individuals have and how they differ from university-affiliated scholars. This would be important background information if independent scholars were to seek future expansion of the TDM exemptions to cover their research needs.

What are the legal ramifications of licensing agreements that ostensibly do not allow scholars to rely on the TDM exemption or fair use?

Contractual override of users' rights–that is, contracts that limit a users' right to engage in fair uses–is an important topic for copyright law. In the U.S., termination rights are thought to be the only rights under copyright law that cannot be waived or altered by contractual terms in advance. By contrast, most scholars believe that private parties have unlimited freedom to contract away the legal safeguards provided by fair use and TDM exemptions. However, it is not entirely clear how private contracts interact with these users' rights. It would benefit TDM scholars to have more clarity on the legal ramifications of licensing agreements that ostensibly do not allow scholars to rely on the TDM exemption or fair use.

What are the most effective strategies for enabling access to out-of-copyright content that TDM researchers wish to use?

Legally speaking, when a work's copyright term expires, the work becomes free for everyone to use. Many say such an out-of-copyright work "enters the public domain" and becomes an integral part of the shared pool of human culture and knowledge that future authors can add to and build upon. Practically, though, many such out-of-copyright materials are not free to access, with many important works behind paywalls or subject to other restrictions. How might law or policy adapt to better safeguard or promote access to out-of-copyright work? What are some

other collaborative or decentralized approaches we could take to promote the use of out-of-copyright works?

Would it be possible to create a public-interest corpus of TV shows, films, ebooks, and other textual materials, that can be used for both TDM research and AI training?
Numerous scholars expressed their desire for a centralized repository that houses any needed research materials; these research materials should be free of any DRM software. Ideally, scholars would receive access to these research materials upon showing that they are affiliated with a university and that they will use the materials only for research and teaching purposes. One researcher states: "*The Library of Congress [should] run some program to verify scholars and allow them access to in-copyright work without DRM software*." A librarian describes such a system as allowing researchers to have "front door" access to the materials. They state: "*The exemption allows academic researchers to break technical locks. I'd rather allow academic researchers to get in through the front door*." In short, scholars are eager for a system where they can access TV shows, films, ebooks, and other textual materials that both recognizes the value of their TDM research and enables them to conduct such research with reasonable (rather than excessive) barriers to accessing the materials.

How can we address the challenges associated with streaming media?
Some scholars advocated for a partnership between TDM scholars and industry, particularly related to the preservation of TV shows, films, and streaming materials more broadly. One scholar, for instance, suggested that Paramount, Disney, Netflix, Amazon Studios, and other studios and production companies create "*a research center that academics could use to study film and TV*." The researcher explained: "*We want to make sure not everything vanishes, and that really important things stay discoverable.*" What are some benefits or drawbacks to a database created and maintained by the content industry? Are there other alternatives to solving the problem with preserving and studying streaming media?

How can we address the challenges associated with social media platforms?
Whose permission is needed, if any, to study internet users' tweets, vlogs, photos, and so on? Should social media platforms adopt X's approach to sell licenses for scholars to conduct TDM research on user-generated content, without any of the profit trickling down to the content creators? Is the TDM of user-generated content a clear case of fair use?