

Electronic Submission via www.regulations.gov

September 28, 2026

**Opposition of the Association of American Publishers to the Petition for Renewal of
the Exemption for Text and Data Mining for Literary Works**

The Association of American Publishers (AAP) objects to the joint proposal filed by the Authors Alliance and the Library Copyright Alliance (together, “Petitioners”) to renew the text and data mining (“TDM”) exemption for literary works as currently codified in 37 C.F.R. § 201.40 (b)(5).¹

Since the last rulemaking, relevant circumstances have changed. Market options tailored for TDM research have multiplied. Additionally, at the time of the previous triennial rulemaking, the rapid growth of AI technologies was only just beginning,² and its impact and reach then were largely unknown.³ Although TDM differs from generative AI in numerous ways, both involve “the creation and curation of datasets.”⁴ The unprecedented increase in demand for large corpora of textual works has turbocharged the licensing market for computational uses of such works, but at the same time has increased

¹ AAP is a not-for-profit trade association representing the leading book, journal, and education publishers in the United States on matters of law and policy. AAP’s members range from major commercial book and journal publishers, to small, non-profit, university and scholarly presses, to leading publishers of educational materials and digital learning platforms, all of which are dedicated to contributing diverse, high-quality content for the benefit and enjoyment of the public. AAP has objected to the grant and renewal of the exemption to circumvention for noninfringing TDM research and teaching in the previous two triennial rulemaking cycles.

² See *generally* Artificial Intelligence and Copyright, 88 Fed. Reg. 59942 (Aug. 30, 2023) (describing rapid developments in generative AI technologies and relevant emerging issues over 2022 and 2023 and asking the public to provide factual information and views on certain copyright policy questions).

³ See, e.g., Alexander Buck, Adam Blandin & David J. Deming, *The Rapid Adoption of Generative AI* 1 (Nat’l Bureau of Econ. Rsch., Working Paper No. 32966, 2025), <https://doi.org/10.3386/w32966> (observing “[u]ncertainty over the impact of genAI on the economy”); Molly Kinder, Xavier de Souza Briggs, Mark Muro & Sifan Liu, *Generative AI, the American Worker, and the Future of Work*, BROOKINGS (Oct. 10, 2024), <https://www.brookings.edu/articles/generative-ai-the-american-worker-and-the-future-of-work/> (noting that “there is little consensus on the nature and scale of generative AI’s potential impacts and how—or even whether—to respond”).

⁴ U.S. COPYRIGHT OFF., COPYRIGHT AND ARTIFICIAL INTELLIGENCE, PART 3: GENERATIVE AI TRAINING 76 (May 2025) (pre-publication version).

copyright owner security concerns regarding technical protections of works and their inclusion in large corpora and reinvigorated large scale piracy of textual works, compounding the harm from such piracy.⁵

Because prior rulemaking records have become unreliable, AAP urges the Copyright Office to decline streamlined renewal of the exemption, require Petitioners to develop a fuller record of evidence, and carefully reconsider the exemption in the subsequent full rulemaking process.

AAP sets out its basis for objection in more detail below.

I. Licensing and market solutions for TDM research continue to emerge and expand, obviating the need for circumvention and diminishing circumvention prohibitions’ adverse effects on researchers’ ability to conduct TDM research and teaching and, thus, the need for the exemption.

To recommend an exemption, the Register of Copyrights must find that “users are either adversely affected, or are likely to be adversely affected, in their ability to make noninfringing uses during the next three years,” and proponents must show that “circumvention is necessary to avoid the alleged adverse effects.”⁶ Adhering to this standard, the Register recommended granting the exemption for TDM research and teaching in 2021 and renewing it in 2024 based on rulemaking records that showed a lack of available alternatives to circumvention.⁷ Specifically, the Register found significant limitations in then-existing TDM solutions, such as the scope of available material, server processing capacity, and other technological restrictions.⁸ Any direct licensing agreements identified in 2024 were also found unsuitable for TDM research because they pertained to

⁵ See Margaux Poueymirou & Maxwell V. Pritt, *Why AI’s Use of Shadow Libraries Should Alarm Us All*, DAILY J. (June 16, 2025), <https://www.dailyjournal.com/articles/386130-why-ai-s-use-of-shadow-libraries-should-alarm-us-all> (“Not too long ago, ‘shadow-libraries’ were dying. . . . Then came AI. Virtually all major companies building LLMs contacted [Anna’s Archive] to train on [its] data.”).

⁶ U.S. COPYRIGHT OFF., SECTION 1201 RULEMAKING: NINTH TRIENNIAL PROCEEDING TO DETERMINE EXEMPTIONS TO THE PROHIBITION ON CIRCUMVENTION: RECOMMENDATION OF THE REGISTER OF COPYRIGHTS 16 (Oct. 2024) [hereinafter 2024 RECOMMENDATION] (citing U.S. COPYRIGHT OFF., SECTION 1201 OF TITLE 17, at 115 (2017)).

⁷ See *id.* at 88; U.S. COPYRIGHT OFF., SECTION 1201 RULEMAKING: EIGHTH TRIENNIAL PROCEEDING TO DETERMINE EXEMPTIONS TO THE PROHIBITION ON CIRCUMVENTION: RECOMMENDATION OF THE REGISTER OF COPYRIGHTS 112-13, 117-19 (Oct. 2021) [hereinafter 2021 RECOMMENDATION].

⁸ See 2021 RECOMMENDATION, *supra* note 7, at 117-18 (explaining that HathiTrust was not an “adequate alternative” due to the “substantial gaps in its collection,” limited server processing capabilities, and security measures in place that “limit researchers’ ability to test and refine” TDM algorithms); 2024 RECOMMENDATION, *supra* note 6, at 88 (finding that Copyright Clearance Center’s RightFind was not a “serious alternative to circumvention . . . due to the relatively limited scope of materials it licenses”).

generative AI rather than “the kinds of research projects at issue.”⁹ Moreover, in 2024, an expansion to the exemption was granted to allow third-party institutional researchers access to TDM corpora prepared by another institution. This expansion was based on the finding that, without it, cross-institutional collaboration and efforts to pursue different projects on an existing corpus would be hampered, hindering “research that reflects more diverse viewpoints, methods, and subjects.”¹⁰

These prior records have become unreliable for streamlined renewal of the exemption as market options and solutions designed for TDM research have substantially increased. For one, more and more TDM licenses and solutions are directly available through publishers. To name a few, Wiley offers an API service and related access tools specifically to facilitate TDM research, introducing a TDM Client Python package in June 2025 to enhance researchers’ experience with the API.¹¹ Similarly, Springer Nature offers a variety of APIs and direct metadata delivery to facilitate TDM activities in addition to including TDM permission as part of its institutional e-book subscriptions at no additional cost.¹² Elsevier also offers API access for TDM research.¹³ In short, there is far less of a shortage in licensing options through which research institutions can obtain literary works and related products tailored for TDM research.

Moreover, ProQuest TDM Studio provides a platform through which publishers can license their works and researchers can access large collections of diverse content. It currently hosts and makes available millions of rights-cleared documents.¹⁴ A wide array of prominent nonprofit research institutions of higher education have already adopted ProQuest TDM Studio as a library service to facilitate TDM research, including Columbia University, University of Washington, University of Chicago, Brown University, and University of Oxford.¹⁵ Since the prior rulemaking, that number has grown, with institutions

⁹ 2024 RECOMMENDATION, *supra* note 6, at 88.

¹⁰ *Id.* at 94-95.

¹¹ See *Text and Data Mining*, WILEY ONLINE LIBR., <https://onlinelibrary.wiley.com/library-info/resources/text-and-datamining> (last visited Sep. 22, 2026).

¹² See *Text and Data Mining at Springer Nature*, SPRINGER NATURE, <https://www.springernature.com/gp/researchers/text-and-data-mining> (last visited Sep. 22, 2026).

¹³ See *Text and Data Mining*, ELSEVIER, <https://www.elsevier.com/about/policies-and-standards/text-and-data-mining> (last visited Sep. 22, 2026).

¹⁴ See *TDM Studio*, CLARIVATE PROQUEST, <https://about.proquest.com/en/products-services/TDM-Studio/> (last visited Sep. 22, 2026) (“TDM Studio provides access to millions of rights-cleared publications to enable timely interrogation using text analysis and visualization tools.”).

¹⁵ See, e.g., *Text Mining: ProQuest TDM Studio*, COLUMBIA UNIV. LIBRS., <https://guides.library.columbia.edu/text-mining/proquest-tdm-studio> (last visited Sep. 22, 2026); *ProQuest TDM Studio Guide: About TDM Studio*, UNIV. OF WASH. LIBRS.,

such as University of Missouri,¹⁶ Rutgers University,¹⁷ University of North Carolina at Chapel Hill,¹⁸ and University of Virginia¹⁹ providing TDM Studio to their researchers during the past three years.

In 2025, ProQuest moved its History Vault product to the ProQuest platform, integrating into its TDM Studio over 25 million pages of diverse historical archival materials “sourced from renowned institutions,” including academic institutions and state and local archives across the U.S.²⁰ Further, ProQuest TDM Studio’s Workbench enables cross-institutional collaboration, allowing researchers to share TDM corpora and derive distinct analyses.²¹ In sum, ProQuest TDM Studio’s centralization of rights-cleared material, and

https://guides.lib.uw.edu/tdmstudio?utm_source=chatgpt.com (last visited Sep. 22, 2026); *Database: ProQuest TDM (Text Data Mining)*, BROWN UNIV. LIBR., <https://libguides.brown.edu/ProQuestTDM> (last visited Sep. 22, 2026); *ProQuest TDM Studio: Home*, UNIV. OF CHICAGO LIBR., <http://guides.lib.uchicago.edu/tdmstudio> (last visited Sep. 22, 2026); *Text and Data Mining: TDM-Ready Collections*, UNIV. OF OXFORD: BODLEIAN LIBRS., https://libguides.bodleian.ox.ac.uk/TDM/current_collections (last visited Sep. 22, 2026); *TDM Studio*, STANFORD GRADUATE SCH. OF BUS.: LIBR., <https://gsb-research-help.stanford.edu/library/faq/338678> (last visited Sep. 22, 2026); *Text and Data Mining with TDM Studio: Home*, CARNEGIE MELLON UNIV. LIBRS., <https://guides.library.cmu.edu/TDMStudio> (last visited Sep. 22, 2026).

¹⁶ See Shannon Cary, *New Text and Data Mining Tool Available to Mizzou Researchers*, UNIV. OF MO. LIBRS. (Aug. 11, 2026), <https://library.missouri.edu/news/resources-and-services/new-text-and-data-mining-tool-available-to-mizzou-researchers>.

¹⁷ See *New This Year: Vastly Expanded ProQuest Content, Plus Text and Data Mining*, RUTGERS UNIV. LIBRS. (Oct. 13, 2025), <https://www.libraries.rutgers.edu/news/new-year-vastly-expanded-proquest-content-plus-text-and-data-mining>.

¹⁸ See *TDM Studio Makes Text and Data Mining Simple*, UNIV. OF N.C. AT CHAPEL HILL: UNIV. LIBR. (Sep. 5, 2025), <https://library.unc.edu/news/tdm-studio-makes-text-and-data-mining-simple/> (announcing the new availability of ProQuest TDM Studio through the UNC Chapel Hill library).

¹⁹ See *Fall 2026: Welcome and Library 101!*, UNIV. OF VA. LIBR. (Sep. 10, 2026), <https://library.virginia.edu/news/2026/fall-2026-welcome-and-library-101> (announcing the new availability ProQuest TDM Studio through the University of Virginia library this year).

²⁰ *History Vault Sets a New Milestone in Accessing Archival Collections*, CLARIVATE PROQUEST: BLOG (Feb. 3, 2025), <https://about.proquest.com/en/blog/2025/history-vault-sets-a-new-milestone-in-accessing-archival-collections/> (stating that the collections were sourced from institutions such as “the National Archives and Records Administration (NARA), the Library of Congress, National Association for the Advancement of Colored People (NAACP), Wisconsin Historical Society, Catholic University of America, Chicago History Museum, the Southern Historical Collection of the University of North Carolina at Chapel Hill,” in addition to the “numerous state and local archives across the United States”).

²¹ See *Supplemental Terms of Use – TDM Studio*, CLARIVATE PROQUEST (July 6, 2026), <https://about.proquest.com/en/about/supplemental-terms-of-use-tdm-studio/> (“Authorized Users are expressly allowed within the designated TDM Studio Workbench to create derived data from the textual content of the eligible databases. . . . Customer may designate collaborating scholars from other institutions as Authorized Users with respect to a specific TDM Studio Workbench.”); CLARIVATE, PROQUEST TDM STUDIO: FREQUENTLY ASKED QUESTIONS 7, https://proquest.libguides.com/ld.php?content_id=74086531 (last visited Sep. 28, 2026) (stating that each Workbench “supports up to 5 researchers at one time” and that it is a “completely shared space” where “any datasets, scripts, or output files can be seen and edited by any other Workbench member”); see also ProQuest TDM Studio, HARVARD BUS. SCH.: BAKER LIBR.,

the continued growth in adoption, have improved the accessibility of wide-ranging works, enabling researchers to find and utilize content with more ease than before, marking a clear shift away from the circumstances described in previous rulemaking records. As a result, the need for the 2024 expansion also has been diminished.

Despite these developments, Petitioners again claim that, while “[c]ommercially licensed text and data mining products continue to be made available to research institutions. . . . [T]hese licensed products do not allow researchers to license the full array of texts needed to engage in the research they seek to do, and those works remain protected from copying by technological protection measures.”²² Because this claim is virtually identical to that made in Petitioners’ renewal petition in the previous rulemaking cycle when alternatives to the exemption were found to be limited, further clarification and explanation are necessary.²³ The full array of TDM products available for nonprofit research and individual publishers who permit nonprofit TDM research on their titles has not been developed in previous rulemaking records. Accordingly, streamlined renewal of the exemption should not be granted.

II. New evidence surrounding the growing practice of illicitly acquired works for building both TDM and generative AI training corpora illuminates the need to reconsider and revise the exemption.

New evidence of datasets built from works acquired through unlawful means also raises concerns about the exemption’s current language and formulation. To prevent the exemption from contributing to unintended consequences, closer scrutiny is warranted.

<https://www.library.hbs.edu/databases-cases-and-more/databases/proquest-tdm-studio> (last visited Sep. 22, 2026) (“Research collaborators can share Workbenches across institutions.”).

²² Authors All. & Libr. Copyright All., Tenth Triennial Rulemaking: Literary Works – TDM Renewal Pet. (Aug. 24, 2026), <https://www.copyright.gov/1201/2027/petitions/renewal/Renewal-Pet-TDM-Authors-Alliance-and-Library-Copyright-Alliance.pdf>.

²³ Authors All., Libr. Copyright All. & Am. Ass’n of Univ. Professors, Ninth Triennial Rulemaking: Literary Works – TDM Renewal Pet. (July 7, 2023), <https://www.copyright.gov/1201/2024/petitions/renewal/Renewal Pet. - TDM - Authors Alliance, Inc., Library Copyright Alliance, and American Association of University Professors.pdf> (“Commercially licensed text and data mining products continue to be made available to research institutions, as they were at the time of the 2021 exemption and as is reflected in the existing record, but these licensed products do not allow researchers to license the full array of texts and films that are needed to engage in the research they seek to do, which remain protected from copying by the TPM technologies that the current exemption specifies.”).

a. The existing exemption is providing inadvertent cover for the use of illicitly acquired works.

While the existing exemption demands that TDM work corpora only consist of literary works “lawfully acquired and owned by the institution,” meaning that they must be obtained through purchases of legitimate physical copies or non-time-limited digital licenses,²⁴ recent evidence shows that this requirement is practically ineffective. Instead, as currently written, the existing exemption is contributing to the use of unlawfully acquired literary works by providing inadvertent cover for institutions to do so.

Indeed, as reported in an August 2026 Journal of Cultural Analytics article, institutional researchers in France recently compiled an open data catalog “MajinBook” for “computational social science and cultural analytics” by scraping pirate sites such as LibGen and Z-Library, linking metadata from those sites with structured bibliographic data from third parties, and making available corpora in multiple languages.²⁵ Notably, the goal of the project was to “unlock the full potential” of existing pirate sites for researchers.²⁶ In justifying their data acquisition, the researchers named the existing TDM anticircumvention exemption as a legal framework through which the legality of “harvesting and computing data from shadow libraries” can be considered.²⁷ In other words, the existing TDM anticircumvention exemption, as written, has unintentionally facilitated a project that may encourage the further acquisition of infringing works from pirate sites. This fact was not considered in previous rulemakings, and it is relevant to the potential negative impact that circumvention of technological measures may have on the market for or value of copyrighted works.

²⁴ See 2021 RECOMMENDATION, *supra* note 7, at 121 (“The Register recommends that circumvention be permitted only on copies of the copyrighted works that were lawfully acquired and that the institution owns or for which it has a non-time-limited license. Circumvention is not permitted to access copies that the institution has rented or borrowed.”); Transcript of United States Copyright Office Section 1201 Public Hearings, at 335 (Apr. 7, 2021) (J. Matthew Williams, Joint Creators and Copyright Owners), <https://www.copyright.gov/1201/2021/hearing-transcripts/210407-Section-1201-Public-Hearing-Class-15-7a-7b.pdf> (“On the ‘lawfully obtained’ point, I think it needs to be made clear that the obtained copy is a copy that essentially is a distributed copy . . . it really should be limited to copies obtained from a licensed distribution, such as a DVD sale, or to a downloadable copy that is not a temporary copy.”).

²⁵ See Antoine Mazières & Thierry Poibeau, *MajinBook: An Open Catalogue of Digitally Mediated World Literature*, J. CULTURAL ANALYTICS (Aug. 6, 2026), <https://culturalanalytics.org/article/id/1164/>.

²⁶ *Id.*

²⁷ See *id.* (“Lastly, harvesting and computing data from shadow libraries as we did can be considered through the lens of several legal frameworks: In the EU, the above-mentioned ‘TDM exception’, and in the US, the fair use doctrine incorporated into the Copyright Act of 1976 complemented by the recent Text and Data Mining exemption to the Digital Millennium Copyright Act (1998).”).

b. The growth of generative AI has increased the demand for large corpora of textual works, raising concerns over the security of unprotected corpora created under the exemption.

The demand for large corpora of textual works only continues to grow given the continuing rapid evolution and adoption of generative AI technologies. Unfortunately, such needs for data have been increasingly fulfilled through unlawful means. Over the past few years, pirate datasets of literary works have become far too easily available, and allegations of AI developers acquiring datasets through scraping pirate sites continue to surface in AI copyright infringement lawsuits.²⁸ Because TDM research and generative AI training both require obtaining and curating literary works at large scale, corpora compiled for TDM research under the exemption can also be valuable for generative AI training. Indeed, due to the increasing prevalence of datasets acquired through unlawful means and the continued use of generative AI technologies in connection with TDM research,²⁹ concerns that unprotected corpora created under the exemption will be siphoned off to train generative AI models are heightened and deserve full consideration in this rulemaking.

Therefore, we urge the Office to undertake a full review of the existing exemption, rather than simply renewing the same. Indeed, the Office should consider revising the exemption to strengthen the security measures protecting TDM work corpora compiled under the exemption to prevent them from being used for generative AI training. The Office should also consider making explicit the distinction between using existing generative AI tools to perform analytical tasks as part of TDM research and training a generative AI model in the first instance with a corpus assembled under this exemption, a distinction the Office identified in its 2024 Recommendation.³⁰ Expressly stating the prohibition in the

²⁸ See, e.g., Melissa Korn & Alexandra Bruell, ‘Sketchy AF’: What to Know About How OpenAI Staff Discussed Book-Pirating, WALL ST. J. (Sep. 19, 2026), <https://www.wsj.com/tech/ai/sketchy-af-what-to-know-about-how-openai-staff-discussed-book-pirating-fdc788a1> (“Starting in 2019, OpenAI began using books from file-sharing site Library Genesis—LibGen for short—to help train its large language model.”); *Kadrey v. Meta Platforms, Inc.*, 788 F. Supp. 3d 1026 (N.D. Cal. 2025); *Bartz v. Anthropic PBC*, 787 F. Supp. 3d 1007 (N.D. Cal. 2025); *In re Mosaic LLM Litig.*, No. 3:24-cv-01451 (N.D. Cal. consolidated Dec. 2, 2024); *In re Google Generative AI Copyright Litig.*, No. 5:23-cv-03440 (N.D. Cal. consolidated Oct. 28, 2024); *Nazemian v. NVIDIA Corp.*, No. 4:24-cv-01454 (N.D. Cal. filed Mar. 8, 2024).

²⁹ See generally Reshmi B.R., *The Future Trends of AI Approaches in Text Mining: Challenges and Innovations*, 10 INT’L J. SCI., ARCHITECTURE, TECH., & ENV’T 32 (Oct. 2025), https://ijsate.com/wp-content/uploads/2025/10/V2I10P04_IJSATE1025004.pdf (providing an overview of the current state and emerging trends of AI-driven TDM).

³⁰ 2024 RECOMMENDATION, *supra* note 6, at 75 (clarifying that the exemption only allows researchers to circumvent TPMs for conducting noninfringing TDM research and teaching and that “[i]t would be outside the

exemption's language would not prevent researchers from utilizing the exemption for noninfringing TDM research and teaching activities because proponents confirmed that TDM researchers did not intend to use the exemption to train LLMs and admitted to the importance of differentiating between using generative AI models for analytical work and generating new works.³¹

Conclusion

As demonstrated above, the prior rulemaking record on which the existing exemption is based can no longer be relied upon for streamlined renewal. New market options and solutions catering to TDM research and teaching have emerged and continue to proliferate, thereby diminishing the potential adverse effects the prohibition on circumvention inflicts on researchers' ability to conduct noninfringing TDM research. Moreover, new evidence of the existing exemption's unintended facilitation of research activities based on illicitly acquired works and recent developments surrounding unlawful data acquisition practices in connection with generative AI training also reveal the need to review and update the exemption to more effectively prevent misuse that would harm the market for or value of copyrighted works.

Therefore, to ensure that the balance between rightsholders' exclusive rights and the public's ability to engage in noninfringing uses is properly preserved, AAP urges the Office to decline Petitioners' streamlined renewal and reevaluate the exemption in the full rulemaking proceeding.

scope of these exemptions to circumvent TPMs in order to compile training data for use in the development of generative AI models and systems or to use corpora previously compiled for text and data mining research for that purpose").

³¹ See Transcript of Section 1201 Public Hearing: Proposed Class 3: Motion Pictures & Literary Works – Text and Data Mining, at 76 (Apr. 17, 2024) (John P. Bell, Dartmouth College and University of Maine, on behalf of Authors Alliance), <https://www.copyright.gov/1201/2024/hearing-transcripts/240417-Section-1201-Public-Hearing-Class-3.pdf> ("I'll just add that I don't think that we are interested in the digital humanities and recreating Sora, right? Nobody wants to make a completely generative system out of this."); *id.* at 82 (David Bamman, University of California, Berkeley, on behalf of Authors Alliance) ("[I]t is important to differentiate between the uses of generative AI."); *id.* at 85 (Bamman) ("So I think that I would not imagine copyrighted works being used to train a large language model like we imagine ChatGPT being.").