Proposed Class 7(b): Literary Works—Text and Data Mining
Submitted by: Association of American Publishers

[ ] Check here if multimedia evidence is being provided in connection with this comment

ITEM A. COMMENTER INFORMATION

This comment is submitted on behalf of the Association of American Publishers ("AAP"). AAP represents the leading book, journal and educational publishers in the United States on matters of law and policy, advocating for outcomes that incentivize creative expression, professional content and innovative educational materials. AAP's members depend first and foremost on a rational and effective copyright system.

ITEM B. PROPOSED CLASS ADDRESSED

Proposed Class 7(b): Literary Works—Text and Data Mining

ITEM C. OVERVIEW

Petitioners seek a sweeping and unprecedented exemption for "researchers" in the "humanities, social sciences and sciences" to "circumvent technological protection measures on lawfully accessed literary works distributed electronically ... in order to deploy text and data mining techniques." Petition for New Exemption Under 17 U.S.C. § 1201, at 2 ("Pet."); Notice of Proposed Rulemaking, 85 Fed. Reg. 65293, 65305 (Oct. 15, 2020). The proposed exemption would encompass copying of the circumvented works so they could be assembled into collections for purposes of text and data mining ("TDM"). Petitioners' Long Comment at 4 ("Petrs. Comment"). Though Petitioners' interest in conducting academic research is a worthy goal, the exemption they seek is extraordinarily broad and unnecessary to achieve their stated objective. More than that, it is unsupported by law, and would pose an unprecedented threat to copyrighted literary works—including books, journals, databases and computer programs—that are distributed in electronic formats.

ITEM D. TECHNOLOGICAL PROTECTION MEASURE(S) AND METHOD(S) OF CIRCUMVENTION

Like other investors in and proprietors of copyrighted works, AAP's members employ technological protection measures ("TPMs"), or "access controls," to prevent unauthorized access to and infringement of their works. TPMs are beneficial for both content owners and consumers. TPMs allow rightsholders to control access to their works, and also encourage copyright owners to make their works electronically available. TPMs also empower consumers, as they enable the development of content delivery systems that allow users to access desired content at a time and place, and on the platform, of their choosing.

Unfortunately, TPMs can be circumvented to remove their protections from ebooks and other literary works. In addition, rogue actors may employ nefarious means to bypass password protections. Once hacked, an unprotected work can be shared freely on pirate sites without authorization from, or remuneration to, the rightsholder. Sites trafficking in stolen ebooks and journal articles include online distribution hubs, cyber lockers and auction sites. *See* United States Trade Representative, *2020 Review of Notorious Markets for Counterfeiting and Piracy, available at* https://ustr.gov/sites/default/files/files/Press/Releases/2020%20Review%20of%20Notorious%20Markets%20for%20Counterfeiting%20and%20Piracy%20(final).pdf (last visited Feb. 2, 2021). The notorious website Sci-Hub, for example, enables users to illegally download PDF versions of scholarly articles—including articles that require subscriptions to access on the journals' authorized sites. Sci-Hub has grown rapidly since its creation in 2011; as of March 2017, its database reportedly contained 68.9% of the 81.6 million scholarly articles registered with the digital registration agency Crossref, and 85.1% of the articles published by subscription journals. Himmelstein et al., *Sci-Hub Provides Access to Nearly All Scholarly Literature,* eLIFE (Feb. 9, 2018), *available at* https://elifesciences.org/articles/32822 (last visited Feb. 2, 2021). In another example, pirate site Library Genesis ("Libgen") claims to host copies of more than 2.4 million nonfiction books, 80 million science magazine articles, 2.2 million fiction books, 0.4 million magazine issues, and 2 million comics strips, a vast number of which are infringing. *See* https://libgen.onl/ (last visited Feb. 2, 2021).

ITEM E. ASSERTED ADVERSE EFFECTS ON NONINFRINGING USES

Petitioners' proposal fails to satisfy the requirements for an exemption.

1. The Proposed Class Is Not Narrowly Tailored

Petitioners seek the ability to circumvent any electronically distributed literary work—including all such works from the 20th and 21st centuries, a vast portion of which are subject to copyright protection—in order to copy and assemble them into collections to conduct TDM activities. Petrs. Comment at 4, 21. "TDM" is broadly described by petitioners as an "umbrella term … used internationally to refer to the use of copyrighted work[s] in computational research." *Id.* at 4 n.1. "Computational research" is not itself defined. Petitioners do not propose any limit to the vast amount of material that would be covered by the exemption; any literary work could be circumvented and reproduced so long as it was distributed electronically. *Id.* at 9-10. Further, in describing the subject works as "lawfully accessed literary works distributed electronically," Proposed Class 7(b) does not require that the user of the exemption (rather than someone else) have "lawful access" to a work, or even that the electronic distribution of the work itself be lawful.

Although the examples provided in Petitioners' Comment and attached letters of support focus on books, by defining the proposed class as "literary works," the proposed exemption would extend to dramatic works, periodicals, databases, websites and computer programs, all of which are considered "literary works" under the Copyright Act. *See* 17 U.S.C. §§ 101, 102(a); U.S. Copyright Office, *Compendium of U.S. Copyright Office Practices* § 721.1 (3d ed. 2021). In other words, any work expressed in words, numbers or symbols would be covered. *See* 17 U.S.C. § 101.

As the Copyright Office and Congress have made clear, a "particular class of copyrighted works" designated for an exemption under section 1201 must be 'a *narrow and focused subset*' of the broad categories of works … identified in section 102 of the Copyright Act," such as literary works. U.S. Copyright Office, *Section 1201 Rulemaking: Seventh Triennial Proceeding to Determine Exemptions to the Prohibition on Circumvention* 13 (2018) (quoting H.R. Rep. No. 105-551, pt. 2, at 38 (1998) ("Commerce Comm. Report") (emphasis by Copyright Office) ("*2018 Rulemaking*"). As the Register of Copyrights has elaborated:

> For example, while the category of "literary works" under section 102(a)(1) "embraces both prose creations such as journals, periodicals or books, and computer programs of all kinds," Congress explained that "[i]t is exceedingly unlikely that the impact of the prohibition on circumvention of access control technologies will be the same for scientific journals as it is for computer operating systems." Thus, "these two categories of works, while both 'literary works,' do not constitute a single 'particular class' for purposes of" section 1201(a)(1).

Id. at 13-14 (quoting Staff of H. Comm. on the Judiciary, 105th Cong., *Section-by-Section Analysis of H.R. 2281 as Passed by the United States House of Representatives on August 4, 1998,* at 7 (1998) ("House Manager's Report"). The Register's observation and legislative history could not be more on point: Proposed Class 7(b) is untenably broad.

2. The Proposed Class Poses Enormous Security Risks

Petitioners have not suggested any qualifying criteria for "researchers," other than that they be in the fields of the "humanities, social sciences, or sciences" and seek to conduct TDM research. Pet. at 2. In other words, essentially any individual or entity espousing an interest in TDM activities could qualify. There is no limit on the size of the corpus of works that any such person or entity could copy and assemble, let alone any requirement as to how or where it would be maintained, who would have access to it, or whether it could be further reproduced, distributed, etc. Anyone with a desire to engage in TDM efforts in connection with, for example, books of fact or fiction, scientific databases or gaming software, would be permitted to circumvent access controls on such works and reproduce them to create a full-text collection of the works, free of any security protections or user restrictions. Stripped of access controls, the collected works would be exposed to unauthorized downloading and distribution over the internet. Proposed Class 7(b) thus presents a specter of unlawful dissemination of copyrighted books, databases, and software programs on a massive scale.

Such a scenario is exactly the opposite of the purpose of section 1201, which is to create a secure environment for the online distribution of copyrighted works. As the Copyright Office has explained, "[i]n enacting section 1201, Congress recognized that the same features making digital technology a valuable delivery mechanism—the ability to quickly create and distribute near-perfect copies of works on a vast scale—also carry the potential to enable piracy to a degree unimaginable in the analog context." *2018 Rulemaking* at 9. Congress adopted section 1201 so copyright owners would be able to rely on TPMs when distributing their works in digital form. *Id.* The adoption of Proposed Class 7(b) would undermine that intent.

3. Petitioners Have Failed to Establish that the Proposed Uses Are Noninfringing Uses

a. *Google Books* and *HathiTrust* do not extend to the activities contemplated by Proposed Class 7(b)

To qualify for an exemption, the proposed acts of circumvention must enable noninfringing uses. 17 U.S.C. § 1201(a)(1)(B). "[T]here is no 'rule of doubt' favoring an exemption when it is unclear that a particular use is a fair or otherwise noninfringing use." *2018 Rulemaking* at 15 (citing U.S. Copyright Office, *Section 1201 Rulemaking: Sixth Triennial Proceeding to Determine Exemptions to the Prohibition on Circumvention* 15 (2015) and U.S. Copyright Office, *Section 1201 of Title 17* 115-16 (2017) ("*Section 1201 Report*")). In other words, "'the rulemaking is not an appropriate venue for breaking new ground in fair use jurisprudence.'" *Id.* (quoting *Section 1201 Report* at 116-17). Under this well-established standard, petitioners have failed to establish that the activities they seek to enable qualify as noninfringing uses for purposes of section 1201.

Petitioners rely heavily on two Second Circuit cases, *Authors Guild, Inc. v. HathiTrust*, 755 F.3d 87 (2d Cir. 2014) ("*HathiTrust*"), and *Authors Guild, Inc. v. Google, Inc.,* 804 F.3d 202 (2d Cir. 2015) ("*Google Books*"), in support of their broad proposal to allow TDM researchers to circumvent access controls on any work that is made available in an electronic format. *HathiTrust* and *Google Books* held that certain TDM activities relying on digitally scanned books, in controlled and allegedly secure environments, constituted fair uses of copyrighted works. By their very terms, however, as further discussed below, the rulings in *HathiTrust* and *Google Books* do not extend to the activities contemplated by Proposed Class 7(b).

Moreover, in a later Second Circuit case involving an unauthorized, searchable database of audiovisual works created by defendant TVEyes, *Fox News Network, LLC v. TVEyes, Inc.*, 883 F.3d 169 (2d Cir. 2018) ("*TVEyes*"), the court—distinguishing *Google Books* and *HathiTrust*—held that various functionalities of TVEyes's database, including the ability to view video clips, were infringing rather than fair uses. *Id.* at 174, 176-81. Observing that the Second Circuit had "cautioned in [*Google Books*] that th[at] case 'test[ed] the boundaries of fair use,'" the *TVEyes* court concluded that defendant TVEyes had "exceeded those bounds." *Id.* at 174 (quoting *Google Books,* 804 F.3d at 206). Concerning the searchable TVEyes database itself, although the issue was not before it, the court made a point of stating that it "expressed no views on the [search function of the database], neither upholding nor rejecting it." *Id.* at 182 n.7.

In sum, U.S. law on TDM uses of copyrighted works—and whether such uses qualify as fair uses—is far from settled. Only a single circuit court has addressed this area in a meaningful way; its decisions have been limited and fact-specific, and have reached differing conclusions on the question of fair use. Contrary to petitioners' claim, there is no general declaration under *Google Books* or *HathiTrust* that the reproduction and/or other uses of copyrighted works for TDM purposes is a noninfringing use. *See* Petrs. Comment at 21-22. In fact, the *Google Books* decision was careful to cabin its holding to the circumstances of that particular case, which, as noted, the court considered to "test[] the boundaries of fair use." *See Google Books*, 804 F.3d at 206-07, 222, 224-25, 229 (qualifying its holding with terminology such as "at least under present conditions," "in these circumstances," "at this time," "at least as presently structured," "as … presently constructed," "at least as … presently designed," "[o]n the present record," etc.). Similarly, the finding of fair use in *HathiTrust* was expressly limited to the specific facts before the court, which noted that its fair use determination was made "[w]ithout foreclosing a future claim based on circumstances not now predictable, and based on a different record." *HathiTrust*, 755 F.3d at 101; *see also* Matthew Sag, *The New Legal Landscape for Text Mining and Machine Learning*, 66 J. Copyright Soc'y of the U.S.A. 291, 294 (2019), *available at* https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3331606 (last visited Feb. 2, 2021) ("Sag") (observing that *Google Books* and *HathiTrust* "were a product of the particular factual circumstances and can only be extended cautiously to other contexts").

Apart from the far different scenario posed by Proposed Class 7(b), it is worth noting that the world has changed since Google began scanning tens of millions of books almost two decades ago—including in the six years since the *Google Books* decision issued. There is considerably more recognition today of the value and marketability of large datasets that lend themselves to TDM activities. A collection of human-authored works can be extremely valuable to users who seek to develop artificial intelligence ("AI") and machine learning capabilities and applications, including academic researchers seeking to partner with commercial entities and/or publicize or seek remuneration for their efforts in this area. "[M]any AI practices involve the ingestion of copyrighted content, including content from journals, newspapers, books and databases …." Copyright Clearance Center, Inc. ("CCC"), *Written Comments of Copyright Clearance Center, Inc.*, Intellectual Property Protection for Artificial Intelligence Innovation, Docket No. PTO-C-2019-0038, U.S. Patent and Trademark Office, Department of Commerce, at 3 (Jan. 10, 2020), *available at* https://www.uspto.gov/sites/default/files/documents/Copyright-Clearance-Center_RFC-84-FR-58141.pdf ("*CCC USPTO Comments*") (last visited Feb. 9, 2021). "In fact, quality data inputs, including inputs of copyrighted content, are now one of the most valuable tools for businesses and other organizations to operate successfully and efficiently." *Id.*

In any event, both *Google Books* and *HathiTrust* are readily distinguishable from the uses petitioners propose to make here. In neither case were full-text copyrighted works made available to researchers. Except in connection with the use of assistive technologies by print-disabled persons, the HathiTrust library (which was generated by Google's book-scanning project) did not display any text from the copyrighted works in its collection. *HathiTrust*, 755 F.3d at 91; *Google Books*, 804 F.3d at 217. As for Google, it limited the display of text in response to user searches to "snippet views" of about three lines of text; this functional limitation on users' ability to see the book was critical to the court's analysis. *See Google Books*, 804 F.3d at 210, 222-23, 226. The *Google Books* opinion flatly distinguished the situation (such as under the exemption proposed here) where full-text works were made available to researchers: "If Plaintiffs' claim were based on Google's converting their books into a digitized form and making that digitized version accessible to the public," the court observed, that claim "would be strong." *Id.* at 225.

Another crucial distinction between the *Google Books* and *HathiTrust* cases and what petitioners propose here were the purported efforts of Google and HathiTrust to secure the copyrighted works in their collections from unauthorized access. In contesting Google's mass scanning project, the *Google Books* plaintiffs argued that the defendants' digitization and storage of their books exposed the books to piracy, undermining the value of their copyrights, and negating fair use. *Id.* at 227. Far from dismissing this concern, the Second Circuit panel confirmed that the claim "ha[d] a reasonable theoretical basis." *Id.* Accordingly, the court proceeded to review the record evidence of the alleged steps taken by Google to keep the digital corpus secure, which included walling the works off from internet access and applying the same "impressive" security measures used by Google to keep its own confidential information safe. *Id.* at 228. After reviewing this evidence, the court determined that Google had "carri[ed] its burden on this aspect of its claim of fair use." *Id.*

The *HathiTrust* decision, likewise, addressed the concern that the corpus of works maintained by HathiTrust could pose an existential threat to the plaintiffs' copyrighted works if the repository were hacked. Again, the court reviewed the security measures undertaken by defendants "to safeguard against the risk of a data breach." *HathiTrust*, 755 F.3d at 100. These included "rigorous" physical security controls, "highly restricted access" to the corpus by library staff, "highly restricted" web access and protocols to prevent downloading of non-public domain works, and a "mass download prevention system" to shut off user access in case of excessive export activity. *Id.* (citing Joint Appendix). On this record, the court upheld defendants' claim of fair use, but was careful to note that it did not foreclose a future claim based on security concerns. *Id.*

In contrast to the security protocols cited and relied upon by the Second Circuit in *Google Books* and *HathiTrust*, petitioners have not proposed any protective measures to safeguard the potentially vast corpora of circumvented copyrighted works that would be generated under their proposal. Even if they had, it does not seem plausible that any such measures could realistically be implemented under such a wide-ranging exemption, let alone monitored or enforced. Petitioners' attempt to brush the security concerns aside by arguing that the circumvented works *could* be secured—rather than *would* be—are unconvincing. *See* Petrs. Comment at 27-28. Notably, in a 2019 article, a leading proponent and practitioner of fair-use based TDM activities, Matthew Sag—who submitted one of the letters of support for Proposed Class 7(b)—had this to say about security under a model where researchers have direct access to full-text materials, as would be the case here:

> The obvious drawback of the direct access model is that the researcher becomes a single point of failure for copyright and security risks…. The security risk in this scenario is that a researcher will improperly reproduce the database of underlying works or allow an unauthorized third party to do so…. [I]t would seem reckless to give [a graduate student] unsupervised access to the entire HathiTrust corpus.

Sag, 66 J. Copyright Soc'y of the U.S.A. at 359; *see also id.* at 294 (both *Google Books* and *HathiTrust* "addressed security issues that might bear upon the fair use claim"). AAP agrees with the above assessment, except that the "recklessness" concern would apply not just to graduate students, but any potential user of the proposed exemption or a corpus created thereunder.

b. The proposed uses have not been shown to be fair uses

As the above discussion demonstrates, petitioners are unable to meet their burden of establishing that the uses contemplated by Proposed Class 7(b) are fair uses of copyrighted works. The sheer breadth of their proposal in terms of the "researchers" who would be able to engage in circumvention and extraordinary scope of affected works; the availability of large bodies of full-text copyrighted works—ranging from books to periodicals to databases to computer software—to individuals and entities without restriction; and the lack of any protection or protocol to prevent unauthorized uses or dissemination of copyright-protected works—even such basic criteria as where, how, and by whom these corpora would be maintained—all militate strongly against fair use.

Section 107 of the Copyright Act prescribes the factors to be considered in evaluating a claim of fair use: (1) the purpose and character of the use, including whether it is of a commercial or nonprofit nature; (2) the nature of the copyrighted work; (3) the amount and substantiality of the portion used; and (4) the effect of the use on the potential market for or value of the copyrighted work. 17 U.S.C. § 107. Under the Supreme Court's decision in *Campbell v. Acuff-Rose Music, Inc.*, 510 U.S. 569 (1994), courts also consider under the first factor whether a use is "transformative," that is, whether it alters the original with "new expression, meaning, or message." *Id.* at 579.

Given the broad swath of activities covered by the proposed exemption, which would include commercial uses, petitioners cannot establish that the overall purpose and character of the uses they seek to facilitate weighs in favor of fair use. The examples provided by petitioners in support of Proposed Class 7(b) with respect to literary works focus on academic research related to books, but the proposal is in no way limited to scholarly or book-related projects. As explained above, it also encompasses nonacademic uses, as well as periodicals, databases and software. And the proposal does not specify the nature or boundaries of TDM activities that would be permitted under the exemption; the vague definition of TDM espoused by petitioners merely requires that the copyrighted works be used in some fashion to engage in "computational research," without any limiting principles.

Moreover, unlike in *Google Books* or *HathiTrust*, users of the exemption would acquire full-text access to the works they chose to circumvent—as apparently would other persons and entities too—without any limitations on use of the expressive content. Such a scenario is plainly incompatible with fair use.

While some of the TDM projects cited in the petition seem to involve actual computational analysis of works (*e.g.*, how often does a certain word appear in particular texts), others appear to contemplate more traditional textual analysis involving expressive aspects of the text (*e.g.,* how and in what contexts are concepts, motifs, or tropes employed and deployed). For example, a group of researchers from Stanford seeks to analyze "written" versus "spoken" language as it appears in "portions" of text in the *Baby-Sitters Club* series of books; consider "how these books treat religion, race, adoption, divorce, and disability"; and catalog material from the series that has been included or adapted in, or excluded from, "new media formats." Petrs. Comment Appendix D (Letter from the Data-Sitters Club), at 1-2. Although it may include computational elements, such a project does not sound merely computational in nature. Even if useful passages are located by a computer, the examination and analysis of expressive passages of text is not a nonconsumptive use of the text. *Cf. HathiTrust*, 755 F.3d at 97 (noting "little resemblance" between underlying books and HathiTrust search results). Nor is it a transformative use of the text, for the text is being used for its expressive qualities. *See* Sag, 66 J. Copyright Soc'y of the U.S.A. at 45 ("TDM metadata may simply be the first stage in a process of knowledge discovery that involves reading a curated selection of the underlying works…. [B]ut [such curated reading] is unlikely to be supported by fair use."); *see also* Kyle K. Courtney, Rachael Samberg & Timothy Vollmer, *Big Data Gets Big Help: Law and policy literacies for text data mining*, 81 Ass'n of College & Research Libraries 4 (2020), *available at* https://crln.acrl.org/index.php/crlnews/article/view/24383/32222 (last visited Feb. 8, 2021) ("Courtney *et al.*") (HahtiTrust model did not include "read[ing] th[e] content").

The concern about exploiting, and profiting from, the expressive value of copyrighted works through TDM activities goes well beyond the Data-Sitters example. As explained above, petitioners' proposal would enable circumvention to conduct such activities for purposes of AI and machine learning. As copyright and technology scholar Benjamin Sobel explains:

> Emerging applications of machine learning challenge … the[] premises of non-expressive use. First, machine learning gives computers the ability to derive valuable information from the way authors express ideas. Instead of merely deriving facts about a work, they may be able to glean value from a work's expressive aspects; as a result, these uses of machine learning may no longer qualify as non-expressive in character.

Benjamin L. W. Sobel, *Artificial Intelligence's Fair Use Crisis*, 41 Colum. J. L. & Arts 45, 57 (2017). That a machine is doing the "reading" does not render the use transformative for purposes of fair use—the fact is, machines, too, may read for expressive content. Or, put another way: "Why should a digital humanities scholar devour millions of texts without compensating their authors, while a more conventional literary hermeneut—or an ordinary reader—must pay for the copyrighted works she interprets?" *Id.* at 82.

Beyond these concerns, petitioners have wholly failed to address the purpose and character of the proposed uses with respect to the literary works other than books to which the proposed exemption would apply. In what manner would researchers be engaging with periodicals, databases or software? And to what end? Petitioners offer no basis whatsoever to support their claim that these uses would be fair.

Wholly absent, as well, is any discussion of the potential commercial use of work product, another consideration under the first fair use factor. 17 U.S.C. 107(1). Certain types of TDM research—especially that directed to AI and machine learning—has significant commercial value, weighing against fair use. Though seemingly focused on (presumably nonprofit) academic research, Proposed Class 7(b) would allow profit-seeking individuals and entities to circumvent copyrighted works just as well as professors of literature.

In short, factor one weighs heavily against petitioners.

Factor two, as well, weighs against petitioners in that all types of works—including highly creative works at the "core" of copyright protection—would be included under the exemption. *See Campbell*, 510 U.S. at 586.

The third fair use factor considers the amount and substantiality of what is taken from the copyrighted work. By definition, for purposes of the exemption, that would be the entire work; and, as discussed above, the entire work would be rendered accessible to the circumventing party and presumably others as well. Such full-work appropriation and sharing is easily distinguished from the takings in *Google Books* and *HathiTrust* and negates a finding of fair use. *See* Courtney *et al.* (noting that *Google, HathiTrust* and *TVEyes* do not support "redistribution" of copyrighted works contained in TDM corpora).

Last but not least to be considered—under the fourth fair use factor—is the current and future market harm that would result from the unlimited circumvention and appropriation of copyrighted works envisioned by Proposed Class 7(b). Proponents of the exemption claim that the creation of unprotected libraries of full-text works cannot cause market harm because the works of researchers that result from such libraries will not compete with the underlying works that were circumvented. *See* Petrs. Comment at 27-28. This argument misses the mark, in more ways than one.

The first question to be answered is whether the researchers' *uses* are usurping an existing or potential market for, or otherwise diminishing the value of, those works. The answer to both aspects of this inquiry is a categorical yes. As explained above, TDM has been very loosely defined by petitioners, and that loose definition could easily encompass consumptive uses of works by circumventing parties. In addition, under the proposed exemption, once works were circumvented, reproduced and/or assembled into collections for TDM purposes, it seems there would be nothing to stop secondary users— whether or not engaged in TDM activities—from gaining access to the full-text works. In short, Proposed Class 7(b) would permit substitutional uses of expressive content in a manner and to a degree that clearly departs from the narrow, nonconsumptive uses permitted under *Google Books* and *HathiTrust*.

Even more, the exemption is broad enough to permit use of literary works for AI and machine learning purposes, including commercial enterprises. Copyright-protected input data is "commonly used to train models to generate similar output." Sobel, 41 Colum. J. L. & Arts at 65. Such output could include "new" expressive works that mimic and perhaps even infringe upon works in the corpus from which they were derived. *Id.* The mining of copyrighted works to generate new, similar works is a function of the expressive qualities of the underlying works, and may well yield competing substitutional works. Such activities thus "present[] a new threat of market substitution that alters the analysis of the fourth fair use factor." *Id.* at 57.

Further, as petitioners acknowledge, publishers of copyrighted literary works currently license such works for TDM purposes. Petrs. Comment at 27 ("To be sure, some major publishers license collections for TDM purposes.") (referencing Gale Primary Sources Platform, https://www.gale.com/primary-sources/platform) (last visited Feb. 2, 2021)). The market for large-scale collections of copyrighted works on which to conduct TDM research activities is nascent, but growing. AAP members have developed and participate in licensing programs to address user demand for corpora on which to carry out such activities. These programs are focused on ensuring the security of copyrighted works through employment of appropriate access controls, as well as on fair remuneration for use of the works, especially in the case of for-profit users. A leading example is the RightFind offering of the CCC, which makes millions of works available in a full-text format for TDM research by paying users. *See* CCC, *RightFind XML for Mining Solution* (including embedded video), *available at* https://www.copyright.com/publishers/rightfind-xml-for-mining-solution/ (last visited Feb. 8, 2021); *CCC USPTO Comments*, at 4 (CCC TDM license covers over 11 million articles from 8,000 journals). Needless to say, third parties permitted to circumvent publishers' works to create their own TDM corpora containing the same works would be competing with such licensing programs.

Lastly, as discussed above, the creation and maintenance of unrestricted, digital rights management ("DRM")-free collections of full-text copyrighted works as envisioned by the proposed exemption would present an unprecedented risk of unauthorized access to, and theft of, vast numbers of copyrighted works. As recognized by both *Google Books* and *HathiTrust*, such unauthorized uses compete with legitimate, paid uses of works and undermine the market for and value of those works. *See Google Books*, 804 at 225-28 (recognizing security risk as part of market harm analysis under fourth factor of fair use). Unlike in those cases, petitioners here have not met—nor could they meet—their burden of establishing that unlimited hacking of ebooks, periodicals, databases and software and compiling them into potentially massive, unprotected collections would not undermine the value of those works.

4. Petitioners Have Not Met the Test for Adverse Impact

a. Petitioners themselves acknowledge existing alternatives to circumvention

A petitioner must be able to point to "'distinct, verifiable and measurable impacts'" in order to demonstrate that TPMs are having an adverse effect on legitimate uses of copyrighted works. *2018 Rulemaking* at 17 (quoting Commerce Comm. Report at 37). In addition, it must be shown that the prohibition on circumvention is causing the asserted adverse effects and preventing the proponents from making noninfringing uses without circumventing access controls. *2015 Rulemaking* at 83. An exemption cannot be granted on the basis of "*de minimis* impacts." *2018 Rulemaking* at 17 (quoting Commerce Comm. Report at 37). In keeping with Congress' intent, the Register of Copyrights has stressed that "'mere inconveniences'" caused by the prohibition do not satisfy the rulemaking standard. *Id.* (quoting House Manager's Report at 6).

Before turning to petitioners' assertions concerning possible alternatives to circumvention, it is worth noting that it is not uncommon for publishers to include contractual terms in licensing and browsewrap agreements for electronically distributed works that forbid the use of works that are accessed for TDM purposes. *See* Courtney *et al*.; CCC, *RightFind XML for Mining, available at* http://www.copyright.com/business/xmlformining/. Accordingly, even if circumvention were permissible, the circumventing party could be violating agreed terms of use. "Researchers and librarians … need to understand circumstances in which contracts they have signed or to which they have assented can control—and even supersede—TDM uses …." Courtney *et al.* In such a circumstance, it is the contractual term—rather than the prohibition on circumvention—that is the source of the limitation on use. Accordingly, the statutory requirement that the section 1201 prohibition be the *cause* of the claimed adverse effect would not be met.

As petitioners acknowledge, TDM researchers already have access to significant resources to conduct their research, including large corpora of text-searchable works maintained by the Gutenberg Project, Google and HathiTrust, as well as publisher-licensed collections. *See* Petrs. Comment at 10, 11, 27. The HathiTrust library alone—the product of Google's digital book scanning project—consists of over 10 million volumes, the majority of which are protected or potentially protected by copyright. Hathitrust, *Hathitrust*, *available at* https://www.hathitrust.org/documents/HathiTrust-Overview-Handout.pdf (last accessed Feb. 2, 2021). HathiTrust's holdings include "digital versions of roughly 50% of the print holdings of every large research library in North America." *Id.* HathiTrust users are able to conduct full-text searches of the entire HathiTrust repository or of "personal collections" they create of selected works. *Id.*

Petitioners, including several researchers in their letters of support, complain that it is challenging to use the HathiTrust search functionalities because HathiTrust requires researchers to employ a "secure data capsule"—*i.e.*, a particular security protocol—and that all research be carried out on HathiTrust servers. Petrs. Comment at 11 (citing various researcher letters). The HathiTrust security protocol may require a researcher to "go in and out of the capsule" repeatedly as he or she makes adjustments to the search methodology. *Id.* This complaint proves too much, as essentially it is an objection to having to abide by the very security procedures that persuaded the *HathiTrust* court to accept the HathiTrust model as a fair use of copyrighted works. Yes, it would be easier to dispense with burdensome security protocols, and operate from unsecured databases. But that is incompatible with fair use.

Petitioners also take issue with "gaps" in the HathiTrust library and speculate that it may not be accessible to particular researchers. *Id.* at 12-13. As evidenced by petitioners' submission, however, to the extent HathiTrust or other existing library resources are inadequate, researchers can and do engage in digital scanning to generate textual materials from physical books for TDM purposes. *Id.* at 13. Indeed, virtually all of the letters in support of the literary works exemption attested to optical character recognition ("OCR") processing as a familiar means by which academic researchers create corpora of works for study. *See generally id.* Appendices. Despite this acknowledgment, the letter writers nonetheless complain that it is burdensome and time-consuming to generate scans and correct scanning errors, and thus view it as an unacceptable alternative to circumvention.

The bottom line is it seems clear that lawfully tailored digital scanning negates the need for circumvention even if it is perhaps less appealing or convenient. As noted above, "mere inconveniences" do not qualify as cognizable adverse effects. The complaints about OCR scanning fall short of persuasive evidence that the inability to circumvent ebooks or other literary works poses an insurmountable, or even significant, barrier to conducting TDM research. The fact that digital scanning was used by Google to build a corpus of tens of millions of searchable copies of books—as well as to create the Google Books search engine and the HathiTrust library—should be proof enough that digital scanning is a viable alternative to circumvention.

To be clear, however, AAP does not endorse unauthorized systematic scanning of full-text copyrighted works, which it believes to be incompatible with fair use. To the extent scanning occurs, it must be performed within the boundaries of fair use, including by being carefully calibrated to a justifiable use and conducted with appropriate security protocols. The point is that the legal framework already provides sufficient legal safeguards, and does not remotely support an argument for the widespread circumvention of access controls on copyrighted literary works. Such circumvention would discourage continued growth of the digital marketplace by undermining publishers' ability to rely on DRM protections. As shown above, the TDM uses proposed by petitioners far exceed the boundaries of fair use.

b. The exemption does not satisfy the statutory criteria

Finally, petitioners have not satisfied the five statutory criteria to be considered for an exemption, as set forth in section 1201(a)(1)(C): (i) the availability for use of copyrighted works; (ii) the availability for use of works for nonprofit archival, preservation, and educational purposes; (iii) the impact that the prohibition on the circumvention of technological measures applied to copyrighted works has on criticism, comment, news reporting, teaching, scholarship, or research; (iv) the effect of circumvention of technological measures on the market for or value of copyrighted works; and (v) such other factors as the Librarian considers appropriate. 17 U.S.C. § 1201(a)(1)(C).

With respect to the first and second factors, for the reasons discussed above, the proposed exemption would in fact discourage the dissemination and availability of copyrighted works in electronic formats by significantly increasing the probability that they would be circumvented and exposed to piracy. Copyright owners might choose to withhold electronic versions rather than take that risk. Regarding the third factor, as discussed above, researchers already have sufficient means to conduct TDM activities for purposes of scholarly and educational uses. Concerning factor four, as explained above, the widespread circumvention of literary works to build unprotected libraries of full-text works would devalue those works by undermining the legitimate market for the works, which is inconsistent with fair use principles. With respect to the fifth factor, AAP submits that the sweeping exemption proposed by petitioners is diametrically opposed to the very purpose of section 1201, which is meant to encourage the digital dissemination of copyrighted works by allowing copyright owners to rely on access controls. The exemption proposed here would be an exception that swallows the whole.

DOCUMENTARY EVIDENCE

AAP requests that the online sources and information cited and/or linked to herein be considered as documentary evidence in support of AAP's comment.

5. Conclusion

The Register should recommend denial of Proposed Class 7(b).

Dated: February 9, 2021 Association of American Publishers

By: ___________________________
Jacqueline C. Charlesworth

Charlesworth Law
15671 Royal Ridge Road
Sherman Oaks, CA 91403
jacqueline@charlesworthlaw.com

Counsel for the Association of American Publishers