Samuelson Law, Technology & Public Policy Clinic

University of California, Berkeley
School of Law
353 Law Building
Berkeley, CA 94720-7200
(510) 643-4800
estallman@clinical.law.berkeley.edu

August 9, 2021

Mr. Mark Gray
Ms. Rachel Counts
U.S. Copyright Office
Library of Congress
101 Independence Ave. SE
Washington, DC 20559-6000

via email

Re: Docket No. 2020-11
Exemptions to Prohibition Against Circumvention of Technological Measures Protecting Copyrighted Works
Class 7(a) & (b)

Dear Mr. Gray and Ms. Counts:

On August 5, 2021, on behalf of Authors Alliance, Erik Stallman, Catherine Crump, and David Bamman met with Kevin Amer and Jordana Rubel to discuss the proposed exemption for Class 7, which addresses text and data mining ("**TDM**") of literary works stored electronically and motion pictures. We provided a brief demonstration of the research project discussed at the April 7 public hearing and discussed circumstances in which access to corpus content is necessary for verifying algorithmic findings. We also discussed security measures and opponents' proposed changes to the exemption.

Text and Data Mining Demonstration

In the meeting, Dr. Bamman demonstrated text and data mining research methods related to his co-authored article *The Transformation of Gender in English-Language Fiction*.[1] He discussed this work during the April 7, 2021, hearing on Proposed Class 7, where the Copyright Office asked Dr. Bamman to explain what a typical TDM project looks like.[2] Dr. Bamman explained that depending on the size and nature of the research corpus, the corpus typically would be stored on either a server or server cluster and then accessed via a virtual terminal or similar computing interface. The corpus in the demonstration comprised 100 pre-1923 novels obtained from Project Gutenberg. Dr. Bamman executed code he had written to identify the number of male and female

[1] Ted Underwood, David Bamman, & Sabrina Lee, *The Transformation of Gender in English-Language Fiction*, Journal of Cultural Analytics, Feb. 13, 2018. DOI: 10.22148/16.019.

[2] Transcript of Section 1201 Public Hearings, Proposed Class 7, at 347–350 (Apr. 7, 2021), https://www.copyright.gov/1201/2021/hearing-transcripts/210407-Section-1201-Public-Hearing-Class-15-7a-7b.pdf.

characters in novels by male and female authors, and the percentage of male and female characters in novels by gender of the author. He also executed code to show the objects most frequently associated with male and female characters.

We then discussed circumstances in which a researcher would need to access text in the corpus to verify research findings. Again drawing from the work related to his article, Dr. Bamman gave two examples. In the first, he executed code that produced all lines of text in Nathaniel Hawthorne's *The Scarlet Letter* that include both female gendered pronouns and capitalized words, and investigated an algorithm's failure to identify any female characters in the novel. In the second, Dr. Bamman executed code that produced all lines of text that included the word "legs" to investigate why this was one of the objects most associated with male characters in the research corpus. Dr. Bamman explained that it would not suffice to perform this verification by resorting to the original works. The scale of many research projects would make verification of anomalous research findings without access to the research corpus prohibitively time-consuming. Second, for many projects, the content of the research corpus is not stored in a manner that correlates to the formats of the original sources, making verification without access to the corpus impossible.

We then discussed how an outright ban on accessing text in the corpus would have made this project impossible because the researchers could not have interrogated the conclusions reached by the code they had developed. Similarly, a fixed limit on the amount of text accessed or blacking out randomly selected passages would block efforts to verify findings or to exclude front matter, forewords, or similar extraneous material from research findings. We explained that TDM in the digital humanities is an evolving field, which should not be constrained by a crabbed definition that reduces it to statistics and page locations.

We also noted that the format of the textual output via the terminal was poorly suited to consumptive or expressive use of the underlying works. Given that researchers must already have lawfully obtained a human-readable copy of the work as a condition of eligibility for the exemption, they would have no reason to make consumptive use of the corpus copy. This condition, along with a requirement that users have an institutional affiliation, makes the access issues arising here distinct from those in *Authors Guild v. Google, Inc.* ("*Google Books*")[3] and *Authors Guild v. HathiTrust* ("*HathiTrust*"),[4] in which the court considered limitations on *public* access to the underlying works.

Security

Our discussion of security covered two areas. First, we explained how *Google Books* and *HathiTrust* are consistent with the Office's past approach to security in its § 1201 recommendations. Second, we discussed potential refinements to the proposed

[3] 804 F.3d 202 (2d Cir. 2015).

[4] 755 F.3d 87 (2d Cir. 2014).

regulatory language that address opponents' legitimate security concerns without undermining the goal of the exemption or unreasonably interfering with institutions' data security management.

Google Books and *HathiTrust*

We began by observing that the approach of existing § 1201 exemptions that require reasonable security measures keyed to particular, identified risks is consistent with *Google Books* and *HathiTrust*. In both cases, plaintiffs contended that defendants' storage of their books exposed them to the risk that hackers would make their books available online for free or at low cost, destroying their economic value.[5] The Second Circuit evaluated this argument in *Google Books* by asking whether Google's conduct would "expose Plaintiffs to an unreasonable risk of loss of copyright value through incursions of hackers."[6] Neither *Google Books* nor *HathiTrust* prescribed or endorsed an exhaustive list of security controls. Instead, the court evaluated whether the security measures in place were adequate to prevent or mitigate the identified risk.[7] In neither case had plaintiffs demonstrated any likelihood of a data breach.[8] In both cases, the court concluded that the security controls in place were adequate.[9]

We explained that the Second Circuit in both cases identified security measures that were reasonable responses to actual risks.[10] This is consistent with past Office recommendations that identify the risk to be guarded against, such as the 2018 recommendation for Class 2: Audiovisual Works–Accessibility.[11] The purpose of that exemption was to permit circumvention of technological measures on motion pictures so that disability services professionals could create accessible versions.[12] The Office noted that record testimony established that accessible versions of motion pictures were made available to students with disabilities through the same delivery mechanisms as content for students without an accommodation, specifically, through classroom display or private distribution platforms.[13] Concerned about the risk of unauthorized dissemination of the works, and cautioning that mishandling of circumvented copies would give rise to infringement liability, the Office recommended regulatory language that required that the

---

[5] *Google Books,* 804 F.3d at 207; *HathiTrust*, 755 F.3d at 100.

[6] *Google Books,* 804 F.3d at 207–08.

[7] *Google Books*, 804 F.3d at 227–28; *HathiTrust*, 755 at 99–101.

[8] *Google Books,* 804 F.3d at 228 ("Nor have Plaintiffs identified any thefts from Google Books (or from the Google Library Project)."); *HathiTrus*t, 755 F.3d at 99–100 (describing HathiTrust's evidence on security as "essentially unrebutted").

[9] *Google Books*, 804 F.3d at 228; *HathiTrust*, 755 at 100–101.

[10] *Google Books*, 804 F.3d at 227–28; *HathiTrust,* 755 F.3d at 99–101.

[11] Acting Register of Copyrights, Section 1201 Rulemaking: Seventh Triennial Proceeding to Determine Exemptions to the Prohibition on Circumvention, Recommendation of the Acting Register of Copyrights, at 110 (Oct. 5, 2018) ("**2018 Recommendation**").

[12] *Id.* at 89.

[13] *Id.* at 110.

accessible versions be "provided to students or educators and stored by the educational institution in a manner intended to reasonably prevent unauthorized further dissemination of a work."[14] The Office identified the risk that exemption users must guard against, but did not prescribe the security controls to guard against it.

We also noted that *HathiTrust* should provide reassurance to rightsholders and the Office regarding the capacity of institutions of higher education to implement high quality security regimes. HathiTrust is the product of a collaboration between academic and research libraries.[15] It is housed at the University of Michigan.[16] Although researchers' limited-access research corpora will not require the same security measures as a public digital library with ten million works, HathiTrust is an example of the ability of institutions of higher education to implement security controls that reasonably address likely risks.[17] Standard practices at these institutions involve assigning a risk level to the data at issue,[18] using as a baseline a set of security controls designated for data at that risk level,[19] and then making additional modifications as justified by the unique features of the data at issue. Common security standards promulgated by organizations such as the National Institute of Standard and Technology ("**NIST**") and the International Organization for Standardization are consistent with what institutions do: identify risks and then adopt a set of controls to guard against those risks.[20]

Finally, we observed that, as in *Google Books* and *HathiTrust*, rightsholders' concerns are speculative. They do not point to any examples where the Copyright Office's previously granted exemptions to research institutions led to unauthorized dissemination, downloading, or access of protected works. Thus, the Office should recognize that institutions of higher education have an established track record of safeguarding research data and copyrighted works.

---

[14] *Id.* at 111 & n.685 (citing *Google Books*, 804 F.3d at 229).

[15] *HathiTrust*, 755 F.3d at 90; Univ. of Mich. Libr., HathiTrust Digital Library, Who We Are, https://www.lib.umich.edu/about-us/our-divisions-and-departments/hathitrust.

[16] Univ of Mich. Library, supra note 15.

[17] HathiTrust contained ten million works at the time of the Second Circuit's decision in 2014. *HathiTrust*, 755 F.3d at 90.

[18] Berkeley Information Security Office, Data Classification Standard (issued Nov. 7, 2019), https://perma.cc/EEW6-4W88.

[19] Berkeley Information Security Office, Minimum Security Standards for Electronic Information (last updated Oct. 11, 2019), https://perma.cc/TE8N-REC9.

[20] Letter from Erik Stallman et al to Regan Smith & Jordana Rubel at 2–3, (May 21, 2021), https://www.copyright.gov/1201/2021/post-hearing/letters/Class%207%20Authors%20Alliance%20and%20Library%20Copyright%20Alliance%20Post-Hearing%20Response.pdf.

Further Refinements to Proposed Exemption Language

We then discussed that, if the Office wishes to add more specific security language to the exemption to address concerns raised by opponents, there are two good options, added in bold to the language we have already proposed:

> the researcher**, in consultation with their institution's information technology office,** uses reasonable security measures to **prevent dissemination, downloading, and unauthorized access,** and to limit access to the corpus of circumvented works only to other researchers affiliated with qualifying institutions for purposes of collaboration or the replication and verification of research findings.

First, Authors Alliance does not object to the inclusion of the requirement that researchers wishing to avail themselves of the exemption consult with their institution's information technology office. As discussed earlier, institutions of higher education are well positioned to provide this kind of advice, and it would ameliorate some of opponents' concerns.[21]

Second, Authors Alliance does not object to the inclusion of language more specifically defining the harms that exemption users must guard against when implementing security controls. This is how the Office has provided more specific security-related guidance in the past.[22] Moreover, in their responses to the post-hearing questions, the security concerns put forward by exemption opponents do not appear to be rooted in the practice of text and data mining itself but rather the possibility that creation of a collection will result in economic loss through dissemination, downloading, and unauthorized access.[23] An exemption that specifically requires users to guard against these risks would address these concerns while leaving appropriate flexibility to allow institutions to integrate these measures with their existing security practices.

In addition, we discussed the various specific security controls and standards opponents advocated for in their post-hearing letters. We explained that while we continue to believe that the Copyright Office's reasonableness approach is the right one, the intended exemption beneficiaries would still be able to avail themselves of the exemption if certain controls are imposed. Others, however, would render the exemption unusable.

---

[21] Letter from Jacqueline C. Charlesworth to Regan Smith & Jordana Rubel at 10 (May 21, 2021), https://www.copyright.gov/1201/2021/post-hearing/letters/Class%207%20AAP%20--%20Post-Hearing%20Response.pdf ("**AAP**") (contending that institutions, rather than individuals, should be responsible for circumvention); Letter from Matthew Williams to Regan Smith at 5 (May 21, 2021), https://www.copyright.gov/1201/2021/post-hearing/letters/Class%207%20Joint%20Creators%20and%20Copyright%20Owners%20Post-Hearing%20Response.pdf ("**JCCO**").

[22] 2018 Recommendation at 111.

[23] AAP at 2.

Opponents proposed several security controls to which Authors Alliance does not object. These include encryption on the server;[24] limiting access to the collection to those with a legitimate and authorized need;[25] deletion of the collection upon conclusion of the applicable research need;[26] and, mechanisms to detect and prevent downloading of stored materials[27]. These security controls are commonly available and will not prevent academic researchers from using the exemption.

Other proposed security controls would render the exemption unusable. The physical separation of the corpus server from other facility servers is not a mainstream security requirement and would be difficult to implement.[28] It would not be appropriate to leave development of security standards to copyright owners, or to form a workgroup of relevant stakeholders to draft best practices.[29] Opponents' opportunity to contribute to consideration of security concerns is through this proceeding. Any further time period would result in undue delay. The proposal that exemption users be compelled to reach out to content creators to seek approval for circumvention is objectionable for the same reason.[30]

We also discussed NIST 800-171,[31] explaining that it is unsuitable as a security standard for text and data mining collections. Compliance is excessively time-consuming and expensive. NIST 800-171 was specifically developed to give federal government contractors a set of standards to follow when securing controlled unclassified information.[32] It contains many dozen security controls divided into 14 families. The standard is not self-executing, meaning that it has different levels and controls associated with different risks and data classifications. Thus, it is not the case that the Office could impose NIST 800-171 as a standard and then assume that all institutions would apply the

---

[24] DVD CCA and AACS LA, Responses of AACS LA and DVD CCA to Post-Hearing Letter For Class 7(a) (Motion pictures—text and data mining) at 1 (May 20, 2021) https://www.copyright.gov/1201/2021/post-hearing/letters/Class%207%20AACS%20LA%20and%20DVD%20CCA%20Post-Hearing%20Response.pdf ("**AACS LA and DVD CCA**").

[25] AACS LA and DVD CCA at 2.

[26] AAP at 3; JCCO at 5.

[27] AAP at 3.

[28] AACS LA and DVD CCA at 2. We do not object, however, to a requirement that exemption users take reasonable steps to provide physical security for servers containing TDM collections.

[29] JCCO at 5; AACS LA and DVD CCA at 3.

[30] Letter from Christopher Mohr to Regan Smith (May 21, 2021), https://www.copyright.gov/1201/2021/post-hearing/letters/Class%207b%20SIIA%20TDM.pdf ("**SIIA**"); AACS LA and DVD CCA at 4.

[31] National Institute of Standards and Technology, U.S. Dep't of Commerce, Protecting Controlled Unclassified Information in Nonfederal Systems and Organizations, Special Publication 800-171 **("NIST 800-171")**.

[32] NIST 800-171 at vii ("The recommended security requirements contained in this publication are only applicable to a nonfederal system or organization when mandated by a federal agency in a contract, grant, or other agreement.").

same set of security controls. Finally, for many institutions the standard is out of reach because of its complexity and cost, particularly in departments, such as those focusing on literature and film, likely to use this exemption.

Opponents' Other Proposed Changes to the Exemption

We discussed the further limitations to the proposed exemptions that opponents put forward in their responses to the Office's post-hearing questions. We noted that several of their proposals are entirely consistent with the intent and text of the proposed exemption as revised in our Reply Comment, and Authors Alliance would agree to them:

- Authors Alliance has no issue with clarifying that scholarly research and teaching must be the "sole purpose" of the exemption.[33]
- Authors Alliance has no issue with expressly prohibiting distribution of copies but believes this is best handled via the security language discussed above.[34]
- Authors Alliance can agree to deletion of the research corpora after completion of all research and verification of research findings.[35] We noted, however, that a fixed three-year duration would in many instances lead to destruction of a corpus before research and verification are complete.[36] We also noted that the limitation for text and data mining in the European Union's Copyright Directive contains no similar time-based requirement to destroy the research corpus.[37]
- Authors Alliance has no objection to prohibiting substitutional, for-profit, or commercial uses of expressive content in the research corpus.[38] However, we noted that the Office in past § 1201 proceedings has observed that the meaning of "commercial" is not always clear.
- Authors Alliance can agree to limiting eligible beneficiaries of the exemption at accredited institutions of higher education to faculty, students working under the supervision of faculty, and staff.[39] However, we noted that those terms are a poor fit for research activities at libraries, archives, and museums not affiliated with an institution of higher education.

---

[33] SIIA at 6; AACS LA and DVD CCA at 8; JCCO at 4.

[34] JCCO at 5.

[35] *Id.*

[36] *See* SIIA at 5-6 (proposing that research corpora be destroyed after three years).

[37] Directive 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market, Art. 3(2), http://data.europa.eu/eli/dir/2019/790/oj.

[38] JCCO at 5, AAP at 9.

[39] AAP at 9.

We then discussed opponents' proposed changes to which Authors Alliance and the researchers seeking this exemption object:

- Authors Alliance objects to limiting the works eligible for circumvention to "literary works of fiction."[40] We noted that humanities scholars often study nonfiction works.
- Authors Alliance objects to excluding from the exemption any institution that enjoys immunity from suit under the Eleventh Amendment.[41] Many existing § 1201 exemptions apply to such institutions, which has not led to misuse of those exemptions.
- Authors Alliance objects to the exclusion of all works obtained pursuant to license to the extent that condition would exclude purchased digital downloads.[42] This could make ebooks ineligible, undermining the exemption's core purpose. We reemphasized that the proposed exemption focuses on works that the researchers or their institutions have purchased and that there is no intent to include, for example, subscription databases of scientific periodicals. We stated that we understood that a granted exemption may exclude all subscription services but noted the problems that exclusion could create for studying contemporary culture as subscription-based distribution models become more exclusive and ubiquitous.

We thank the Copyright Office for its time and its willingness to consider refinements to the proposed exemption that address opponents' concerns without undermining the exemption. We are happy to answer any additional questions the Office may have.

Sincerely,

/s/Erik Stallman
/s/Catherine Crump
Counsel to Authors Alliance

[40] *Id.* at 6.
[41] SIIA at 6.
[42] *Id.* at 4.