Samuelson Law, Technology
& Public Policy Clinic

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

Regan Smith
General Counsel
Jordana Rubel
Assistant General Counsel
U.S. Copyright Office
Library of Congress
101 Independence Ave. SE
Washington, DC 20559-6000

May 21, 2021

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

Dear Ms. Smith and Ms. Rubel:

On behalf of Authors Alliance, thank you for the opportunity to respond to the questions outlined in your April 16, 2021 letter. While different information security standards take different approaches to assigning requirements, objectives, or security control baselines, they share in common a sensitivity to the details of the data, network, and institution those measures seek to protect. That sensitivity is consistent with the Copyright Office's own approach to security measures. We outline these approaches and then identify examples of security controls that would be reasonable for text and data mining ("**TDM**") research corpora. The European Union's approach to securing research corpora is still a work in progress, and there are good reasons not to follow it here. Finally, we clarify that while researchers do not need this exemption for the purpose of viewing the full text or images of the works that they or their institutions have already obtained lawfully, researchers must be able to verify their research methods and research results, which requires some ability to view corpus text or images. That ability is consistent with the research environments of both HathiTrust Data Capsules and Google Book Search. Further, this viewability is consistent with fair use precedent. A blanket prohibition on viewing text or images would comprehensively undermine TDM research relying on the exemption and would provide little added value or protection given the other restrictions in the proposed exemption.

1. Security Measures

The flexible process that information security and data management professionals at research institutions use to select and apply security controls to research data tracks the processes laid out in international standards and federal agency procedures. Accepted risk assessment frameworks are superior to a globally applied fixed list of minimum security requirements and are consistent with the Office's approach to information security in previous exemptions. They are also superior

to the process currently under construction in the European Union. While there are security controls common among all of these models, there is generally more than one control available to accomplish a security objective. Thus, the Office should provide an open list of reasonable security measures rather than imposing a fixed list of required controls. We provide some examples of reasonable security measures below.

a. Security Standards and Controls

The security standards cited at the April 7 hearing vary in whether they prescribe minimum requirements, baselines, or security objectives. What they share, however, is a common understanding that the appropriate combination of security controls for an information system or dataset must take into account details of the system, data, and institution in question. We briefly review these standards, how they align with discussions of security measures in previous exemptions, and how they might apply here.

The International Organization for Standardization ("**ISO**") and the National Institute for Standards and Technology ("**NIST**") have developed similar procedures for performing risk assessments and using those assessments to identify security controls. The ISO/IEC 27001 standard guides organizations through a process of developing a risk treatment plan to identify and select appropriate options and security objectives to address the particular risks identified.[1] This standard works in conjunction with ISO/IEC 27002, which categorizes security controls according to the specific objectives they accomplish.[2] Importantly, these standards do not prescribe a fixed list of minimum required security controls, as the security policy must be sensitive to the interfaces and dependencies of the information system in question.[3]

NIST's *Standards for Security and Privacy Controls for Information Systems and Organizations*[4] and related publications, which apply to many federal information systems, take a similar approach. The NIST standards provide a framework for categorizing information and information systems by identifying relevant security objectives (confidentiality, integrity, and availability) and the potential impact of unauthorized access and use.[5] The results of this security categorization "help guide and inform the selection of security control baselines to protect systems and information."[6] The baseline is then "a starting point for the subsequent tailoring activities that are

[1] ISO/IEC 27001, Information Technology—Security Techniques—Information Security Management Systems—Requirements (2d ed. Oct. 1, 2013), §§ 6.1.2, 6.1.3.

[2] ISO/IEC 27002, Information Technology—Security Techniques—Code of Practice for Information Security Controls (2d ed. Oct. 1, 2013).

[3] ISO/IEC 27001, § 4.3(c).

[4] National Institute of Standards and Technology, U.S. Dep't of Commerce, Standards for Security Categorization of Federal Information and Information Systems, Federal Information Processing Standards Publication 199 (Feb. 2004), https://perma.cc/H3YC-GQCS ("**NIST Standards**").

[5] NIST Standards, at 6.

[6] National Institute of Standards and Technology, U.S. Dep't of Commerce, Control Baselines for Information Systems and Organizations, NIST Special Publication 800-53B 6 (Oct. 2020), https://doi.org/10.6028/NIST.SP.800-53B.

applied to the baseline to produce a targeted or customized security and privacy solution[.]"[7] Like ISO/IEC 27002, a separate NIST publication groups the security controls into twenty families, which address issues such as access control, physical and environmental protection, and system and information integrity.[8] As with the interlocking ISO standards, the result of the NIST process is a set of customized and context-sensitive security controls appropriate to the specific information, information system, and institution in question.

An example of institution-specific security controls based on a risk assessment process much like the ones laid out in the above standards is the Minimum Security Standards for Electronic Information ("**MSSEI**") used by the University of California, Berkeley.[9] These standards first require the individual responsible for information technology resources supporting researchers (the "IT Resource Proprietor") to assign to the research data the appropriate protection level based on potential "business impact."[10] Based on the protection level appropriate for the data in question, the MSSEI then provides a list of required and recommended security controls. The highest protection level is reserved for personally identifiable information, financial and medical information, passwords and other "authentication secrets," and information relating to "control systems that affect life and safety."[11] A TDM research corpus would be unlikely to fall in this category. But even for lower protection levels, the MSSEI assigns a comprehensive set of relevant security controls such as controlled access, device physical security, and encryption.

In their sensitivity to the data, information system, and institution involved, the Berkeley, ISO, and NIST standards align with the Copyright Office's approach to information security in previous exemptions. In the past, the Copyright Office has required users or their institutions to implement reasonable security precautions without being overly prescriptive about the nature of those precautions. As the Office concluded when considering security measures relevant to potential § 108 reforms, "attempting to prescribe detailed digital security requirements tailored to each kind of use would result in an unduly burdensome requirement. Whether an institution's particular digital security measure is 'reasonable' will largely depend upon what measures other institutions of similar size and mission have adopted."[12] The Office took this approach in the 2015 and 2018 § 1201 rulemakings as well. As in those prior rulemakings, it is more consistent with sound

[7] NIST Special Publication 800-53B at 5.

[8] National Institute of Standards and Technology, U.S. Dep't of Commerce, Security and Privacy Controls for Information Systems and Organizations, Special Publication 800-53 Rev. 5 8 (Sept. 2020), https://perma.cc/8LFZ-8K2X.

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

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

[11] *Id.*

[12] U.S. Copyright Office, Section 108 of Title 17: A Discussion Document of the Register of Copyrights, at 21 (Sept. 2017).

information security practices for the Office to require use of security controls that "reasonably prevent unauthorized further dissemination of a work."[13]

There are good reasons to allow the information security offices of eligible institutions to select specific controls rather than prescribe them. First, the appropriate controls will depend on the existing information security system and nature of the research corpus. Securing a massive trove of literary works made partially available online to third parties—an apt description of Google Book Search or the HathiTrust Digital Library—is a fundamentally different exercise, and requires more security precautions, than a single researcher's project requiring a collection of 25 post-1994 works with characters who have autism.[14] Second, in a landscape of constantly evolving technology and security challenges, any prescription would fall out of date once an information system reached the end of its "lifecycle."[15] Third, prescribing a specific security control is inadvisable because there is often more than one way to accomplish a security objective. For example, an organization can decide to prohibit employees from accessing information systems via privately owned mobile devices or it can adopt security measures to protect those systems when accessed via privately owned devices.[16] This is in part why standards like ISO/IEC 27002 categorize security controls by the control objective and then identify multiple controls that can accomplish that objective.[17] The optimal set of controls depends on factors such as the future utility of the data, the specifics of the network and storage, and the risks associated with disclosure.

In his hearing testimony, Christopher Hoffman, the Associate Director of Research Information Technology for the University of California, Berkeley, explained that he regularly works with researchers and scholarly communications offices to develop security plans and protocols for their research data. In developing those plans, researchers and data management professionals draw from minimum security requirements and security controls typical for securing sensitive research data. Based on input from both information technology professionals and researchers, common and effective security controls used in many research settings include user authentication, use of encryption, event logging, and maintaining physical security of the resources housing the data.[18] The Office should identify these controls as examples of reasonable security measures, while leaving room for information security departments and researchers to fine-tune the precise security controls used to the specifics of the research corpus and the information system in which it is housed.

---

[13] Exemption to Prohibition on Circumvention of Copyright Protection Systems for Access Control Technologies, 83 Fed. Reg. 54,010, 54,017–18 (Oct. 26, 2018) (criticism and commentary in online courses); *id.* at 54,019 (accessible films in education).

[14] Letter from Jes Lopez, Appendix H to Authors Alliance et al., Class 7(a) and 7(b) Initial Comment, at 2.

[15] See ISO/IEC 27002, § 0.5.

[16] See ISO/IEC 27002, § 6.2.1.

[17] ISO/IEC 27002, § 4.2 (Control Categories).

[18] *See* ISO/IEC 27002, § 9.2 (user access management), § 10.1 (cryptographic controls), § 11.1 (secure areas), § 12.4.1 (event logging). Note that any event logging should be implemented in a manner consistent with academic institutions' and libraries' respect for academic freedom and patron privacy.

b. Comparison to the European Union's Directive on Copyright in the Digital Single Market.

Allowing research institutions to determine the set of applicable security controls is also superior to the approach mandated in Article 3 of the European Union's ("**EU**") Directive on Copyright in the Digital Single Market ("**CDSM**").[19] At the outset, we note that EU member states are still in the early stages of transposing Article 3 into national law,[20] and the text of the article and related recitals are open to interpretation and likely to be contested in that process. Thus, while there are commendable aspects of the EU's efforts to provide an exemption for TDM, it is too early to view Article 3 as a model for security measures.

Article 3(2) states that copies of works "shall be stored with an appropriate level of security and may be retained for the purpose of scientific research, including for the verification of research results."[21] This clause seems to contemplate the same flexible and customizable approach to security controls reflected in international standards and prior § 1201 exemptions. Similarly reasonable is the suggested requirement in Recital 15 that any retained copies of works "should be stored in a secure environment."[22]

Article 3(3) allows rightsholders to "apply measures to ensure the security and integrity of the networks and databases where the works or other subject matter are hosted."[23] Recital 16 makes clear that the purpose of this provision is to allow rightsholders to protect the security and integrity of *their own* networks and databases that might be strained by the "potentially high number of access requests to, and downloads of, their works or other subject matter[.]"[24] Notably, Recital 16 recognizes that there is more than one way to accomplish this security objective, as it identifies both IP address validation and user authentication as ways to protect the security and integrity of systems and databases.[25] Recital 16 also clarifies that the security measures rightsholders use "should remain proportionate to the risks involved, and should not exceed what is necessary to pursue the objective of ensuring the security and integrity of the system and should not undermine the effective application of the exception."[26] Thus, like the ISO and NIST standards, Recital 16

---

[19] 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, http://data.europa.eu/eli/dir/2019/790/oj.

[20] *See* CREATe, *Copyright in the Digital Single Market Directive—Implementation: An EU Copyright Reform Resource*, https://perma.cc/UF5C-85AW (last visited May 16, 2021) (summarizing status of member state transpositions).

[21] CDSM, Art. 3(2).

[22] CDSM, Recital 15.

[23] CDSM, Art. 3(3).

[24] CDSM, Recital 16. *See also* Christophe Geiger, Giancarlo Frosio, and Oleksandr Bulayenko, *Text and Data Mining: Articles 3 and 4 of the Directive 2019/790/EU*, *in* PROPIEDAD INTELECTUAL Y MERCADO ÚNICO DIGITAL EUROPEO 54-55 (Concepción Saiz García and Raquel Evangelio Llorca eds., 2019) (analogizing this provision to the provisions in the Telecoms Single Market Directive allowing telecom operators to manage congestion on their networks).

[25] CDSM, Recital 16.

[26] *Id.*

counsels a flexible, risk-based approach to security. Importantly, the recital also seeks to ensure that security is not used as a pretext to undermine the benefit of the exemption.

Problems emerge, however, with Article 3(4)'s requirement that member states "encourage rightholders, research organisations and cultural heritage institutions to define commonly agreed best practices concerning" the security measures described in Articles 3(2) and 3(3).[27] While good can come from rightsholders and research institutions sharing their information security practices, asking them to identify commonly agreed security controls overlooks the needs for the flexibility and customization discussed above. Further, past efforts to develop best practices related to matters such as notice-and-takedown or "standard technical measures" under 17 U.S.C. § 512 have been largely unsuccessful. Particularly with an exemption that may last only three years, it is difficult to see how best practices could be developed and implemented in a timeframe that allows researchers to benefit from the exemption. Finally, it is unclear which rightsholders or how many rightsholders would need to be involved in these efforts. Depending on the corpus, the number could be vast, and those rightsholders could be difficult—and in some cases, impossible—to identify and locate.

In sum, the approach to security measures used in prior § 1201 exemptions mirrors the approach in prevailing national and international security standards, and even some aspects of the EU Copyright Directive. That approach provides needed flexibility while appropriately protecting works from unauthorized and unlawful dissemination and use. The Office should adopt that approach here, while supplying examples of reasonable security measures drawn from relevant standards.

2. Prohibiting Researchers from Viewing Text and Images

An outright ban on viewing any content in a research corpus would comprehensively undermine TDM research projects based on the exemption. It is a basic principle of scientific research that researchers must be able to verify their research methods and findings. For TDM, this means that researchers must be able to view the output of models and algorithms, and verify that the results of those models and algorithms are accurate. This verification cannot take place if researchers cannot view the underlying data at all. Should the Office include a limitation on viewability, it should follow the model and policies of established TDM research environments like HathiTrust Data Capsules. However, even that limitation is unnecessary in view of other limitations in the proposed exemption.

We clarify here the testimony of exemption proponents at the Office's April 7 hearing: the purpose of the exemption is not to create additional, human-readable or viewable copies of works in the research corpus for expressive use. Under the terms of the exemption, the researchers must already have a lawfully obtained copy of those works in hand, but those copies are not useful for TDM purposes. Instead, the exemption is needed to create machine-readable copies of those works to be

[27] CDSM, Article 3(4).

used in TDM. However, researchers must be able to view enough of the text and images included in the corpus to verify their research methods and research results.

To use the example discussed at the hearing,[28] a researcher might use a high-performance computing ("**HPC**") cluster to examine a large number of motion pictures to develop an algorithm that identifies scenes of violence and assigns films overall violence scores. The interfaces for the HPC clusters do allow for viewing and revising code and analyzing algorithmic outputs, but those command-line interfaces are not designed for and would be unsuitable for watching motion pictures. If an algorithm tells the researcher that frame #133292 of a corpus copy has a high probability of being a scene of violence, and that frame corresponds to a scene in the film *Pulp Fiction*, the researcher would not watch a copy of the DVD or digital download in its original format to verify that finding. But at some point, either the researcher or peer reviewer may need to locate and examine frame #133292, a designation that exists only in the corpus copy, to verify the algorithmic finding. The exemption should not foreclose this verification. Similarly, if an algorithm identifies "Jarndyce" as a person in a corpus of 100 different texts, the researcher must be able to verify that the word "Jarndyce" appearing at the location in the corpus identified by the algorithm refers to a person rather than, say, a lawsuit.[29] This is no more and no different than a Google Book Search word query producing a snippet so the searcher can verify that "Einstein" refers to a physicist instead of cat.[30]

Notably, HathiTrust Research Center ("**HTRC**") Data Capsules also permit this verification. The purpose of Data Capsules is to enable non-consumptive research, but part of that research involves producing small portions of the text from the corpus. HTRC implements its non-consumptive research restriction via a Non-Consumptive Use Research Policy clarifying that Data Capsules are not to be used to read or display "substantial portions" of in-copyright works.[31] The policy also clarifies that "[e]xamples of acceptable in-capsule uses of corpus text that may facilitate non-consumptive research include referrals to specific passages in order to verify or evaluate results, to develop and revise algorithms for processing the text, and to select appropriately short quotes as necessary examples in reporting the research, as may be supported by fair use."[32] The HTRC Non-Consumptive Use Research Policy is a workable model for viewability restrictions the Office may be contemplating because it allows researcher to view text and images in a corpus as necessary to verify their research methods and findings.

---

[28] We are unable to cite to the record as the transcript and video of the hearing are unavailable from the Copyright Office at the time of filing.

[29] *See* David Bamman, *An Annotated Dataset of Literary Entities*, PROCEEDINGS OF THE NAACL-HLT 2019, 2138, 2140 (2019), https://www.aclweb.org/anthology/N19-1220.pdf.

[30] *Authors Guild v. Google, Inc.*, 804 F.3d 202, 217–18 (2d Cir. 2015) ("*Google Books*").

[31] HathiTrust Digital Library, *Non-Consumptive Use Research Policy*, (approved Feb. 20, 2017), https://perma.cc/8WRP-7532 (defining "substantial portion" to mean "a portion of an individual volume sufficient in quality or quantity to provide a substitute for access to the volume's expressive content. A portion that merely reveals factual information (about the work or about the world) is not thereby a substitute for access to the volume's original expressive content.").

[32] *Id.*

This allowance also mirrors provisions in the EU Copyright Directive accommodating verification of TDM research findings. Article 3 expressly permits retaining copies of works "for the purposes of scientific research, including for the verification of research results."[33] And Recital 15 further clarifies that research organizations and cultural institutions should be able to retain research corpora "for subsequent verification of scientific results" so long as the copies are stored in a secure environment.[34]

An outright ban on viewing text or images in a research corpus is also unnecessary because of other limitations in the proposed exemption. First, only lawfully obtained copies of works are eligible for circumvention and inclusion in a research corpus.[35] Given that researchers will already have the entire, viewable copy of the work in hand, they will have little incentive or desire to view that same content in a format intended for machine processing beyond what is necessary "to develop and revise algorithms for processing the text," or verify results.[36] The raw .txt files used for machine processing are simply inferior to works in electronic publication (".epub") or similar formats developed specifically for consumptive use with electronic readers. And because reformatting text for TDM degrades the human readability of the works, it is very unlikely that anyone would choose to read a full work in this format rather than the original format.

Second, the proposed exemption does not permit the further dissemination of copies, or even access beyond collaborators and peer reviewers.[37] Thus, there is no danger of an individual using the exemption to create copies of works for consumptive purposes.

Finally, viewing text or images in a research corpus for the purpose of validating research methods and findings still falls in the category of "mak[ing] available significant information *about those books*," which both *Authors Guild v. HathiTrust* and *Authors Guild v. Google, Inc.* found to be "the sort of transformative purpose described in *Campbell* as strongly favoring satisfaction of the first [fair use] factor."[38] Again, the goal of the research performed by exemption proponents and others in their field is "to derive information on how nomenclature, linguistic usage, and literary style have changed over time"[39] or comparable information in the area of film studies. Viewability here is solely an adjunct to the TDM analysis directed at that purpose. Accordingly, nothing about viewing text and images in the corpus for verification disturbs the status of TDM for scholarly research purposes as fair use.

In sum, researchers need to view enough of a corpus to verify their methods and their findings. Nothing about viewing corpus content for that purpose poses a risk of consumptive use or further dissemination of the circumvented works, or disturbs the conclusion that the underlying use of the

[33] CDSM, Article 3(2).
[34] CDSM, Recital 15.
[35] Authors Alliance et al. Class 7(a) & 7(b) Reply Comment at 6.
[36] HathiTrust Digital Library, *Non-Consumptive Use Research Policy*, https://perma.cc/8WRP-7532.
[37] Authors Alliance et al. Class 7(a) & 7(b) Reply Comment at 6.
[38] *Google Books*, 804 F.3d at 217.
[39] *Id.* at 209.

work is for a transformative purpose and a fair use. While proponents do not believe an express viewability limitation is warranted, should the Office choose to include one, it should use the model of the HTRC Non-Consumptive Use Policy rather than an outright ban that would comprehensively undermine the usefulness of the exemption.

* * *

We thank the Office for allowing us to address these questions and would be happy to answer any further questions you may have.

Sincerely,

Erik Stallman, Associate Director
Catherine Crump, Director
Tait Anderson, Clinical Law Student
Counsel to Authors Alliance

Jonathan Band
Counsel to the Library Copyright Alliance