



UNITED STATES COPYRIGHT OFFICE

Long Comment Regarding a Proposed Exemption Under 17 U.S.C. § 1201

ITEM A. COMMENTER INFORMATION

Commenters:

This Comment has been submitted on behalf of the Authors Alliance, the American Association of University Professors, and the Library Copyright Alliance.

(1) **Authors Alliance** is a nonprofit organization with the mission to advance the interests of authors who want to serve the public good by sharing their creations broadly. We create resources to help authors understand and enjoy their rights and promote policies that make knowledge and culture available and discoverable. For more information, visit <http://www.authorsalliance.org>

Represented by:

Samuelson Law, Technology & Public Policy Clinic
University of California, Berkeley, School of Law
Jennifer M. Urban, Clinical Professor of Law
Zhudi Huang and Mathew Cha, Clinical Law Students
jurban@clinical.law.berkeley.edu

(2) **The American Association of University Professors** (“AAUP”) is a nonprofit membership association of faculty and other academic professionals. Since our founding in 1915, the AAUP has helped shape American higher education by developing the standards and procedures that maintain quality in education and academic freedom in this country’s colleges and universities. We define fundamental professional values and standards for higher education, advance the rights of academics, particularly as those rights pertain to academic freedom and shared governance, and promote the interests of higher education teaching and research.

Represented by:

Risa Lieberwitz, AAUP General Counsel, rlieberwitz@aaup.org
Aaron Nisenson, AAUP Senior Counsel, anisenson@aaup.org
Edward Swidriski, AAUP Assistant Counsel, eswidriski@aaup.org

Privacy Act Advisory Statement: Required by the Privacy Act of 1974 (P.L. 93-579)

The authority for requesting this information is 17 U.S.C. §§ 1201(a)(1) and 705. Furnishing the requested information is voluntary. The principal use of the requested information is publication on the Copyright Office website and use by Copyright Office staff for purposes of the rulemaking proceeding conducted under 17 U.S.C. § 1201(a)(1). NOTE: No other advisory statement will be given in connection with this submission. Please keep this statement and refer to it if we communicate with you regarding this submission.

(3) **The Library Copyright Alliance** (“**LCA**”) consists of two major library associations—the American Library Association (“**ALA**”) and the Association of Research Libraries (“**ARL**”)—that collectively represent over 100,000 libraries in the United States.

Represented by:

Jonathan Band
policybandwidth
jband@policybandwidth.com

Table of Contents

Item A. Commenter Information	1
Item B. Proposed Class Addressed	5
Item C. Overview	5
1. The current exemption has begun to enable valuable digital humanities research and teaching via text and data mining.	6
2. Researchers and teachers need the proposed expansion to fully carry out the purposes of the current exemption.	8
a. The ambiguity of the term “collaboration” prevents researchers and teachers from effectively using the current exemption.	8
b. The current exemption’s limits on sharing hamper valuable research because of the difficulty and costs associated with preparing usable corpora from lawfully acquired copies, and because the inability to share corpora for new inquiries impairs research quality and sustainability.	10
3. Granting the proposed expansion will enable additional valuable research.	16
Item D. Technological Protection Measure(s) and Method(s) of Circumvention	19
Item E. Asserted Adverse Effects on Noninfringing Uses	19
1. The proposed class includes at least some works protected by copyright.	19
2. Conducting TDM research is likely to be noninfringing under Title 17.	19
3. Academic researchers and students currently are, and are likely to be in next three years, adversely affected in their ability to conduct research and study humanities using TDM techniques.	23
a. As in the 2021 Triennial Proceeding, this expansion would not curtail the availability for use of copyrighted works.	24
b. The proposed expansion would increase the amount, quality, and societal value of TDM research and education.	24
c. The inability to use corpora compiled by other higher education institutions for different projects has negative consequences on TDM teaching, scholarship, and research.	28
d. The proposed expansion is not likely to affect the licensing market for or the value of copyrighted motion pictures and literary works.	30

4. The statutory prohibition on circumventing access controls is the cause of the adverse effects.....	33
Documentary Evidence	35

ITEM B. PROPOSED CLASS ADDRESSED

The Proposed Classes include:

Proposed Class 3(a) Motion Pictures—Text and Data Mining: Lawfully accessed motion pictures where circumvention is undertaken in order to deploy text and data mining techniques.

Proposed Class 3(b) Literary Works—Text and Data Mining: Lawfully accessed literary works distributed electronically where circumvention is undertaken in order to deploy text and data mining techniques.

Proponents are submitting a single comment addressing both motion pictures and literary works because the relevant factual and legal issues as to the two classes of works—including the nature of the proposed research activities, the relevant markets for the works, and the lack of available potential alternatives to circumvention—are highly similar. Supporting evidence of adverse effects in the absence of the proposed expansion with respect to both classes is provided below.¹

ITEM C. OVERVIEW

Proponents seek an expansion to the currently existing exemptions 37 C.F.R. § 201.40(b)(4) and 37 C.F.R. § 201.40(b)(5) from the 17 U.S.C. § 1201 prohibition on circumventing technological protection measures (“**TPMs**”) to facilitate text and data mining (“**TDM**”).

Specifically, we seek limited modifications to the post-circumvention limitations currently imposed by the current exemption when researchers affiliated with one nonprofit institution of higher education share corpora with researchers affiliated with different institutions of higher education. Under the current exemption, such sharing must be “solely for the purposes of collaboration or replication of the research.”² We propose that the current exemption be amended to also allow sharing with researchers affiliated with different nonprofit institutions of higher education for purposes of conducting *independent* text and data mining research and teaching, where those researchers are also in compliance with the current exemption. To be clear, this means that independent researchers falling under this proviso would still need to comply with all other exemption requirements—they must be affiliated with an institution of higher education as defined in the exemption, that institution must itself own lawfully acquired copies of the underlying works, and the institution must comply with security standards as defined in the regulation.

In particular, we ask that 37 C.F.R. § 201.40(b)(4)(i)(D) and (b)(5)(i)(D) be modified as follows (additions italicized):

The institution uses effective security measures to prevent further dissemination or downloading of literary works in the corpus, and to limit access to only the persons identified in paragraph [(b)(4)(i)(A) or (b)(5)(i)(A)] of this section or to researchers affiliated with other institutions of higher education solely for purposes of

¹ Exemptions To Permit Circumvention of Access Controls on Copyrighted Works, 88 Fed. Reg. 72013, 72025 (proposed Oct. 19, 2023) (to be codified at 37 C.F.R. pt. 201).

² 37 C.F.R. § 201.40(b)(4)(i)(D), (b)(5)(i)(D).

collaboration or replication of the research; *or for the purposes of conducting independent text and data mining research and teaching, where those researchers are in compliance with this exemption.*

1. The current exemption has begun to enable valuable digital humanities research and teaching via text and data mining.

The current exemption³ is key to accomplishing socially important research because it allows researchers to use text and data mining to study copyrighted literary and motion pictures in order to explore important questions about our culture and society. It also allows scholars who teach to use these methods in their classes. Since the 2021 Triennial Proceeding, researchers have been able to use the current exemption to ask and answer new and exciting questions about literary and audiovisual works.

More than a dozen organizations and researchers or research teams have provided letters of support describing exciting, socially valuable TDM research enabled by the current exemption.⁴ As the Mellon Foundation, which funds many TDM research projects through its Public Knowledge program, explains:

[W]e have already seen the transformative potential of TDM research under the existing exemption, allowing researchers to expose a more nuanced understanding of history, culture, and society. Continued TDM of in-copyright content helps ensure that *contemporary* history, culture, and society are not omitted from the scholarly record. Importantly, because the exemption allows researchers to interrogate modern, culturally relevant in-copyright materials, it has allowed them to make their research more relevant and accessible for current social and civic concerns.⁵

In their letters, researchers provide many rich examples of this work, including the following:

- Mark Algee-Hewitt, Assistant Professor of Digital Humanities in the English Department at Stanford University and the Director of the Stanford Literary Lab, is working on a project analyzing the discourse of identity through 20th-century academic textual works. In

³ Throughout we refer to both 37 C.F.R. § 201.40(b)(4)(i)(D) and (b)(5)(i)(D) collectively as the “exemption” unless otherwise noted. Similarly, we refer to our proposed expansions to both 37 C.F.R. § 201.40(b)(4)(i)(D) and (b)(5)(i)(D) as the “proposed expansion” unless otherwise noted.

⁴ Appendix A: Letter from the Association for Computers and Humanities; Appendix B: Letter from Mark Algee-Hewitt; Appendix C: Letter from David Bamman; Appendix D: Letter from John Bell; Appendix E: Letter from Joel Burges and Emily Sherwood; Appendix F: Letter from Brandon Butler; Appendix G: Letter from Allison Cooper; Appendix H: Letter from Hoyt Long; Appendix I: Letter from Matthew Sag; Appendix J: Letter from Rachael Samberg and Timothy Vollmer; Appendix K: Letter from Lauren Tilton and Taylor Arnold; Appendix L: Letter from Henry Alexander Wermer-Colan; and Appendix M: Letter from the Mellon Foundation.

⁵ App. M: Letter from the Mellon Foundation, at 1 (emphasis in original).

addition, he is using the current exemption to teach graduate students how to conduct their own research.⁶

- David Bamman, Associate Professor in the School of Information at the University of California, Berkeley, is using the current exemption to analyze the representation of gender, race, and guns in approximately 2,000 films spanning from 1980 to 2022.⁷ He describes the decision to grant the current exemption as a “fundamental turning point” for his research.⁸
- John Bell, Program Director at Dartmouth College’s Data Experience and Visualizations Studio, is using the current exemption to study how acting styles have developed in the 20th century from an adaptation of stage plays to performing for a lens. The current exemption has enabled him to extract and analyze three-dimensional pose and motion data from a broad selection of United States film and television works.⁹
- Joel Burges, Director of Mediate and Associate Professor English and Visual & Cultural Studies at the University of Rochester, and Allison Cooper, Director of Kinolab at Bowdoin College, are, in a joint research project, examining the use of the close-up in narrative film and television and its racial and sexual history.¹⁰ Burges has also used the current exemption to teach hundreds of students in multiple classes to engage critically with audiovisual media, receiving two teaching awards for his work and mentorship in this area.¹¹
- Laura McGrath, Assistant Professor of English at Temple University, and Henry Alexander Wermer-Colan, Interim Academic Director and Digital Scholarship Coordinator at Temple University Libraries’ Loretta C. Duckworth Scholars Studio, are analyzing at a mass scale the content and characteristics of books banned in the 21st century. The current exemption is enabling them to build a corpus and do this important research.¹²
- Lauren Tilton, Professor of Liberal Arts and Digital Humanities, and Taylor Arnold, Associate Professor of Data Science, directors of the Distant Viewing Lab at the University of Richmond, are using the current exemption to study film and television at a mass scale.¹³

⁶ App. B: Letter from Mark Algee-Hewitt, at 1.

⁷ App. C: Letter from David Bamman, at 1.

⁸ *Id.* at 2.

⁹ App. D: Letter from John Bell, at 1.

¹⁰ App. E: Letter from Joel Burges and Emily Sherwood, at 1; App. G: Letter from Alison Cooper, at 1.

¹¹ App. E: Letter from Joel Burges and Emily Sherwood, at 1.

¹² App. L: Letter from Henry Alexander Wermer-Colan, at 1.

¹³ App. K: Letter from Lauren Tilton and Taylor Arnold, at 1.

Several of these researchers wrote in support of the current exemption in the 2021 Triennial Proceeding.¹⁴ Today, they are using the granted exemption to conduct important research.

In evaluating the petition from the 2021 Triennial Proceeding, the Register “recognize[d] the academic and societal benefits that could result from TDM research.”¹⁵ Today, these academic and societal benefits have begun to be realized—and researchers will continue to engage in research enabled by the current exemption, to society’s further benefit. As the Mellon Foundation explains, this work “helps to build an informed, heterogeneous, and civically engaged society[,]” and can help “ensure that more authentic, reflective, and nuanced stories are revealed, preserved, and told.”¹⁶

2. Researchers and teachers need the proposed expansion to fully carry out the purposes of the current exemption.

While, as described above, the current exemption has begun to enable valuable research and teaching, certain limitations prevent fully effective use of the exemption by TDM researchers studying motion pictures and literary works. First, researchers are stymied by the uncertainty surrounding what is and what is not allowed in the current exemption’s rules for corpora sharing. Second, the inability to share the corpora with researchers affiliated with other higher education institutions, except in enumerated narrow circumstances, has prevented researchers who would otherwise qualify for the current exemption from engaging in TDM research, in light of the high costs associated with compiling corpora. In turn, the limitation on corpora sharing in the current exemption has imposed roadblocks to highly valuable research and effective teaching methods.

- a. The ambiguity of the term “collaboration” prevents researchers and teachers from effectively using the current exemption.

Currently, the exemption allows researchers to share corpora with researchers at another higher education institution for “collaboration.”¹⁷ Unfortunately, which research activities would qualify as “collaboration”—and which would not—remains undefined.¹⁸ This leaves researchers unsure about the level of individual contribution to a project, goals, duration, or scale of research that is required for a “collaboration.” For example, Professor Mark Algee-Hewitt notes that the range of

¹⁴ Authors All., Am. Ass’n of Univ. Professors, & Libr. Copyright All., Round 1 Comment on Notice of Proposed Rulemaking on Exemptions To Permit Circumvention of Access Controls on Copyrighted Works at App. B (Letter from David Bamman), App. M (Letter from Lauren Tilton and Taylor Arnold), App. P (Letter from Henry Alexander Wermer-Colan) (Dec. 13, 2020), https://www.copyright.gov/1201/2021/comments/Class%2007a%20and%2007b_InitialComments_Authors%20Alliance,%20American%20Association%20of%20University%20Professors,%20and%20Library%20Copyright%20Alliance.pdf [https://perma.cc/RZK6-WRLD].

¹⁵ Shira Perlmutter, Register, *Section 1201 Rulemaking: Eighth Triennial Proceeding to Determine Exemptions to the Prohibition on Circumvention*, U.S. Copyright Off. (Oct. 19, 2021) (hereinafter 2021 Triennial Proceeding), 121.

¹⁶ App. M: Letter from the Mellon Foundation, at 1.

¹⁷ 37 C.F.R. § 201.40(b)(4)(i)(D), (b)(5)(i)(D).

¹⁸ *Id.*

activities that constitute “collaboration” in academia can vary widely, including, for example, “formal collaborations under the auspice of a grant, ad hoc collaborations that result from two teams discovering that they are working on similar material to the same ends, or even discussions at conferences between members of a loose network of scholars working on the same broad set of interests.”¹⁹ The Mellon Foundation, speaking from its broader perspective as a research funder, similarly notes, “We have seen in a number of other grant areas the tremendous value of collaborative efforts to build, share, and innovate upon corpora. Often these efforts do not begin with specific or well-defined collaborative research questions[,]” though they may result in valuable research.²⁰

The ambiguity surrounding “collaboration” has negative consequence for digital humanities researchers, who tend to be conservative in their interpretations of what is allowed. In our experience, researchers who use this exemption are extremely conscientious about following the text of the exemption and respecting its limits. This makes sense, as academic norms demand that researchers be transparent about sources of data in their research methodology. Because of this, researchers are prone to interpret ambiguity conservatively in order to avoid any interpretation that would cast a shadow on the methodology or results of their research. For example, Professor Joel Burges and Director Emily Sherwood note that the ambiguity has led to duplicated effort and extra labor in their research, as they were unsure whether students at the University of Rochester were allowed to break encryption as part of a project with Bowdoin College and Kinolab.²¹

This ambiguity stymies valuable research.²² As Professor Mark Algee-Hewitt notes:

“When we collaborate on understanding an archive of text through TDM methods, it sometimes means that we will be cooperating in applying the same methods to the same texts, and sometimes indicates that we will be taking diverging approaches to the same set of materials. . . . As the exemption is unclear what counts as a collaboration for the purpose of sharing extracted data, we have had to be exceptionally cautious about sharing in-copyright material with any collaborators at all, much to the detriment to our research, and the field as a whole.”²³

This is a particular loss in the digital humanities field, which is young and small, making academics’ ability to engage in robust collaborative and complementary activities crucial to research development.²⁴

Additional barriers arise from due to the nature of academic work. Questions can be raised when researchers—as they commonly do—move from one university to another and when technical staff

¹⁹ App. B: Letter from Mark Algee-Hewitt, at 2.

²⁰ App. M: Letter from the Mellon Foundation, at 2.

²¹ App. E: Letter from Joel Burges and Emily Sherwood, at 2.

²² App. B: Letter from Mark Algee-Hewitt, at 2.

²³ *Id.* at 2.

²⁴ *Id.* at 1.

work in parallel with academic appointees.²⁵ For example, as director of the Stanford Literary Lab, Professor Algee-Hewitt is concerned about the ability to ensure the continuation of projects when researchers in his lab move to another institution.²⁶ As Professor Algee-Hewitt emphasizes, this ambiguity has a disparate impact on younger and non-tenured researchers.²⁷ Early-career researchers, due to the structure of academic jobs like “post-doc” appointments, are unlikely to stay at a single higher education institution for the entire duration of projects.²⁸ When university administrators apply conservative legal interpretations, “these researchers can no longer continue their projects as they can no longer access the data,”²⁹ despite the fact that the researchers lead their own projects, “independently formulate hypotheses, design their own experiments, and analyze their results.”³⁰

Besides the negative effects on researchers’ work and careers, losing young researchers from research projects can also negatively affect the quality and public value of research. Younger researchers can bring fresh perspectives to cultural works. They are also more likely than older, more settled generations to come from the diverse backgrounds that can prompt distinct research questions and uncover valuable findings. Interrupting these researchers’ work when they change institutions means a loss of their unique perspectives and the specific research questions they may ask of corpora.³¹

In addition, as further detailed below, researchers have expressed uncertainty about the current exemption’s allowance for teaching by allowing students to use decrypted materials “at the direction of [a] researcher” in the same institution of higher education.³² This is a particular loss for data science education, because student experience working with “messy” digital humanities data is especially valuable.³³

Granting the proposed expansion would, as detailed below, address these ambiguities in the current exemption and ameliorate their current chilling effect on socially valuable research and teaching.

- b. The current exemption’s limits on sharing hamper valuable research because of the difficulty and costs associated with preparing usable corpora from lawfully acquired copies, and because the inability to share corpora for new inquiries impairs research quality and sustainability.

²⁵ *Id.* at 2.

²⁶ *Id.*

²⁷ *Id.* at 2–3.

²⁸ *Id.*

²⁹ *Id.* at 2.

³⁰ *Id.*

³¹ *Id.* at 2–3.

³² App. E: Letter from Joel Burges and Emily Sherwood, at 1.

³³ App. K: Letter from Lauren Tilton and Taylor Arnold, at 2.

Ambiguity is one issue. However, researchers have indicated that the ability to share their corpora so that others may *independently* analyze the works and ask different questions of a corpus is also crucial to the digital humanities. Likewise, researchers have also stated that they wish to study corpora built by others in order to conduct independent research. But, as the Mellon Foundation explains, “Because the current exemption limits the ways in which one research project can share access to their data with others, it has meant that subsequent research projects must start as if from a blank slate, effectively reinventing the wheel[.]”³⁴ This practical inability to share corpora with researchers engaging in independent inquiries hampers valuable research. At a minimum, it limits the research questions that can be asked, compromises research sustainability, creates inequities in research opportunities, and potentially biases research findings.

To begin, beyond the costs of lawfully acquiring copies of the studied works—which higher education institutions and researchers are prepared and willing to bear—lie the difficulty and costs associated with building and preparing usable corpora. These separate costs create large barriers to conducting TDM research. Moreover, because researchers are so limited in their ability to share corpora with researchers at other higher education institutions, research and data sustainability are harmed when institutions must bear the costs of needlessly duplicated labor. For example:

- In its letter of support, the Association for Computers and the Humanities describes the process of preparing corpora as “slow, painful, and expensive,” even for well-resourced institutions.³⁵ Beyond acquiring the works, bypassing TPM and preparing the materials in a research-ready format is nontrivial. The Association further states that “[r]equiring subsequent groups of scholars to undertake the technically complex and tedious processes of actually converting the material . . . does not provide any additional profit to copyright holders; it merely imposes an additional labor burden both on the scholars and the vendors who distribute these works. It delays research and simply serves as a disincentive for scholars to use text and data mining methods on in-copyright work.”³⁶
- Director Brandon Butler, in a forthcoming, co-authored analysis of how copyrighted works are used in research, notes that the concerns about sharing data “in some way impaired [] research” for 43% of researchers, with 23% changing the design of the project, 14% avoiding taking on a project, and 6% forced to abandon a project.³⁷ Scholars also expressed concerns that the “inability to work with more contemporary materials in digital humanities courses was making it more difficult to cultivate students’ interest in these courses, and even in the humanities more generally. As one researcher observed, ‘[s]tudents would be so much more engaged if we could use more contemporary literature.’”³⁸

³⁴ App. M: Letter from the Mellon Foundation, at 1.

³⁵ App. A: Letter from the Association for Computers and Humanities, at 1 (quoting a forthcoming paper from Stanford Literary Lab).

³⁶ *Id.* at 2.

³⁷ App. F: Letter from Brandon Butler, at 2.

³⁸ *Id.*

- Mark Algee-Hewitt has described concerns about how reduplicating the work of preparing a corpus “compromis[es research] as small differences between data sets often have an outsized impact on the research results.”³⁹
- John Bell writes about how his progress in his project analyzing three-dimensional pose and motion data in 20th-century film has been hampered by the time and cost of preparing the works for analysis and breaking the TPM—this process took most of the time of a year-long grant, requiring a request for an unfunded extension to complete the study.⁴⁰ He notes that the resulting corpus has value outside his specific research focus—but because it cannot be used by other researchers, another researcher would have to repeat the same effort.⁴¹
- Joel Burges and Emily Sherwood explain how they were required to engage in “massive redundancy” in labor costs to prepare the works for research and data analysis.⁴²
- Lauren Tilton mentions how, for audiovisual works, annotation data prepared by researchers is closely coupled with the underlying video file.⁴³ If the annotation uses timestamps to record specific information, then anyone wishing to build upon another researcher’s annotations must confirm that the compilation of the corpus did not introduce minute alterations that significantly affect the data. Through sharing the corpus, other researchers could ensure that the underlying data is exactly the same as the one Lauren Tilton used to create annotations, avoiding this issue.
- Henry Alexander Wermer-Colan cites the cost of preparing for research a dataset of books banned in the United States was “tens of thousands of dollars”, and notes that having a researcher bear this cost every time they wanted to engage in TDM research would be too costly to support further research in the digital humanities.⁴⁴
- David Bamman notes, simply, “The act of digitization is a laborious one[,]” and that a “hindrance to the larger goals of science . . . is the lack of our ability to share DRM-broken⁴⁵ materials with other researchers who in all other respects are following the protocol of 37 CFR 201.40(b)(4).”⁴⁶

The costs of taking a set of DVDs or e-books, breaking the TPM, and preparing them for research are substantial. The Mellon Foundation notes, “As a funder of these efforts, the Mellon Foundation

³⁹ App. B: Letter from Mark Algee-Hewitt, at 2.

⁴⁰ App. D: Letter from John Bell, at 1–2.

⁴¹ *Id.* at 2.

⁴² App. E: Letter from Joel Burges and Emily Sherwood, at 2.

⁴³ App. K: Letter from Lauren Tilton and Arnold Taylor, at 2.

⁴⁴ App. L: Letter from Henry Alexander Wermer-Colan, at 2.

⁴⁵ Professor Bamman is referring to “digital rights management” technologies, known in this proceeding as “technological protection measures” or “TPMs.”

⁴⁶ App. C: Letter from David Bamman, at 2.

has seen first-hand how expensive and complicated it is to build a corpus—which requires technical staff and expertise, as well as computing resources and tools.”⁴⁷

Yet, the Mellon Foundation points out, under the current exemption “each project must break independently the “digital locks” of [TPMs], process data, and build a corpus in a form that is useful for research.”⁴⁸ Under the current exemption, the costs of “effectively reinventing the wheel”⁴⁹ in this way would accrue over and over, because researchers must prepare corpora anew to study new questions, outside of direct collaborations or within a single higher education institution. In many cases, these new questions would simply go unanswered, as it is simply too expensive to put together corpora anew, again and again. These issues are especially acute in the humanities, where institutional support is often relatively limited.⁵⁰

Researchers are prepared to obtain their own copies of the underlying copyrighted works when a corpus is shared with them, and are prepared to comply with effective security measures. However, the technical and logistical capacity required to transform a collection of legally acquired works into a research-ready corpus, combined with the lack of ability to share corpora, requires researchers and institutions to waste resources duplicating efforts. These costs can be significant, and can prevent the research community from fully exploring what any given corpus has to tell us. Henry Alexander Wermer-Colan, for example, explains that he cannot support researchers engaging in TDM if he has to incur duplicated costs for every researcher with a corpus of interest.⁵¹ Similarly, Rachael Samberg, Scholarly Communication Officer and Program Director of UC Berkeley Library’s Office of Scholarly Communication Services (“OSCS”), and Timothy Vollmer, Scholarly Communication and Copyright Librarian at OSCS, support researchers’ efforts and thus have a birds-eye view of researchers’ struggles. They describe an exciting project they are supporting that is developing a corpus of “2,500 films and 800 television seasons to use in TDM research,”⁵² but note that:

Even if a corresponding scholar at another institution complied with the requirements of the TDM Exemptions and purchased the very same corpus works for study, the scholar would still have to pay thousands of additional dollars to set up a similar process to engage in what is ultimately duplicative circumvention and quality-checking. In many scholarly disciplines, these funds simply are not available.⁵³

In addition, Joel Burges and Emily Sherwood describe the limits of the current exemption as a “massive redundancy in both labor and purchasing costs” and that “too much of this kind of

⁴⁷ App. M: Letter from the Mellon Foundation, at 1.

⁴⁸ *Id.*

⁴⁹ *Id.*

⁵⁰ See App. F: Letter from Brandon Butler, at 3 (noting a “pattern of a lack of professional support [that] makes it even more important to expand the current exemption to remove barriers for researchers”).

⁵¹ See *id.* at 2.

⁵² App. J: Letter from Rachael Samberg and Timothy Vollmer, at 3.

⁵³ *Id.* at 3–4.

redundancy is intellectually and institutionally inefficient.”⁵⁴ These costs create barriers even though researchers and their institutions are able and willing to lawfully acquire the relevant works and to ensure that the security requirements and other exemption requirements are followed.

The current exemption’s limitations also shape how research is conducted when it does occur. Its limitations can lead to research bias in terms of the questions asked and prevent researchers from effectively critiquing and building on each other’s research. Professor Bell is concerned that the problem may “end[] many potential investigations before they can even begin.”⁵⁵ As established in the 2021 Triennial Proceeding, creative works by women, gender minorities, and artists of color are much more prevalent in cultural works from the 20th and 21st centuries compared to previous centuries.⁵⁶ The unnecessary, duplicative costs imposed by requiring corpora to be compiled anew for every new study pursued outside of the original institution disproportionately affect research into these works and these groups’ perspectives.

Worse, these unnecessary costs create barriers that can prevent smaller and less-well-resourced institutions from conducting TDM research at all. Without being able to use existing, pre-processed corpora, researchers at less-well-resourced institutions can be limited, or even precluded, from engaging in valuable research projects, despite owning the copies of the relevant motion picture or literary works.⁵⁷ The Mellon Foundation explains that, regrettably, “These costs have meant that TDM research that engages works protected by TPMs has largely been limited to projects at institutions that have the resources to compensate and maintain technical staff and infrastructure, supplemented by grants like those we have supported.”⁵⁸

And costs are not the only concern. Requiring each new team of researchers to start over can negatively affect research in the field for substantive reasons as well. The bodies of works researchers choose to study are not cobbled together arbitrarily. Rather, and in addition to their resource-intensive nature, corpora prepared for TDM projects possess immense academic value in and of themselves. Professor Cooper, for example, describes the careful methodology required to select works for the Kinolab corpus, which relies on her and Professor Burges’ academic expertise and research in order to review and apply existing scholarship on “race, ethnicity, gender, and sexuality in American film and television” in order to curate the collection.⁵⁹ Professor Algee-Hewitt also notes that researcher-designed corpora have more transparent content and organization and that researchers are concerned about logic for assembling them, history of how they were assembled, reasons they were assembled, and why the choices were made.⁶⁰ As he puts it, these

⁵⁴ App. E: Letter from Joel Burges and Emily Sherwood, at 2.

⁵⁵ App. D: Letter from John Bell, at 2.

⁵⁶ App. A: Letter from the Association for Computers and Humanities, at 2.

⁵⁷ *See, e.g.*, App. B: Letter from Mark Algee-Hewitt, at 2 (“What is possible for us at Stanford, for example, with the assistance of our well-funded library, would not be possible for scholars working at less well-funded public institutions.”).

⁵⁸ App. M: Letter from the Mellon Foundation, at 1.

⁵⁹ App. G: Letter from Allison Cooper, at 2.

⁶⁰ App. B: Letter from Mark Algee-Hewitt, at 3.

corpora “are purpose-built for research and documented for research use.”⁶¹ This allows later researchers to ask different questions of the corpus with full knowledge of the methodological strengths and limitations of the corpus, aiding “research consistency and sustainability.”⁶²

Currently, however, these promises remain unfulfilled. As Professor Long, of the University of Chicago, explains, “A single research team simply does not have the time to pursue the full range of questions that a several-volume collection of novels might open up.”⁶³ Though Professor Long’s Textual Optics Lab “constantly receive[s] requests from researchers from other university faculty and graduate students who are hoping to pursue their own research projects”—projects for which “there is no question that they are worthy of being pursued.”⁶⁴ But “these are not projects that members of [Professor Long’s] lab have the expertise to pursue or collaborate on.”⁶⁵

Accordingly, many questions remain unanswered, and the public loses out on valuable research. This loss is attributable to the limits on sharing for independent research. As the Mellon Foundation notes, “The barrier to sharing fosters a siloed approach to TDM efforts and prohibits projects from benefiting from shared understandings and learnings, which can often lead to innovation.”⁶⁶

The current exemption’s limits on corpora sharing also undermine the quality of research and obstruct the advancement of the research field by preventing researchers from evaluating and analyzing TDM methods themselves. When evaluating the efficacy of a computational technique, it is extremely difficult to comparatively evaluate the performance of two different methods if they are applied to two different data sets, as it introduces confounding factors that inhibit clean comparisons of the methods. But researchers lack confidence that this type of independent methods review is necessarily “collaboration” or “replication.”⁶⁷ Even since the 2021 Triennial Proceeding, there has been an explosion in the growth of computational methods that can be used to analyze creative works. When researchers are unable to evaluate how well these methods work, their research utility greatly decreases. Ultimately, Professor Bell explains, “[t]he evolving methodology of entire disciplines is being held back by the requirement to restrict a prepared corpus to its original research group.”⁶⁸

As an important example, researchers have developed techniques that can help combat bias commonly found in machine learning algorithms.⁶⁹ These projects may not necessarily qualify as “replication” or “collaboration.” At a minimum, this creates uncertainty around using these techniques to independently test corpora; at worst, it prevents it altogether. Yet employing bias-

⁶¹ *Id.*

⁶² *Id.*

⁶³ App. H: Letter from Hoyt Long, at 2–3.

⁶⁴ *Id.* at 3.

⁶⁵ *Id.*

⁶⁶ App. M: Letter from the Mellon Foundation, at 2.

⁶⁷ 37 C.F.R. § 201.40(b)(4)(i)(D), (b)(5)(i)(D).

⁶⁸ App. D: Letter from John Bell, at 2.

⁶⁹ See App. G: Letter from Allison Cooper, at 2.

combatting techniques is a highly socially important research activity, crucial for developing an accurate and complete understanding of our cultural heritage.

Relatedly, researchers and institutions have voiced concerns about research transparency and record-keeping.⁷⁰ While supporting researchers, higher education institutions also consider how to maintain a research environment that is financially sustainable and that can enable further research. These important academic practices conflict with the need to repeatedly spend time, energy, and capital in order to recreate corpora multiple times in order to engage in fair use via research.⁷¹

These roadblocks together affect how research and teaching are conducted and foreclose research that would be valuable to the public. Researchers in institutions that are unable to support corpora preparation must give up their projects. This meaningfully inhibits research by preventing the research community from asking new questions and offering different perspectives on the body of work contained in a particular corpus, depriving the public of valuable findings. Without being able to use existing corpora, researchers at less-well-resourced institutions can be limited, or even precluded, from engaging in valuable research projects, even though they own the copies of the relevant motion picture or literary works and have a secure research environment.⁷² Ultimately, these barriers diminish the quality of research and prevent the public gaining valuable knowledge. As the Mellon Foundation notes, “It does not benefit the ‘progress of science and the useful arts’ when technical barriers mean that this type of research can be done only by researchers with ample resources.”⁷³

3. Granting the proposed expansion will enable additional valuable research.

The current exemption has helped researchers in the digital humanities engage in new and socially valuable research. Granting the proposed expansion will address the roadblocks described above, enabling researchers to both more effectively continue their existing research and ask new questions in the field, “catalyz[ing] the speed and quality of TDM research”⁷⁴ into literary works and motion pictures. As the Mellon Foundation explains, “We have seen in a number of other grant areas the tremendous value of collaborative efforts to build, share, and innovate upon corpora.”⁷⁵ The proposed expansion could ensure that TDM research can provide the same value.

⁷⁰ App. E: Letter from Joel Burges and Emily Sherwood, at 3.

⁷¹ See App. L: Letter from Henry Alexander Wermer-Colan, at 2.

⁷² App. B: Letter from Mark Algee-Hewitt, at 2 (“What is possible for us at Stanford, for example, with the assistance of our well-funded library, would not be possible for scholars working at less well-funded public institutions.”); App. M: Letter from the Mellon Foundation, at 1.

⁷³ App. M: Letter from the Mellon Foundation, at 1.

⁷⁴ *Id.* at 2.

⁷⁵ *Id.*

For example, researchers have built a rich set of corpora to study, such as a collection of fiction written by African American writers,⁷⁶ a collection of books banned in the United States,⁷⁷ and a curated corpus of movies and television with an “emphasis on racial, ethnic, sexual, and gender diversity.”⁷⁸ Some have received requests from other researchers for corpora they have created, but are currently unable to assist due to lack of capacity to engage in additional collaborations.⁷⁹ And some, in turn, desire to work with other researchers’ corpora in order to further explore and contribute to knowledge about our cultural heritage.⁸⁰ The proposed expansion would thus increase both the quantity and the quality of research.

In the same way a single literary work or motion picture can evince multiple meanings based on the lens of analysis used, when different researchers study one corpus, they are able to pose different research questions and apply different methodologies, ultimately revealing new and original findings. A single group of researchers will, by necessity, build a corpus to “pursue a limited set of research questions.”⁸¹ Enabling broader sharing and thus, increasing the number of researchers that can study a corpus, will allow a body of works to be better understood beyond the initial “limited set of research questions.”

For example, Hoyt Long, co-director of the Textual Optics Lab at the University of Chicago, has received multiple requests from researchers desiring to study his corpora of general U.S. fiction and of African American writers.⁸² Proposed research directions include “exploring representations of climate” and “studying how African American English is expressed in by African American writers.”⁸³ While, as noted above, Professor Long is currently unable to support all the researchers who could develop useful research using the Textual Optics Lab corpus, the proposed expansion would allow researchers to independently explore these new questions.

In addition, sharing corpora more broadly would enable the development of better methods, thus improving the quality of research across the field. For example, John Bell notes that text and data mining research is a fast moving field and that “between the time when we finish our analysis and publish our results, our methods will be obsoleted by new technology and the first thing our readers will want to do is rerun the analysis using new models to produce more accurate results or examine a related research question that could not be addressed using current inference models.”⁸⁴ Similarly, Brandon Butler explains that “having a shared corpus where researchers can test

⁷⁶ App. H: Letter from Hoyt Long, at 3.

⁷⁷ App. L: Letter from Henry Alexander Wermer-Colan, at 1.

⁷⁸ App. G: Letter from Allison Cooper, at 2.

⁷⁹ See App. H: Letter from Hoyt Long, at 3

⁸⁰ See App. E: Letter from Joel Burges and Emily Sherwood, at 3–4; App. G: Letter from Allison Cooper, at 2–3.

⁸¹ App. H: Letter from Hoyt Long, at 3.

⁸² *Id.*

⁸³ *Id.*

⁸⁴ App. D: Letter from John Bell, at 2.

different methods and ask different questions would thus not only increase the efficiency of TDM research, but also reduce authority bias.”⁸⁵

The proposed expansion will also facilitate cross-institutional research and thus, increase the quality of research across the field. The field of digital humanities commands limited funding compared to some other fields, rendering acute the need for corpora sharing to advance the field.⁸⁶ Further, a single researcher or collaborative team has neither the expertise to fully plumb a single corpus, nor the capacity to engage in the collaborations it would take to do so. In its letter, the Association for Computers and the Humanities explains that “[t]he kinds of research questions that scholars who use text and data mining methods ask are often expansive, requiring more than one scholar’s expertise to responsibly and thoroughly interpret the data.”⁸⁷

As a specific example, Professor Cooper directs Kinolab at Bowdoin College, which is “a digital humanities laboratory for the analysis of narrative film and television.”⁸⁸ Kinolab develops “an at-scale, representative digital collection of narrative film and television clips,” including clips of close-up and other cinematic techniques and film languages.⁸⁹ Professor Cooper states that the Kinolab corpus would benefit from machine learning analysis, but that such research would require partnership with other, larger, institutions.⁹⁰ Likewise, Professor Burges and Director Sherwood note that the hundreds of hours of labor that were devoted to annotating and labeling their corpus could help train algorithms developed by other researchers.⁹¹ And similarly, Professor John Bell explains that a corpus could “serve as a basis for everything from deep examinations [of] what sets films made by diverse creators apart from mainstream films to technical analysis of cinematography to machine vision investigations uncovering hidden histories that human critics have overlooked.”⁹²

Therefore, the proposed expansion, by permitting researchers to share corpora with researchers at other institutions of higher education studying independent questions, “would create a more efficient research pipeline and speed up discovery and the advancement of knowledge.”⁹³ Moreover, the proposed expansion would improve the quality of digital humanities research by creating an environment where works are more thoroughly studied, methodology can more rapidly improve and adapt, and cross-institutional research is enabled. This, in turn, allows the field to

⁸⁵ App. F: Letter from Brandon Butler, at 2.

⁸⁶ See App. D: Letter from John Bell, at 2 (“In the arts and humanities fields where I work, funding is already difficult to come by and the labor required to build a corpus large enough to provide statistically significant datasets ends many potential investigations before they can even begin.”); App. K: Letter from Lauren Tilton and Taylor Arnold, at 1 (“There are significant funding limits in humanities subjects versus STEM in higher education”).

⁸⁷ App. A: Letter from the Association for Computers and Humanities, at 1.

⁸⁸ App. G: Letter from Allison Cooper, at 1.

⁸⁹ *Id.*

⁹⁰ *Id.* at 2–3.

⁹¹ App. E: Letter from Joel Burges and Emily Sherwood, at 3.

⁹² App. D: Letter from John Bell, at 2.

⁹³ App. J: Letter from Rachael Samberg and Timothy Vollmer, at 4.

better contribute to public knowledge and develop a fuller understanding of “the vibrant last century of cultural production,”⁹⁴ to further “increase equitable access to deep knowledge that helps to build an informed, heterogeneous, and civically engaged society . . . and ensure that more authentic, reflective, and nuanced stories are revealed, preserved, and told.”⁹⁵

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

The technological protection measures and methods of circumvention at issue for this proposed expansion include those measures and methods applicable to motion pictures and literary works distributed electronically. The proposed expansion does not materially alter the nature and basic operations of the relevant technological protection measures and methods of their circumvention as compared to the current exemption.

ITEM E. ASSERTED ADVERSE EFFECTS ON NONINFRINGING USES

1. The proposed class includes at least some works protected by copyright.

The proposed expansion includes copyrighted works because it includes the same types of works included in the current exemption.⁹⁶ Conducting TDM research in the digital humanities field requires creating a dataset of works of interest, which typically involves “motion pictures contained on DVDs or Blu-ray discs, or transmitted through streaming services, as well as literary works distributed electronically.”⁹⁷ As the Copyright Office noted in the 2021 Triennial Proceeding, “[t]here is no dispute that at least some of these works are protected by copyright.”⁹⁸

2. Conducting TDM research is likely to be noninfringing under Title 17.

As the Copyright Office concluded when recommending the current exemption, and affirmed in October 2023 when it recommended renewal of the current exemption, conducting TDM for scholarly research and teaching purposes is fair use.⁹⁹ This remains true for any research facilitated by the requested expansion. The requested expansion performs the same function as the current exemption itself: allowing researchers and students affiliated with a nonprofit institution of higher education to conduct TDM research, provided that the institutions abide by certain requirements,

⁹⁴ App. A: Letter from the Association for Computers and Humanities, at 2. *Id.*

⁹⁵ App. M: Letter from the Mellon Foundation, at 1.

⁹⁶ *See* 37 C.F.R. § 201.40(b)(4) & (5).

⁹⁷ 2021 Triennial Proceeding, at 103, 105–06.

⁹⁸ *Id.* at 106.

⁹⁹ *Id.* at 117 (“Balancing the four fair use factors, with the limitations discussed, the Register concludes that the proposed use is likely to be a fair use.”); Exemptions To Permit Circumvention of Access Controls on Copyrighted Works, 88 Fed. Reg. 72013, 72018 (Oct. 19, 2023).

including maintaining security and obtaining lawful copies of the works to be used.¹⁰⁰ Thus, the fair use analysis from the 2021 Triennial Proceeding regarding the current exemption applies with equal force in the proposed expansion.

The first factor considered in the fair use analysis is the “purpose and character” of the use.¹⁰¹ The proposed use covered by the proposed expansion is the same as the use covered by the current exemption, namely, to create a searchable collection of works for the purpose of TDM research.¹⁰² In the 2021 Triennial Proceeding, the Copyright Office concluded that this use is transformative when it is solely for noncommercial scholarly research and teaching purposes.¹⁰³ The Copyright Office based its conclusion on *Authors Guild, Inc. v. HathiTrust* (“**HathiTrust**”) and *Authors Guild v. Google* (“**Google Books**”), in which the courts held that the creation of a full-text searchable database is transformative because the resulting database is different in purpose and character from that of the original literary work.¹⁰⁴ The Copyright Office reasoned that, because “the intended purpose of the proposed activity is to provide information about works by identifying trends or calculating statistics, which differs from the expressive or informative purposes of the original works, the proposed use is similar to Google’s ngrams tool” and “to the report provided by HathiTrust.”¹⁰⁵ Accordingly, the Copyright Office concluded that “the proposed use is non-commercial and likely transformative.”¹⁰⁶

The proposed expansion covers use that is still solely for noncommercial scholarly research and teaching purposes. TDM research, which “is used to produce statistics and facts about copyrightable works,” is still non-expressive and highly transformative.¹⁰⁷ The reasoning and conclusions of *HathiTrust* and *Google Books* apply. Recently, the Supreme Court addressed the transformative inquiry in fair use analysis in *Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith* (“**Warhol**”).¹⁰⁸ The Court noted that the transformative purpose or character is a matter of degree, “and the degree of difference must be balanced against the commercial nature of the use.”¹⁰⁹ As Professor Sag notes in his letter, the *Warhol* decision “reinforces the importance of focusing on the particular use made by the defendant and the prospect that the use might result in

¹⁰⁰ See 37 C.F.R. § 201.40(b)(4)(i), (b)(5)(i).

¹⁰¹ 17 U.S.C. § 107(1).

¹⁰² 2021 Triennial Proceeding, at 104.

¹⁰³ *Id.* at 109.

¹⁰⁴ *Id.* at 109; *HathiTrust*, 755 F.3d at 97 (holding that creation of a full-text searchable database that did not show the user any of the text of the copyrighted works was transformative); *Google Books*, 804 F.3d at 217 (holding that the “snippet view,” which showed portions of unaltered, copyrighted text, was transformative because it “add[ed] important value to the basic transformative search function” by allowing users to verify that the list of books returned by the database was responsive to the user’s search).

¹⁰⁵ 2021 Triennial Proceeding, at 109–110.

¹⁰⁶ *Id.* at 109.

¹⁰⁷ App. I: Letter from Matthew Sag, at 2.

¹⁰⁸ 598 U.S. 508 (2023).

¹⁰⁹ *Id.* at 532–33.

competitive substitution for the plaintiff's expressive work.”¹¹⁰ Thus, in *Warhol*, the Supreme Court affirmed that “[d]eriving uncopyrightable information and insights from copyrighted expression is not just transformative, it is highly transformative.”¹¹¹

Here, both of the purposes at issue—conducting TDM research, and teaching—are highly transformative. Courts have consistently held that copying for the purpose of accessing information about the works, which is itself unprotectable, is highly transformative.¹¹² Thus, the *Warhol* decision supports a finding of fair use for the proposed use in this petition.

Moreover, uses that would be made under the proposed expansion are noncommercial, and for research or teaching purposes. These are quintessential fair use purposes explicitly listed in section 107 of the Copyright Act, both in its preamble and under the first factor.¹¹³ Therefore, in light of the statutorily favored, noncommercial, and highly transformative nature of the uses, the first factor weighs heavily in favor of fair use.

The second fair use factor considers the “nature of the copyrighted work.”¹¹⁴ Research corpora often include a mixture of works, depending on the topic being studied. Many of the works TDM researchers investigate—for example, movies or novels—are highly creative, while others—for example, scholarly works or factual reporting—may be less so. Courts have weighed the second factor against fair use when more creative works are copied.¹¹⁵ However, as the Copyright Office emphasized in the 2021 Triennial Proceeding, the nature of the work factor “is of limited

¹¹⁰ App. I: Letter from Matthew Sag, at 3.

¹¹¹ *Id.* at 2.

¹¹² *See, e.g.,* A.V. v. iParadigms Liab. Co., 544 F. Supp. 2d 473, 482 (E.D. Va. 2008) (“This Court finds the “purpose and character” of iParadigms’ use of Plaintiffs’ written works to be highly transformative. Plaintiffs originally created and produced their works for the purpose of education and creative expression. iParadigms, through Turnitin, uses the papers for an entirely different purpose, namely, to prevent plagiarism and protect the students’ written works from plagiarism. iParadigms achieves this by archiving the students’ works as digital code and makes no use of any work’s particular expressive or creative content beyond the limited use of comparison with other works.”); *AV Ex Rel. Vanderhye v. iParadigms, LLC*, 562 F. 3d 630, 640 (4th Cir. 2009) (“The district court, in our view, correctly determined that the archiving of plaintiffs’ papers was transformative and favored a finding of “fair use.” iParadigms’ use of these works was completely unrelated to expressive content and was instead aimed at detecting and discouraging plagiarism.”); *Authors Guild, Inc. v. HathiTrust*, 755 F. 3d 87, 97 (2d Cir. 2014) (“[W]e conclude that the creation of a full-text searchable database is a quintessentially transformative use.”); *Authors Guild, Inc. v. Google, Inc.*, 804 F.3d 202, 216–17 (2d Cir. 2015) (“We have no difficulty concluding that Google’s making of a digital copy of Plaintiffs’ books for the purpose of enabling a search for identification of books containing a term of interest to the searcher involves a highly transformative purpose, in the sense intended by Campbell.”).

¹¹³ 17 U.S.C. § 107 (“Notwithstanding the provisions of sections 106 and 106A, the fair use of a copyrighted work, including such use by reproduction in copies or phonorecords or by any other means specified by that section, for purposes such as criticism, comment, news reporting, teaching (including multiple copies for classroom use), scholarship, or research, is not an infringement of copyright.”); 17 U.S.C. § 1201 (a)(1)(c) (iii) (“[T]he Librarian shall examine. . . 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[.]”).

¹¹⁴ *Id.* § 107(2).

¹¹⁵ *See e.g.,* *Harper & Row v. Nation Enterprises*, 471 U.S. 539, 563 (1985) (“The law generally recognizes a greater need to disseminate factual works than works of fiction or fantasy.”).

significance in the analysis of this class.”¹¹⁶ This is consistent with precedent cases where other factors—most importantly, the first factor—weigh in favor of fair use. In *Google Books*, for example, the Second Circuit weighed the second factor in favor of fair use when “the secondary use transformatively provides valuable information about the original” regardless of whether the underlying works are factual or creative in nature.¹¹⁷ Thus, the second factor should not weigh against fair use.

The third fair use factor considers “the amount and substantiality of the portion used in relation to the copyrighted work as a whole.”¹¹⁸ Compiling TDM corpora to generate information about the underlying works often requires copying a substantial amount or the entirety of works. This amount of copying is necessary to fulfill the highly transformative purpose of extracting and generating information about the works in question, alone and in relation to one another. As the Copyright Office concluded in the 2021 Triennial Proceeding with regard to TDM research, “copying the entire work [to obtain data about the works] is likely reasonable.”¹¹⁹ The same reasoning applies with equal force to the proposed expansion, as the proposed expansion furthers the exact same “legitimate purpose”¹²⁰—copying to obtain information about the copied works.

The fourth fair use factor assesses the use’s impact on “the potential market for or value of the copyrighted work,”¹²¹ and whether secondary work serves as a substitute for the original work.¹²² The proposed expansion would cause neither actual nor cognizable market harm to the copyrighted works.

First, the proposed expansion would not actually harm the original market for the copyrighted work. In fact, the proposed expansion may actually increase the demand for the copyrighted work. Under the proposed expansion, the beneficiary intuitions would still be required to obtain lawful copies of the underlying works or licenses without a time limitation on access. Thus, an institution that does not already own a lawfully acquired copy of the subject works would be required to lawfully obtain them before it could accept the corpora. Accordingly, allowing corpora sharing between institutions that both already own copies or licenses of the underlying work would not harm the market for the works.¹²³ Rather, the proposed expansion facilitates research into the copyrighted works and encourages researchers to pay for copies or licenses of e-books and DVDs that their institutions may not otherwise buy.

Second, the proposed expansion would not result in cognizable market harm. As the Copyright Office concluded when recommending the current exemption, cognizable market harm is limited

¹¹⁶ 2021 Triennial Proceeding, at 111.

¹¹⁷ *Authors Guild, Inc. v. Google, Inc.*, 804 F.3d 202, 220 (2d Cir. 2015).

¹¹⁸ 17 U.S.C. § 107(3).

¹¹⁹ 2021 Triennial Proceeding, at 111.

¹²⁰ *Id.*

¹²¹ 17 U.S.C. § 107(4).

¹²² 2021 Triennial Proceeding, at 112.

¹²³ See App. I: Letter from Matthew Sag, at 3 (“The proposed expansion would not give any individual or institution access to a work that they did not already have.”).

to “market substitution,” which is unlikely here.¹²⁴ As the Office explained, “with the limitation that researchers may not use the copies of the copyrighted works in the corpus for their expressive purposes, the copies would not serve as substitutes for the original works.”¹²⁵ The current exemption limits use of the copyrighted work to nonexpressive purposes only.¹²⁶ The proposed expansion applies only to the same nonexpressive use. The person receiving the corpora still may only view of “the contents of the works solely for the purpose of verification of the research findings.”¹²⁷ Accordingly, as for the current exemption, the use enabled by the proposed expansion would not serve as a substitute for the original work. As the Copyright Office determined in the 2021 Triennial Proceeding, any claim of lost TDM licensing revenue is thus not cognizable market harm for purposes of the fourth factor analysis.¹²⁸

Third, the proposed expansion is not likely to cause unauthorized downloading and distribution of copyrighted works that would harm the market for the underlying works. The institutions receiving the corpora would still be required to comply with the security measures required by the current exemption.

Thus, the fourth factor analysis is unchanged from the 2021 Triennial Proceeding and should weigh in favor of fair use.

Therefore, balancing the four fair use factors, uses of works made under the proposed expansion are likely to be fair use.

3. Academic researchers and students currently are, and are likely to be in next three years, adversely affected in their ability to conduct research and study humanities using TDM techniques.

The Copyright Office considers adverse effects of the prohibition on circumvention on proposed uses under five statutory factors.¹²⁹ The five statutory factors are: “(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.”¹³⁰

¹²⁴ 2021 Triennial Proceeding, at 112 (*citing* *Campbell v. Acuff-Rose Music, Inc.*, 510 U.S. 569, 591 (1994)).

¹²⁵ *Id.*

¹²⁶ App. I: Letter from Matthew Sag, at 2 (“The text data mining relevant to this petition is used to produce statistics and facts about copyrightable works. These statistics and facts are not same as, or even substantially similar to, the original expression in the underlying works[.]”).

¹²⁷ 37 C.F.R. § 201.40(b)(4)(i)(c), § 201.40(b)(5)(i)(c).¹²⁷ 37 C.F.R. §§ 201.40(b)(4)(i)(c), 201.40(b)(5)(i)(c).

¹²⁸ 2021 Triennial Proceeding, at 113.

¹²⁹ § 1201(a)(1)(C).

¹³⁰ *Id.*

- a. As in the 2021 Triennial Proceeding, this expansion would not curtail the availability for use of copyrighted works.

Under the first factor, the Copyright Office considers the whether the proposed use would decrease the ability for use of the copyrighted works.¹³¹ In the 2021 Triennial Proceeding, the Copyright Office found this factor not against the grant of the current exemption, on the ground that “the proposed use is narrowly tailored to scholarly research, and it is unlikely that copyright owners would entirely withhold electronic versions of their works from the market” due to the exemption.¹³²

The proposed expansion in this petition does not deviate from the current exemption in its purpose and effect. It is a modest extension from a copyright perspective, which simply allows researchers who have otherwise complied with the current exemption to share corpora with researchers from another institution for a different research project.¹³³ Like the current exemption itself, the proposed use of the proposed expansion is still limited to academic research at defined institutions of higher education.¹³⁴ Institutions enabled by the proposed expansion must also own lawful copies or licenses without a time limitation on access.¹³⁵

Further, the expected number of uses is not likely to be substantial. Digital humanities is a relatively specialized field, with a small number of researchers trained in quantitative humanities research methods.¹³⁶ The proposed expansion merely allows researchers in this field to use corpora compiled by another institution for their own fair use. Therefore, the proposed expansion is not likely to create or increase incentives for rightsholders to withhold electronic versions of their work, and it is not likely to disrupt a licensing market that provides the same benefit even if such a market exists.

- b. The proposed expansion would increase the amount, quality, and societal value of TDM research and education.

In the 2021 Triennial Proceeding, the Copyright Office concluded that the current exemption’s proponents had established that the exemption would lead to new copyrighted works, and that it would lead to the statutorily favored activities of scholarship, research, and teaching.¹³⁷ For example, the Copyright Office noted that professors could teach TDM techniques to students using

¹³¹ 2021 Triennial Proceeding, at 119–20.

¹³² *Id.* at 120.

¹³³ App. I: Letter from Matthew Sag, at 4.

¹³⁴ 37 C.F.R. § 201.40(b)(4)(ii), (b)(5)(ii)

¹³⁵ *See* 37 C.F.R. § 201.40(b)(4)(i), (b)(5)(i).

¹³⁶ App. B: Letter from Mark Algee-Hewitt, at 1 (“As a relatively young field, the Digital Humanities remains quite small, and there are particularly few scholars with comprehensive training in all of the various methods offered by these fields.”).

¹³⁷ 2021 Triennial Proceeding, at 120.

“contemporary, diverse, and inclusive works.”¹³⁸ Accordingly, the Register weighed the second and third statutory factors in favor of granting the current exemption.¹³⁹

Here, granting the proposed expansion would result in more digital humanities research and thereby generate additional copyrighted works. It would also, following the Copyright Office’s reasoning in the 2021 Triennial Proceeding, further the statutorily favored purposes of scholarship, research, and teaching.¹⁴⁰

First, the proposed expansion would increase the quantity of TDM scholarship and research. When one set of researchers builds a corpus at their home institution, “it is typically built to pursue a limited set of research questions.”¹⁴¹ “A single research team simply does not have the time to pursue the full range of questions” a rich collection of material may open.¹⁴² The proposed expansion would reduce the unnecessary barriers that currently prevent researchers—especially those from less-well-resourced institutions—from conducting valuable digital humanities research. For instance, Professor Long at the Textual Optics Lab at the University of Chicago receives requests from other university faculty and graduate students to use the “large collection of general US fiction and a corpus of novels by African-American writers”, who are hoping to pursue their own research projects.¹⁴³ These proposed research questions include “extracting characters from text and their narrative framing; measuring narrative coherence and its degree of correlation to reader preferences; exploring representations of climate; constructing sentiment and emotion arcs; studying how [African-American English] is expressed in fiction by African-American writers; and investigating the construction of metaphorical language in the same body of fiction.”¹⁴⁴ However, because of the limited capacity to directly collaborate with “every researcher who wishes to work with these corpora,” many of these “worthy” questions go without being answered.¹⁴⁵ The proposed expansion “would instantly make possible the wide array of projects for which we have received requests.”¹⁴⁶

Second, the proposed expansion would increase the quality and substantive value of TDM research. As emphasized by many digital humanities researchers, different researchers’ ability to apply new perspectives, methods, or questions to an existing corpus is key to the success of the study of digital humanities.¹⁴⁷ The proposed expansion would result in research that reflects more

¹³⁸ *Id.*

¹³⁹ *Id.*

¹⁴⁰ *Id.*; 17 U.S.C. § 107.

¹⁴¹ App. H: Letter from Hoyt Long, at 2 (noting this issue in the textual corpus context).

¹⁴² *Id.*

¹⁴³ *Id.* at 3.

¹⁴⁴ *Id.*

¹⁴⁵ *Id.*

¹⁴⁶ *Id.*

¹⁴⁷ *See id.* at 3; App. K: Letter from Lauren Tilton and Taylor Arnold, at 2.

diverse viewpoints, methods, and subjects by allowing this additive practice.¹⁴⁸ Professor Bell notes, for example, that for a corpus incorporating common U.S. motion pictures, various research questions can be raised, ranging “from deep examinations what sets films made by diverse creators apart from mainstream films to technical analysis of cinematography to machine vision investigations uncovering hidden histories that human critics have overlooked.”¹⁴⁹ Professors Tilton and Arnold also express interest in studying the history of TV in the past fifty years using corpora at different institutions.¹⁵⁰

Additionally, as Professor Algee-Hewitt notes, the proposed expansion would allow younger and more diverse researchers to engage in TDM projects by removing the ambiguity surrounding the definition of “collaboration.”¹⁵¹ Similarly, the Mellon Foundation states that, “It is our belief that by expanding the number of institutions that can benefit from the technical work of breaking TPMs for TDM research, the proposed expansion of the exemption in the regulations would result in a more diverse and rich set of research projects.”¹⁵² These projects and new perspectives would be enabled by the proposed expansion, furthering the study of humanities and enriching our understanding of culture and society.

Third, the proposed expansion would also increase the quality of TDM research by facilitating studies of research methods and independent verifications of research findings.¹⁵³ Unlike older forms of humanities research, which mostly rely on qualitative techniques, TDM research uses “algorithmic analysis” that is quantitative in nature.¹⁵⁴ Similar to quantitative research in the STEM fields, TDM research “demands. . . practices like independent validation of results”¹⁵⁵ Yet, the current exemption’s limitations present significant barriers for researchers to investigate each other’s work through different research methods. For example, Professor Burges and Director Sherwood highlight the research value of comparing the close viewing approach of their TDM project “with a distant viewing analysis such as the one currently being led by David Bamman at UC Berkeley on a related corpus of film and television and/or ones historically led by Taylor Arnold and Lauren Tilton at the Distant Viewing Lab at the University of Richmond.”¹⁵⁶ To complement their close viewing method, Professor Burges and Director Sherwood explain that “cutting-edge methods such as distant viewing . . . would advance multiple fields of study” in their

¹⁴⁸ App. A: Letter from Association for Computers and the Humanities, at 2 (“Barriers to computational scholarship on in-copyright works functionally amount to limits of the diversity of what scholars can research using these methods.”); App. M: Letter from the Mellon Foundation, at 1–2 (“It is our belief that by expanding the number of institutions that can benefit from the technical work of breaking TPMs for TDM research, the proposed expansion of the exemption in the regulations would result in a more diverse and rich set of research projects.”).

¹⁴⁹ App. D: Letter from John Bell, at 3.

¹⁵⁰ App. K: Letter from Lauren Tilton and Taylor Arnold, at 2.

¹⁵¹ App. B: Letter from Mark Algee-Hewitt, at 3.

¹⁵² App. M: Letter from the Mellon Foundation, at 1–2.

¹⁵³ See App. D: Letter from John Bell, at 2.

¹⁵⁴ *Id.*

¹⁵⁵ *Id.*

¹⁵⁶ App. E: Letter from Joel Burges and Emily Sherwood, at 2–3.

project, which they plan to further investigate in the next three years, if the proposed expansion is granted.¹⁵⁷

For all TDM research, and perhaps especially for the studies of motion pictures, the ability to use existing research corpora is paramount to fully investigating others' research findings.¹⁵⁸ In part because DVDs often underwent different production runs over time, and may no longer be available at all, "it's hard to find the exact DVD that another group generated their data from."¹⁵⁹ Right now, researchers can share only metadata with unaffiliated researchers who are not working on the same project or for research practices that ensure quality control but that are not necessarily encompassed by "replication."¹⁶⁰ But metadata is not useful for independent validation purposes without the underlying images.¹⁶¹ Accordingly, researchers are deterred from critically and effectively examining each other's work. Without a robust system of independent validation of research finding, as Professor Bell observes, "[t]he evolving methodology of entire disciplines is being held back by the requirement to restrict a prepared corpus to its original research group."¹⁶²

In addition, the proposed expansion would increase the pedagogical value of the current exemption itself. While the current exemption has enabled the development of valuable curricula, the high costs of decryption and compiling corpora present a meaningful barrier for institutions of higher education to provide TDM-related courses. As Director Butler and his survey co-authors observe, scholars who are also teachers at higher education institutions worry that "the inability to work with more contemporary materials in digital humanities courses was making it more difficult to cultivate students' interest in these courses, and even in the humanities more generally."¹⁶³ The proposed expansion would enable more higher education institutions to provide the computational methods classes that are crucial for preparing students to work in "a growing number of data-driven sectors."¹⁶⁴ Corpora compiled for humanities research are particularly beneficial for data science teaching, because humanities data, such as that extracted from TV and film files, is messy.¹⁶⁵ The ability to work with complex, messy data is an important skill for the "next generation of data scientists" to develop.¹⁶⁶

Relatedly, while the current exemption allows students to use decrypted materials "at the direction of [a] researcher" in the same institution of higher education, researchers have expressed

¹⁵⁷ *Id.* at 3.

¹⁵⁸ App. K: Letter from Lauren Tilton and Taylor Arnold, at 2.

¹⁵⁹ *Id.*

¹⁶⁰ 37 C.F.R. § 201.40(b)(4)(i)(D), (b)(5)(i)(D).

¹⁶¹ App. K: Letter from Lauren Tilton and Taylor Arnold, at 2.

¹⁶² App. D: Letter from John Bell, at 2.

¹⁶³ App. F: Letter from Brandon Butler, at 2.

¹⁶⁴ App. A: Letter from Association for Computers and the Humanities, at 2.

¹⁶⁵ App. K: Letter from Lauren Tilton and Taylor Arnold, at 2.

¹⁶⁶ *Id.*

uncertainty about how to fit student research projects into in the current exemption.¹⁶⁷ The proposed expansion would remove the uncertainty that chilled many beneficial teaching projects and enlarge the pedagogical impact of the current exemption.

To conclude, as Professor Tilton and Arnold succinctly put it, the proposed expansion would “increase the quality and quantity of digital humanities research, and contribute to our understanding of history and culture.”¹⁶⁸ Thus, the second statutory factor weighs strongly in favor of the proposed expansion.

- c. The inability to use corpora compiled by other higher education institutions for different projects has negative consequences on TDM teaching, scholarship, and research.

As detailed above, both the quantity and quality of digital humanities research would suffer without the proposed expansion.

First, the current lack of ability to use existing corpora for different projects reduces the overall *quantity* of digital humanities research. In some cases, researchers who could produce valuable research using shared corpora are forced to give up their project entirely.¹⁶⁹

As the Association for Computers and the Humanities noted in its letter, the process of compiling a corpus for TDM research is “slow, painful, and expensive.”¹⁷⁰ Given how “expensive and complicated it is to build a corpus,”¹⁷¹ as detailed above, the cost of curating the corpora can be prohibitive for a significant number of researchers.¹⁷²

For example, for the Mellon-funded close-up project between the University of Richmond and Bowdoin College, costs including the “significant time, technology, and wages to [the studied works’] digitization at both institutions” are extremely high.¹⁷³ The current exemption’s limits on cross-institution corpora sharing mean that researchers who want to raise different research questions on the same corpora must incur “massive redundancy in both labor and purchasing costs”

¹⁶⁷ See, e.g., App. E: Letter from Joel Burges and Emily Sherwood, at 1. (“This burden not only slows down technologically innovative and often transdisciplinary research, but also contributes to unsustainable conditions for imaginative pedagogy and scholarly inquiry in, to name the fields in which Mediate has played a role at UR, the digital humanities, film and media studies, visual studies, cultural studies, musicology, and linguistics.”).

¹⁶⁸ App. K: Letter from Lauren Tilton and Taylor Arnold, at 1–2.

¹⁶⁹ App. F: Letter from Brandon Butler, at 2 (“One respondent was clear and direct, ‘I have stopped research on projects where copyright is confusing or otherwise impedes sharing of data.’”).

¹⁷⁰ App. A: Letter from Association for Computers and the Humanities, at 1.

¹⁷¹ App. M: Letter from the Mellon Foundation, at 1.

¹⁷² App. J: Letter from Rachael Samberg and Timothy Vollmer, at 4 (“Even if a corresponding scholar at another institution complied with the requirements of the TDM Exemptions and purchased the very same corpus works for study, the scholar would still have to pay thousands of additional dollars to set up a similar process to engage in what is ultimately duplicative circumvention and quality-checking. In many scholarly disciplines, these funds simply are not available.”).

¹⁷³ App. E: Letter from Joel Burges and Emily Sherwood, at 2.

of preparing the corpus.¹⁷⁴ Since “[t]here are significant funding limits in humanities subjects versus STEM in higher education,”¹⁷⁵ this kind of redundancy is particularly wasteful in digital humanities.¹⁷⁶ As a result, valuable discoveries about our culture and society are inhibited.

Second, the *quality* of digital humanities research would suffer if the proposed expansion were not granted, for several reasons. When small groups of well-funded research teams are functionally the only ones able to afford the high costs of the corpora curation process, the depth and breadth of the field will be limited and there is a great risk of research bias ensuing. As noted above, the costs of corpora building for institutions that otherwise comply with the current exemption means that valuable corpora are “locked up” at better-endowed institutions.¹⁷⁷ Researchers approach a corpus “with a variety of unique perspectives and techniques that all contribute to the understanding of that corpus and the texts within it.”¹⁷⁸ Thus, limiting the numbers and types of institutions that are capable of conducting TDM research would negatively affect the understanding of the materials within a corpus.

Moreover, for quantitative studies like TDM research, the quality of research depends on independent validation of the research methods, including the scope and content of the materials in the collection.¹⁷⁹ Thus, being able to examine the underlying corpus is particularly important to compare research methods in order to identify those that may reach different results and gain the knowledge necessary to increase accuracy and reduce research bias.

In computational studies of motion pictures, for example, “sample bias . . . has been well documented in machine learning.”¹⁸⁰ To combat this bias, in curating the materials for “A Digital History of the Close-Up in Narrative Film and Television,” Professor Cooper notes that they decided to build their own corpora of annotated film and television clips “to build a deliberate collection of close-up clips.”¹⁸¹ Specifically, Professor Cooper documents that the process “began with a wide-ranging review of existing scholarship and writing on the representation of race, ethnicity, gender, and sexuality in American film and television[,]” and involves weekly curatorial meetings with principal investigators and student curators across two institutions.¹⁸² This type of deliberative process is costly both in money and human resources. Indeed, a majority of the \$100,000 grant from the Mellon Foundation for this project is allocated to building a deliberate

¹⁷⁴ App. E: Letter from Joel Burges and Emily Sherwood, at 2.

¹⁷⁵ App. K: Letter from Lauren Tilton and Taylor Arnold, at 1.

¹⁷⁶ App. E: Letter from Joel Burges and Emily Sherwood, at 2 (“[T]oo much of this kind of redundancy is intellectually and institutionally inefficient.”).

¹⁷⁷ App. H: Letter from Hoyt Long, at 3; *see also* App. K: Letter from Lauren Tilton and Taylor Arnold, at 1 (“[O]nly the most affluent institution such as Stanford have the resources to spend millions of dollars on one data set, which is creating a huge inequality in access to data.”).

¹⁷⁸ App. H: Letter from Hoyt Long, at 3.

¹⁷⁹ *See, e.g.*, App. D: Letter from John Bell, at 2; App. F: Letter from Brandon Butler, at 2.

¹⁸⁰ App. G: Letter from Allison Cooper, at 2.

¹⁸¹ *Id.*

¹⁸² *Id.*

corpus of close-up clips, as opposed to the decryption itself. In 2023 alone, student curators for this project spent “nearly 600 hours” on the curation process.¹⁸³ This thoughtfully compiled corpus of close-ups, “with its emphasis on racial, ethnic, sexual, and gender diversity could” be used to “counter the kind of sample bias” that is common in machine learning.¹⁸⁴ This shows that the corpus itself is an aspect of TDM research that deserves independent review from other researchers to ensure that quantitative analyses of humanities are not based on biased samples. Without the opportunity for independent review, the quality of TDM research would suffer. But by the same token, as Director Butler highlights, “[h]aving a shared corpus where researchers can test different methods and ask different questions would . . . reduce authority bias.”¹⁸⁵

In addition, the quality of research would be inhibited without the proposed expansion because researchers are limited in their ability to create a larger impact beyond their direct research findings. The limitations on corpora sharing discourage full dialogue among researchers, prevent researchers from contributing to new areas of study in the field, and preclude both researchers and the public from receiving the full benefits of scholarly inquiry into corpora.

For example, Professor Cooper notes that the current exemption’s limitation on corpora sharing “is an impediment to Kinolab’s objective of developing its relatively simple film language data model into . . . a complex data model with the potential to represent more than just the visible and/or audible technical practices and aesthetic techniques in narrative film and media.”¹⁸⁶ Specifically, Professor Cooper is interested in developing “[a] film language ontology,” which “would be an expansive and detailed representation of [the] field’s collective knowledge about film language that could represent broad concepts such as cinematic space and cinematic time, relationships such as that of the sequence shot to the long take, the affective attributes of the close up, and more.”¹⁸⁷ It is impossible for one research team to achieve such a large-scale project. Despite the fact that “Kinolab’s platform would provide an ideal test bed” for the project,¹⁸⁸ the current exemption prevents Professor Cooper from offering her computational expertise and Kinolab’s collection to the project, without officially collaborating with the existing project. This limitation significantly hinders what could be an impactful research project—and many other projects could be explored if the corpora-sharing restrictions are modified.

- d. The proposed expansion is not likely to affect the licensing market for or the value of copyrighted motion pictures and literary works.

The proposed expansion is not likely to negatively affect the market for or the value of the studied works.

¹⁸³ *Id.*

¹⁸⁴ *Id.*

¹⁸⁵ App. F: Letter from Brandon Butler, at 2.

¹⁸⁶ App. G: Letter from Allison Cooper, at 3.

¹⁸⁷ *Id.*

¹⁸⁸ *Id.*

First, there is still no market for research corpora that meets researchers' needs. Accordingly, researchers with whom corpora would be shared under the proposed expansion, like researchers developing corpora under the current exemption, don't have somewhere else to go to license a corpus.

Second, there will be no actual impact on the market for original works. Under the proposed expansion, higher education institutions receiving the corpora would still be required to have or obtain lawful copies of the underlying works.¹⁸⁹ Indeed, the proposed exemption is merely a "very modest expansion" to allow corpora sharing between "institutions who would independently qualify for the current exemption," without researchers needlessly going through the process of DRM themselves.¹⁹⁰ If anything, as noted above, the proposed expansion is likely to positively influence the market for the underlying copyrighted work, by encouraging institutions to acquire more copies of literary works and motion pictures.

Third, the Copyright Office's reasoning in the 2021 Triennial Proceeding that non-expressive uses of copyrighted material would not "serve as substitutes for the original or interfere with licensing markets"¹⁹¹ applies with equal force to the present petition. Copies would be used under the proposed expansion in the same way, and for the exact same purpose, as under the current exemption. Particularly, in the 2021 Triennial Proceeding, the Copyright Office rejected opposition arguments based on lost licensing revenue in its fair use factor four analysis.¹⁹² Under *HathiTrust*, lost licensing revenue can only be considered in the analysis of the fourth fair use factor only when the use serves as a substitute for the original, not when the use is transformative.¹⁹³

As noted above in the fair use analysis, the proposed expansion covers a purpose—to undertake TDM for research and teaching—that is still non-infringing nonexpressive, and highly transformative.¹⁹⁴ Thus, lost licensing revenue—if there is any at all—is not considered cognizable harm under *HathiTrust* and the Copyright Office's previous reasoning.

Fourth, the proposed expansion would not result in uncontrolled dissemination of the copyrighted work that would affect the market for or the value of the copyrighted work. Under the proposed expansion, higher education institutions receiving a corpus from another institution would need to comply with the security measure of the current exemption by adopting "effective security measures to prevent further dissemination or downloading of literary works in the corpus."¹⁹⁵ In the 2021 Triennial Proceeding, the Copyright Office emphasized that "[t]he requirement to employ

¹⁸⁹ 37 C.F.R. § 201.40(b)(4)(i)(B), (b)(5)(i)(B).

¹⁹⁰ App. I: Letter from Matthew Sag, at 3.

¹⁹¹ 2021 Triennial Proceeding, at 120.

¹⁹² *Id.* at 113.

¹⁹³ *Id.* at 113 (*citing* Authors Guild, Inc. v. HathiTrust, 755 F. 3d 87, 100 (2d Cir. 2014)).

¹⁹⁴ *See* App. I: Letter from Matthew Sag, at 3.

¹⁹⁵ 37 C.F.R. § 201.40(b)(4)(i)(D), (b)(5)(i)(D).

robust security measures will further reduce the risk of public access and distribution of the copyrighted works.”¹⁹⁶ That is equally true for the proposed expansion.

Indeed, researchers almost uniformly note the efforts they put into complying with the security requirement when conducting TDM research pursuant to the current exemptions.¹⁹⁷ If the institution failed to comply with the required security measures, then it may still be liable if resulting disseminations of copyrighted works render the copies of materials in the corpora no longer fair use. But it is important to recognize, as noted by the Second Circuit Court in *Google Books*, that the mere possibility of misuse does not render the corpora sharing unfair.¹⁹⁸ In *Google Books*, the Court “recognize[d] the possibility that libraries may use the digital copies Google created for them in an infringing manner,” but nonetheless refused to “impose liability on Google for having lawfully made a digital copy for a participating library so as to enable that library to make non-infringing use of its copy.”¹⁹⁹ The Court reasoned that the mere “speculative possibility that the library may fail to guard sufficiently against the dangers of hacking” is not sufficient to hold “Google liable for its creation of a digital copy of a book submitted to it by a participating library so as to enable that library to make fair use of it.”²⁰⁰

As in *Google Books*, there is no more than a speculative possibility that the proposed expansion would result in misuse by the institutions receiving the corpora, particularly in light of the fact that the existing security requirements apply to recipient institutions just as they apply to sharing institutions. Thus, the proposed expansion is not likely to result in uncontrolled dissemination of the copyrighted work that would affect the market of the copyrighted work.

Fifth, like the current exemption itself, the proposed expansion reduces the perverse incentive for research to seek out and use unlawfully “liberated” texts. For example, Brandon Butler notes that illegal standardized corpora such as Books2 and 3 do exist and are well-known in the community.²⁰¹ The proposed expansion would similarly reduce the incentive for researchers without the resources needed to prepare their own corpora to look to shadow research libraries. Accordingly, the proposed expansion would at least disincentivize bad actors to break the protection measures in the first place.

Finally, the statute allows consideration of “such other factors as the Librarian considers appropriate.”²⁰² By making the current exemption more available, a more diverse set of researchers and institutions will be able to engage in text and data mining research and contribute to a fuller understanding of our literary heritage. In the 2021 Triennial Proceeding, in granting the current exemption, the Copyright Office recognized that using copyrighted literature and motion pictures

¹⁹⁶ 2021 Triennial Proceeding, at 120.

¹⁹⁷ See App. A: Letter from the Association for Computers and the Humanities, at 1.

¹⁹⁸ See *Authors Guild v. Google, Inc.*, 804 F.3d 202, 229 (2d Cir. 2015).

¹⁹⁹ *Id.*

²⁰⁰ *Id.*

²⁰¹ See App. F: Letter from Brandon Butler, at 2.

²⁰² 17 U.S.C. § 1201(a)(1)(C)(v).

to conduct TDM research is fair use.²⁰³ Yet the ability to conduct research using this exemption depends largely on the resources available at each institution.²⁰⁴ For researchers and students at less-well-resourced institutions, the costs of building corpora can be prohibitive.²⁰⁵ This should not be the case. As Professor Long highlighted, the proposed expansion would “go a long way to closing this equity gap and ensuring that TDM research can productively be carried out by researchers of diverse backgrounds and perspectives.”²⁰⁶ Accordingly, the proposed expansion is supported by the Constitutional principles of copyright and will enable more complete, diverse, and accurate research, to the ultimate benefit of the public.

4. The statutory prohibition on circumventing access controls is the cause of the adverse effects.

TDM researchers have offered ample evidence that the statutory prohibition on bypassing access controls—and nothing else—is the “but-for” cause of the adverse effects outlined above. But-for the prohibition on circumvention, there would be no need for the current exemption, as the research and teaching activities allowed under the current exemption are fair use under copyright law.²⁰⁷ And but-for the prohibition on circumvention and the limitations to the current exemption, researchers would be able to share their corpora with other researchers in different institutions for TDM research and teaching, purposes the Copyright Office has concluded to be fair use.²⁰⁸

Finally, as established in the 2021 Triennial Proceeding—and unfortunately continuing today—there are no reasonable alternatives to circumvention that would address the shortcomings of the current exemption and allow its full promise to be fulfilled. As established in multiple letters of support, “ensur[ing] that that *contemporary* history, culture, and society are not omitted from the scholarly record”²⁰⁹ requires researchers to circumvent encryption in order to compile corpora that represent the full scope of culture, into and through the 20th and 21st centuries.²¹⁰ There has been

²⁰³ 2021 Triennial Proceeding, at 117.

²⁰⁴ See App. A: Letter from the Association for Computers and the Humanities, at 1 (“Even in the most ideal context of well-resourced institutions with robust data-storage and computational infrastructure, and staff expertise on both the library and computing side, these scholars have found the process ‘slow, painful, and expensive’”); App. B: Letter from Mark Algee-Hewitt, at 2 (“What is possible for us at Stanford, for example, with the assistance of our well-funded library, would not be possible for scholars working at less well-funded public institutions.”); App. H: Letter from Hoyt Long, at 3 (“Permitting a corpus to only be used with a single researcher’s perspective leads to needlessly duplicated effort in preparing the same corpus and to growing disparities between institutions.”); App. K: Letter from Lauren Tilton and Taylor Arnold, at 1 (“Add that only the most affluent institution such as Stanford have the resources to spend millions of dollars on one data set, which is creating a huge inequality in access to data.”).

²⁰⁵ *Id.*

²⁰⁶ App. H: Letter from Hoyt Long, at 3.

²⁰⁷ See 2021 Triennial Proceeding, at 107–17 (analyzing fair use based on precedents).

²⁰⁸ *Id.* at 117.

²⁰⁹ App. M: Letter from the Mellon Foundation, at 1.

²¹⁰ App. A–M.

no material change to the lack of availability of licensed alternatives,²¹¹ and public domain works do not represent the full scope of “society’s complex, intertwined humanity.”²¹² Allowing the proposed, modest expansion to the current exemption, however, would significantly increase the breadth, depth, and diversity of TDM research, to the benefit of society.

²¹¹ Exemptions To Permit Circumvention of Access Controls on Copyrighted Works, 88 Fed. Reg. 72013, 72018–19 (proposed Oct. 19, 2023) (to be codified at 37 C.F.R. pt. 201).

²¹² App. M: Letter from the Mellon Foundation, at 1.

DOCUMENTARY EVIDENCE

[Please see attached Appendices]

APPENDICES

Appendix A: Letter from the Association for Computers and Humanities

Appendix B: Letter from Mark Algee-Hewitt

Appendix C: Letter from David Bamman

Appendix D: Letter from John Bell

Appendix E: Letter from Joel Burges and Emily Sherwood

Appendix F: Letter from Brandon Butler

Appendix G: Letter from Allison Cooper

Appendix H: Letter from Hoyt Long

Appendix I: Letter from Matthew Sag

Appendix J: Letter from Rachael Samberg and Timothy Vollmer

Appendix K: Letter from Lauren Tilton and Taylor Arnold

Appendix L: Letter from Henry Alexander Wermer-Colan

Appendix M: Letter from the Mellon Foundation

Appendix A
Letter from the Association for Computers and Humanities



The Association for
Computers and the
Humanities

November 13, 2023

Dear Librarian of Congress,

On behalf of the members of the Association for Computers and the Humanities (ACH), the US-based professional organization for digital humanities scholars, we advocate in favor of the proposed expansion of the exemptions to DMCA § 1201 -- 37 C.F.R. 201.40(b)(4 & 5), covering text and data mining use of literary works and motion pictures. We support the sharing corpora with researchers affiliated with a different nonprofit educational institution, assuming the researchers meet the same requirements as those who compiled the corpus. This is an essential mitigation of a process that is slow, painful, and expensive, due to a combination of barriers imposed by content vendors, and the security requirements set by the exemption itself.

While traditional humanities scholarship is most commonly undertaken by solitary scholars, cross-institutional collaboration has been a distinctive trait of digital humanities as an interdisciplinary field since its inception. The kinds of research questions that scholars who use text and data mining methods ask are often expansive, requiring more than one scholar's expertise to responsibly and thoroughly interpret the data. It is not unremarkable for projects to span several languages or centuries. At the same time, the relative newness of digital humanities means that not every institution is likely to have multiple scholars who do computational work on literary texts or films -- let alone ones working on the necessary language or time period for any given project. Consequently, it is highly likely that any ambitious project that aims to address societal change at a large scale, or draws upon specific linguistic or cultural knowledge from several communities, will have team members spread across institutions.

While the DMCA exemptions 37 C.F.R. 201.40(b)(4 & 5) make it legally possible to use motion pictures and texts distributed electronically for non-consumptive text and data mining research, this does not mean the process of acquiring the materials, circumventing the technological protection measures, and converting the materials into a format compatible with computational analysis (e.g. plain-text computer files for literature, video files for motion pictures) is trivial. Several of our members have been working on putting the DMCA exemption to use. Even in the most ideal context of well-resourced institutions with robust data-storage and computational infrastructure, and staff expertise on both the library and computing side, these scholars have found the process "slow, painful, and expensive", to quote the title of a forthcoming paper from the Stanford Literary Lab. The restrictive security measures that come with treating in-copyright works as commensurate to the University's other "highly secure data" add non-trivial workflow complications to the process of circumventing the technological protection measures. For ebooks, acquiring the data in a usable format may, depending on the vendor through whom the institution has purchased the ebook, require a time-consuming back-and-forth with the vendor, adding months of delay.

The proposed expansion requires the collaborators at another institution to meet the same requirements around security and legal, institutional ownership of the data as the people who are sharing the computationally-ready corpus. As such, there is no loss of profit on the side of the copyright holders through the proposed expansion to cover sharing corpora, and the data will continue to be stored in with the same extreme level of security as set forth in the exemption as it exists. Requiring subsequent groups of scholars to undertake the technically complex and tedious processes of actually converting the material or communicating with vendors to get usable copies does not provide any additional profit to copyright holders; it merely imposes an additional labor burden both on the scholars and the vendors who distribute these works. It delays research and simply serves as a disincentive for scholars to use text and data mining methods on in-copyright works. As we stated in our previous letter of support for the current exemption, creative works by women, gender minorities, and artists of color were published commercially in the 20th century at rates far surpassing previous centuries. Barriers to computational scholarship on in-copyright works functionally amount to limits of the diversity of what scholars can research using these methods. Furthermore, teaching computational methods to students can enable them to seek employment in a growing number of data-driven sectors. When the only materials that can be studied using these methods are from the far-distant past and represent a very narrow range of life experiences and perspectives, it is more likely to turn them off than spark their imagination. The long-term consequences for the future of the academy in producing new scholars, and for society in developing a computationally-savvy workforce, are serious.

An expansion to the exemption would not address the fact that it remains slow, painful, and expensive to use the exemption -- but it would, at least, make the experience somewhat less slow and painful for subsequent scholars who work with that same data, in a context where the slowness and pain do not lead to any concrete benefit for the rightsholders while actively discourage research on the vibrant last century of cultural production.

Best,

A handwritten signature in black ink, appearing to read "Quinn Dombrowski", with a stylized, cursive script.

Quinn Dombrowski
co-President, Association for Computers and the Humanities

Appendix B
Letter from Mark Algee-Hewitt

December 12, 2023

Dear Register of Copyrights,

I am writing in strong support of the petition to extend the Text and Data Mining exemption to the Digital Millennium Copyright Act. I am a professor of Digital Humanities at Stanford University, where I run a collaborative research group called the Stanford Literary Lab. Our work applies computational models to corpora of literature for the purpose of researching questions of humanities interest. My work therefore depends on the availability of in-copyright works to perform our data mining-based research and it is directly affected by the current exemption to the DMCA, which makes much of my current research program possible.

In the past year, I have received a Public Knowledge Grant from the Andrew Mellon Foundation to undertake a research project in my lab on the history of literary theory. The research involved in this grant requires me to build a corpus of in-copyright academic texts, extracting the underlying text using the DMCA exemption. Because of the exemption, our lab has been able to evaluate 20th century literary theory and criticism through a new lens as we examine how the discourse of identity has evolved from its origins in theoretical academic texts to its dissemination into literature. In addition, I have also made use of the exemption to teach TDM techniques to students, helping them engage in their own projects and train the next generation of digital humanities scholars. Without the exemption, such research and teaching would not be possible.

As a result, I have recent, first-hand, experience with the practical application of the exemption to academic research, particularly in regards to the challenges that still remain. As important as the current exemption is to the possibility of future research and scholarship in the Digital Humanities, there are necessary modifications to the exemption in order to make the work that I, and my students, do possible.

The work of text and data mining in a humanities context is an inherently collaborative undertaking. The research involved requires expertise in a number of different areas, including, for my research team, literary study, statistics, computer science, and the social sciences. As a relatively young field, the Digital Humanities remains quite small, and there are particularly few scholars with comprehensive training in all of the various methods offered by these fields. As such, it is vital that any research project have members who come from a variety of backgrounds. Ideally, we would assemble such a research team from the scholars at a single university; however, this is often not possible given the relatively few academics with any research expertise in the Digital Humanities. Because of this, I have collaborated on projects with scholars both nationally and internationally. In all of these cases, it was crucial that we share data among the different constituent members of the research team. While frequently this involves small groups of people working on the same project, it often takes the form of multiple teams working on a set of related problems. In cases like this, it is essential that the same resources are made

available to all of the members of all of the teams. We are, however, unsure what “collaboration ... of the research” constitutes under the exemption. Our collaborations range greatly in scope and distance, from two team members working closely together, to multiple groups of scholars working independently to verify each other’s work through different quantitative methods. When we collaborate on understanding an archive of text through TDM methods, it sometimes means that we will be cooperating in applying the same methods to the same texts, and sometimes indicates that we will be taking diverging approaches to the same set of materials. These can be formal collaborations under the auspice of a grant, ad hoc collaborations that result from two teams discovering that they are working on similar material to the same ends, or even discussions at conferences between members of a loose network of scholars working on the same broad set of interests. As the exemption is unclear what counts as a collaboration for the purpose of sharing extracted data, we have had to be exceptionally cautious about sharing in-copyright material with any collaborators at all, much to the detriment to our research, and the field as a whole.

When we are working on projects involving in-copyright data, it is unreasonable to expect all of the collaborating teams to source their own data using the DMCA exemption given the differences in resources and capacity of different institutions. What is possible for us at Stanford, for example, with the assistance of our well-funded library, would not be possible for scholars working at less well-funded public institutions. Even when it is possible for each team to assemble their own dataset, the research is still compromised as small differences between data sets often have an outsized impact on the research results. As such, it is vitally important that we are able to share the same data set between groups of researchers. Even in cases where we are not collaborating on the same project, making the same data available for replication and validation studies is a crucial part of the research process.

An even more central problem in our inability to share in-copyright data with other scholars can be seen in the nature of my lab. Like most academic research organizations, my lab is primarily staffed with graduate students, post-doctoral scholars, and other junior academics who work collaboratively on shared research projects. These students bring important intellectual and personal diversity to both the field and the projects that they work on. While all members of my lab provide mutual support (and I, in the role of the director, mentor many of the projects), these students, as junior researchers, independently formulate hypotheses, design their own experiments, and analyze their results. Ideally, these students complete the PhD degree, or their post-doctoral fellowship, and move elsewhere to their own research positions at other institutions. Rarely, however, does such a move coincide with the completion of the research projects that they are working on with my lab. In these cases, the former students often continue working with me (and my group), transitioning from students at my university, to academics (post-docs, professors, sometimes staff) at other Universities. Given the current provisions of the exemption, however, it is unclear whether or not they are permitted to continue working on *their* research projects, which they have already put a significant amount of time into. In many cases, the university administration has taken a conservative legal interpretation, and these researchers can no longer continue their projects as they can no longer access the data that they gathered as members of my Lab. If we are unable to share the in-copyright resources that are often the basis of these projects, it

puts an artificial cap on the ability of these young scholars to complete their research in ways that are beneficial both to their careers and to the field as a whole.

In all, while the exemption has proven to be a crucial step in the right direction in terms of enabling text and data-mining research into humanities subjects of the 20th and 21st centuries, thereby giving students of human culture vital resources research up-to-the-minute phenomena, many challenges still remain. The lack of the ability to share material between research groups—particularly if members of a new group are former members of the team that originally obtained the in-copyright work—severely hampers our ability to do research and negatively impacts both the career paths of young scholars and the diversity of the field. Similarly, the lack of clarity around the security requirements of storage and use of the material we extract limits what we are able to do with the material, particularly in cases where the university adopts the most conservative reading of the exemption. Finally, the disjunction between the legal ability to extract text and image data from purchased media and the terms of service of many distributors that prevent that activity makes it even hard to assemble a workable corpus of material.

If the expansion were to be granted, not only would that clarify the ambiguities that have prevented researchers from examining these texts, but it would also enable new fields of research. Corpora created by other academics are typically assembled with careful consideration and expertise—they are purpose-built for research and documented for research use. In fact, the assembly of corpora itself has become a subject of scholarly concern with the Digital Humanities community and properly bibliographically sourced and responsibly maintained corpora are the product of tremendous scholarly labour on the part of experts in the field. The composition of any particular corpus contains the information of how the works were assembled, reasons the works were chosen, and why the choices were made. This information is valuable by itself, and aids research consistency and sustainability. Further, any single corpus will present more questions than a single researcher can study. The ability to share these corpora within the research community, rather than having them siloed at their respective institutions, would further develop our understanding of our shared literary culture.

For these reasons, I greatly support any efforts to amend and extend the current TDM exemption to the DMCA. The exemption promises to be transformative to whole branch of current research. In its current form, however, it still introduces unsupportable limits into the research that it purports to enable. If I can provide any additional information or evidence as to the utility of the current exemption or the desperate need to extend and amend it, then I would be more than happy to provide it.

Many thanks for your attention in this matter.

Yours,

A handwritten signature in black ink, appearing to read 'Mark Algee-Hewitt', is written over a light gray rectangular background.

Mark Algee-Hewitt
Director, Stanford Literary Lab
Director of Graduate Studies, Program in Modern Thought and Literature
Associate Professor of English and Digital Humanities

Appendix C
Letter from David Bamman

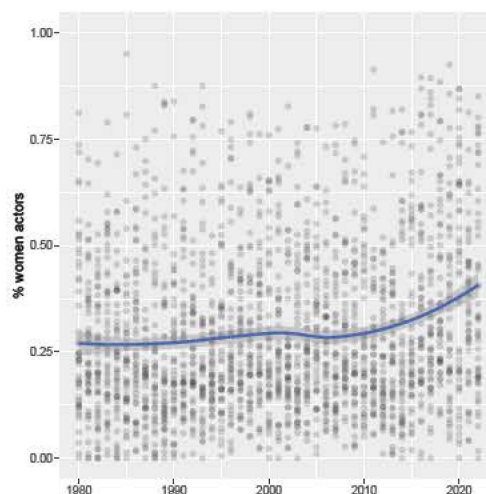


SCHOOL OF INFORMATION
102 SOUTH HALL #4600
BERKELEY, CALIFORNIA 94720-4600

October 17, 2023

I am writing this letter in support of the Authors Alliance's petition to the Copyright Office for an expansion of the current text and data mining exemption to §1201. I am an associate professor in the School of Information at UC Berkeley (with an affiliated appointment in the Department of Electrical Engineering and Computer Sciences), a senior fellow at the Berkeley Institute of Data Science, and faculty member of the Berkeley Artificial Intelligence Research Lab (BAIR). My research is centered on the areas of natural language processing and cultural analytics, where I focus on two complementary goals: improving the state of the art for computational methods for literary and cultural objects¹ and applying NLP and machine learning to empirical questions in the humanities and social sciences.² My work predominantly explores the affordances of empirical methods for the study of literature and culture, and has been recognized by the National Endowment for the Humanities, the National Science Foundation, and an NSF CAREER award. I offer these views in my individual capacity as a researcher working in text data mining and cultural analytics, and not on behalf of any organization.

The Eighth Triennial Rulemaking of 2021 allowed researchers to lawfully break DRM on DVDs in order to carry out research in text and data mining, subject to a number of restrictions outlined in §37 CFR 201.40(b)(4). This current exemption has allowed us to carry out substantial research investigating the representation of gender, race, and guns in contemporary movies over the period 1980-2022, applying computational models to the analysis of approximately 2,000 films. While this work is still ongoing, we are already seeing impactful findings as a result of this ruling; the figure to the right illustrates the rate with which women appear on-screen in the top films by U.S. box office over this period, with men appearing three times more often than women. This work confirms prior findings on the representation of women on-screen, but allows us to take more granular measurements than previous work has considered. This research is leading to fundamentally new facts about film and would simply not be possible without the 2021 exemption to §1201; I have in large part restructured my research agenda around this new ability to examine culture in film using empirical methods, so I cannot



¹See, for example: Sandeep Soni, Amanpreet Sihra, Elizabeth F. Evans, Matthew Wilkens and David Bamman (2023), "Grounding Characters and Places in Narrative Text," ACL; Andrew Piper, Richard Jean So and David Bamman (2021), "Narrative Theory for Computational Narrative Understanding," EMNLP; David Bamman, Olivia Lewke and Anya Mansoor (2020), "An Annotated Dataset of Coreference in English Literature," LREC 2020; Matthew Sims, Jong Ho Park and David Bamman (2019), "Literary Event Detection," ACL 2019; David Bamman, Sejal Popat and Sheng Shen (2019), "An Annotated Dataset of Literary Entities," NAACL 2019;

²See: Matthew Sims and David Bamman (2020), "Measuring Information Propagation in Literary Social Networks," EMNLP 2020; and Ted Underwood, David Bamman, and Sabrina Lee (2018), "The Transformation of Gender in English-Language Fiction," *Cultural Analytics*.

overstate the impact of the last decision has had—I see it as a fundamental turning point that will contribute new knowledge to the public good for years to come.

The exemption is not without its friction. For my own work, I break DRM twice on each DVD (carrying out the process of digitization twice to make use of the affordances of two different exemptions for different research purposes), and the ability to break DRM on ebooks has unfortunately not materialized into research since many ebooks from providers like Amazon have contractual terms of service that prevent breaking DRM, which this ruling does not override. One further hindrance to the larger goals of science, however, is the lack of our ability to share DRM-broken materials with other researchers who in all other respects are following the protocol of 37 CFR 201.40(b)(4) — i.e., purchasing the DVDs at their own institutions, computing in secure research environments, etc. The act of digitization is a laborious one, and the ability to share digitized materials subject to the original restrictions would help further accelerate the research being carried out in this space.

Sincerely,

A handwritten signature in black ink, appearing to read 'David Bamman', with a stylized, flowing script.

David Bamman
Associate Professor
School of Information
University of California, Berkeley

Appendix D
Letter from John Bell

December 4, 2023

I am writing to support the Authors Alliance's petition to expand the Text and Data Mining Exemption to the Digital Millennium Copyright Act. I am the program director for Dartmouth College's Data Experiences and Visualizations Studio, where I also conduct research on a variety of digital arts and humanities topics and teach as a Lecturer in the Film and Media Studies Department. Elsewhere, I teach in the University of Maine's Digital Curation graduate program and am a Senior Researcher at the Still Water Lab studying network art and culture. Though it is a result of my experience in all of these roles, my support for expanding the Text and Data Mining Exemption is my own and I am not writing on behalf of any institution or organization.

My most relevant work for the purposes of this petition is my ongoing collaborative research investigating film and television history. I have been part of Dartmouth's Media Ecology Project ("**MEP**") for more than a decade. MEP is dedicated to working with archives to both provide better access to historic moving image collections for scholars as well as return metadata to those archives that improve discoverability within their collections. Many libraries and archives have large collections of moving images that are incompletely cataloged, and in today's society of information overload, a film that is not digitally cataloged and easily discoverable is destined to remain unseen and unappreciated. Even in a best-case scenario where a film has been digitized and has a descriptive metadata record, that metadata only describes the film at a very high level and is composed of some basic identifying information, a few topic tags, and maybe a paragraph or two of text. Granular metadata that describes the content of the film down to individual scenes and shots is exceedingly rare, but that is exactly the type of information needed for scholars who want to use quantitative analysis methods to better understand how small, recurring details repeated across a large corpus form the building blocks of aesthetic movements. Machine generation of metadata is the only realistic way to examine these collections at scale, making the Text and Data Mining Exemption a critical tool for the preservation and quantitative analysis of our cultural heritage.

For example, we are leveraging the current version of the Text and Data Mining Exemption in a project MEP has been developing for the last year called Deep Screens. For this project we are using cutting-edge deep learning systems to extract three-dimensional pose and motion data from actors in a broad selection of two-dimensional titles from US film and television history. Our goal is to better understand how acting styles progressed from the early days of film, when most actors were trained for the stage, to today, when the craft of acting for the lens rather than a live audience is well understood. We are also comparing movements across genres, creators, and movements, from socially influential indie films to massive Hollywood blockbusters. To do so requires generating data from a large corpus, so we are extracting these films from more than 800 DVDs and Blu-Ray discs in Dartmouth's collection.

The existing provisions of the Text and Data Mining Exemption make this study possible because we would not be able to legally extract the films from these discs without it. While the current Exemption makes Deep Screens possible, though, it does not make it easy. The grant funding this research was originally set to last for one year; the first four months of that year ended up being spent on setting up a basic environment that conformed to the rules for using the Exemption. Ambiguity in the Exemption's description of a security model for storage of the corpus led to confusion among our research infrastructure group and, in an effort to comply, we built an entirely custom method of storage for the Deep Screens project that was more restrictive than what we use for even personally identifiable health information we use in medical research. Once the environment was set up, technical issues related to extracting the corpus from our source discs further delayed our research as some publishers and mastering methods protect disc content in ways that are more difficult to work around. These and other logistical problems that resulted meant we were more than halfway through the original grant period before we could even begin substantive work on the subject of the study itself. As of this writing, data extraction for Deep Screens is still in progress.

Though these films are some of the most critical cultural touchstones of the last century—everything from well-known classics like *Citizen Kane* to canonical films like those associated with the LA Rebellion that launched so many African American filmmakers in the late 20th century—there was no existing research-ready corpus we could draw on. If such a shared resource were legally possible, these are certainly the films and shows that would be in it. Despite restricting ourselves to media already owned by Dartmouth, we had to spend the majority of time originally allotted for research on the basics of gathering and organizing films and now are forced to seek an unfunded extension to complete the study itself. The fact that our requested extension only provides additional time, not additional money, means Deep Screens will inevitably overrun its budget as we have to pay researchers and students for two years rather than one.

This experience is, in my opinion, entirely unnecessary. As mentioned, the films and shows we are looking at are common objects of study and would undoubtedly be included in any shared corpus encompassing US moving image history. If researchers could share corpora, then these common films could serve as a basis for everything from deep examinations of what sets films made by diverse creators apart from mainstream films to technical analysis of cinematography to machine vision investigations uncovering hidden histories that human critics have overlooked. Instead, much of the time and effort dedicated to every individual study must be spent recreating a corpus composed of many of the same films that other researchers have already collected for their own work. In the arts and humanities fields where I work, funding is already difficult to come by and the labor required to build a corpus large enough to provide statistically significant datasets ends many potential investigations before they can even begin. Adding a provision to the Text and Data Mining Exemption allowing media corpora to be shared does not just make existing research easier—in many cases, it would make research possible that could not even be considered without it.

Not being able to share a media corpus across institutions and studies holds back another critical area of investigation as well: the ability for different researchers to verify and build upon each other's work. Validation of results is common practice among STEM researchers who use quantitative methods, but critical scholars in the humanities have historically based their work on more qualitative techniques. The algorithmic analysis supported by the Text and Data Mining Exemption is quantitative in nature, though, and demands that humanists adopt practices like independent validation of results. Given the difficulty in building a large corpus even for original resource, it is not realistic to expect independent researchers to recreate a corpus built at another institution just to prove or disprove somebody else's work, or even to extend that work in unique ways by pursuing related research questions within a common corpus. It is not just individual projects that are held back by not being able to share a media corpus with other researchers. The evolving methodology of entire disciplines is being held back by the requirement to restrict a prepared corpus to its original research group.

For example, the corpus we are building for Deep Screens is entirely composed of commercial media protected by copyright. The three dimensional pose analysis methods we are using, however, are so new that they are being updated or entirely replaced on a weekly basis. It is quite likely that, between the time when we finish our analysis and publish our results, our methods will be obsoleted by new technology and the first thing our readers will want to do is rerun the analysis using new models to produce more accurate results or examine a related research question that could not be addressed using current inference models. This is not a failing of the original methods, but instead an indication of how quickly the technology those methods are based on is moving.

The Media Ecology Project has been working with archives for more than a decade and, counterintuitively, research access to moving image collections has become more difficult in that time rather than easier. Digital Rights Management software devised to protect content from piracy is

becoming more widely adopted even for content that has far more cultural than commercial value. Many textual, image, and video assets that are available for academic use are only offered in paywalled garden environments like Artstor. While this type of mediated access is better than no access, it is typically not appropriate for computational analysis of large corpora because the materials are restricted to access through designated portals with no capacity to work at quantitative scale. To work around these restrictions we have long sought to work directly with archives to produce a scholarly tier of access to these materials based on federated logins tied to institutional accounts verifying their academic status. One of the main sticking points archives have mentioned in these discussions, though, is that they have materials that are either fully protected by copyright or have ambiguous orphaned status. They may be comfortable with sharing such materials in streaming format, but streaming is of little use for most machine learning applications due to their scale and speed. In my ideal world, an Exemption that allowed for sharing of corpora would also provide archives with the confidence needed to begin working on such a federated access system without fear of legal action that would prove ruinous to their work.

The clock is ticking on building such a system. As the entertainment industry transitions away from physical media and toward streaming, the current DMCA 1201 Exemption is not going to age well. New commercial programs will see fewer releases on physical media, and the physical media we have today is already falling to technical obsolescence, delaminating discs, and disintegrating film. The time to build shared corpora based on 1201 is now, while physical media and drives are still viable data sources.

Sincerely,

A handwritten signature in dark ink, appearing to read 'John Bell' in a cursive, stylized script.

John Bell
Program Director, Data Experiences and Visualizations Studio
Lecturer, Film and Media Studies
Dartmouth College

Appendix E
Letter from Joel Burges and Emily Sherwood

December 8, 2023

To the Register of Copyright:

We are Joel Burges, Director of Mediate and Associate Professor of English and Visual & Cultural Studies, and Emily Sherwood, Project Manager for Mediate and the Director of Digital Scholarship, at the University of Rochester (UR). We write in support of an expansion to DMCA § 1201 to enable text and data mining (TDM) of in-copyright materials, especially moving images on DVDs, BluRays, and those only available on streaming platforms. As it exists, DMCA § 1201 is an impediment to more expansive teaching and research with Mediate at and beyond UR. Both would benefit immensely from ending the currently restricted practice of corpora sharing with researchers who otherwise comply with the exemption but who are at a different institution and are outside of direct collaboration. Due to this restriction, researchers at distinct institutions are burdened with recreating the underlying data—a corpus of film and television about the history of the close-up, for instance—over and over again to teach with and/or pursue new inquiries about it. This burden not only slows down technologically innovative and often transdisciplinary research, but also contributes to unsustainable conditions for imaginative pedagogy and scholarly inquiry in, to name the fields in which Mediate has played a role at UR, the digital humanities, film and media studies, visual studies, cultural studies, musicology, and linguistics.

Mediate is a time-based digital annotation tool for audiovisual material that supports both individual and collaborative teaching and research. Students, faculty, and scholars use the tool to develop media literacies of various kinds and pose manifold research questions about audiovisual material often protected by copyright, especially film and television. At UR, it has been used by hundreds of students in multiple classes, engendering a humanistic network of mutual mentoring among students, faculty, librarians, and staff. Given its impact, it has led two teaching awards for Burges from UR in 2022: the Judith Kerman Faculty Teaching and Mentoring Award in Culture and Technology and the Edward Peck Curtis Award for Excellence in Undergraduate Teaching. UR has also supported Mediate with an Educational IT Innovation Grant and a Student Course Development Project Grant, and Mediate is currently central to two externally funded projects: the Rochester Digital Annotation Project, funded by a Digital Justice Seed Grant from the American Council of Learned Societies, and “A Digital History of the Close-Up in Narrative Film and Television,” funded by the Public Knowledge Program of the Mellon Foundation. Despite the internal and external enthusiasm for Mediate, however, restrictions on sharing corpora inhibit further technological and scholarly development that institutions and grant foundations are interested in supporting as part of a more open market of free-flowing ideas in the United States. Lifting the restriction on corpora sharing in addition to reaffirming the current exemptions would go a long way towards shoring up this market of ideas in U.S. higher education and contributing to the global competitiveness of the nation.

As we have shown in two articles, “Audiovisualities out of Annotation: Three Case Studies in Teaching Digital Annotation with Mediate” (Burges et al 2021, *Digital Humanities Quarterly*) and “Collective Reading: Shot Analysis and Data Visualization in the Digital Humanities” (Burges et al 2016, *Cinema Journal Teaching Dossier*), Mediate is anchored in a practice that we now call *close viewing*. In contrast to the practice of *distant viewing*, which emphasizes the at-scale computational analysis of digital images through machine learning (Arnold and Tilton, *Distant Viewing: Computational Exploration of Digital Images*, 2023), close viewing in Mediate involves human users manually annotating audiovisual materials to produce data about media such as film and television. While individuals can use Mediate for research projects, close viewing in Mediate is collaborative by design in ways that call out for it to be used across institutions so students and researchers can analyze media in a community of inquiry investigating shared corpora.



Two recent cross-institutional projects demonstrate both Mediate's potential for direct collaboration and the challenges that restrictions on sharing corpora engender for such collaborations in Mediate. Both projects were direct collaborations with Kinolab, which is led by Allison Cooper at Bowdoin College. Supported by the Kerman Award and the Student Course Development Grant listed above, the first was a fall 2022 teaching project in which we co-taught a contemporary film history course using Mediate for synchronous and asynchronous collaboration. Working together across campuses, Bowdoin and UR students annotated groups of films about which they had a research question, uploading the films onto the platform and generating data about the formal elements of their film language. For example, one group explored the road movie and another annotated films about immigration to the US. Ultimately, all groups published clips from these films to Kinolab. A parallel project occurred in Burges's fall 2022 course "The Poetics of Television," though this class was not cross-institutional. Supported by the Mellon grant listed above, the second cross-institutional project is ongoing. Again in partnership with Kinolab, we are investigating the significance of digital annotation and data analysis for understanding the history of LGBTQ+ and BIPOC representation vis-a-vis the close-up in narrative film and television from 1950-2000. To date, this Mediate-Kinolab collaboration has yielded over 40,000 time-based data points about the close-up in Mediate and over 700 clips—the underlying corpus for the project—related to the racial and sexual history of the close-up. These clips have been published to Kinolab based on a Diversity in Media Representation annotation schema developed for this project and we are currently analyzing the 40,000 data points through visualization and other modes to design the next phase of research for this project.

In the cases of both teaching and researching across institutions, we have run into overlapping uncertainties, redundancies, and obstacles—some of them costly—because of the restriction of corpora sharing:

1. In the case of teaching jointly across campuses, Cooper and Burges were careful to be the ones breaking the encryption on DVDs and Blu-rays before sharing digitized copies for annotation in Mediate with students. We also communicated the legal constraints which bound not only us, but also our students to the latter. Nonetheless, this created a significant amount of extra labor for Cooper and Burges because of uncertainty about whether students could also break the encryption as part of a project that would lead to publication at Kinolab unless they were paid research assistants for Kinolab and Mediate. We increasingly assume students are covered by the same restrictions and exemptions, but this is a persistent ambiguity. It also impacted Burges in the aforementioned "The Poetics of Television."
2. In the case of both projects above, especially the Mellon-funded close-up grant, both UR and Bowdoin have had to invest in physical copies of all the films and television series annotated, store these copies, and devote significant time, technology, and wages to their digitization at both institutions. In other words, the grant has required massive redundancy in both labor and purchasing costs, representing money and time that could have been spent on more data mining of the film and television being annotated and more sustained exploration of the best ways to systematize and visualize that data to share with a wider community of inquiry across fields such as the digital humanities, media studies, visual studies, data science, and information science. There is, moreover, no guarantee that the budgets to which we currently have access will continue, so too much of this kind of redundancy is intellectually and institutionally inefficient.
3. During the process of close viewing for the Mellon-funded project, we realized that our research would benefit from a comparison with a distant viewing analysis such as the one currently being led by David Bamman at UC Berkeley on a related corpus of film and television and/or ones historically led by Taylor Arnold and



Lauren Tilton at the Distant Viewing Lab at the University of Richmond. Without access to cutting-edge methods such as distant viewing, we must rely on speculation rather than systematic collaboration to produce comparative analyses that would advance multiple fields of study. Ideally, we would like to do so in the next three years, but the legal and consequently financial impediments to doing so may be insurmountable at present.

4. Similarly, researchers in distant viewing might benefit from training their algorithms based on close viewing data sets. While we could share the data generated on the corpus, the researchers would be forced to recreate our corpus before they were able to benefit from the shared secondary data to improve their algorithms. Further, part of the interest of our corpus is that our data is time-based: a user generates annotations that are keyed to specific moments unfolding in a clip in the interface. While we can extract that time-based data for interpretation and visualization, we cannot share the annotated clips easily with other researchers and institutions interested in distant or close viewing. Without the ability to share our underlying data in annotated form with others, other researchers and institutions cannot benefit from the years of time and capital that have gone into the development and use of Mediate. Again, we would ideally like to pursue such projects in the next three years, but the legal and consequently financial impediments to doing so may be insurmountable at present.
5. A related obstacle is the current restriction on digitizing streaming materials. This is one reason the close-up project was forced to stop in 2000, since after that date, streaming became more and more the norm for film and television. Without access to this material, and without the ability to share it across institutions were it to become accessible, achieving a representative history, for instance, of the close up in narrative film and television beyond 2000 is impossible. This strands the Kinolab-Mediate collaboration going forward within an admittedly rich fifty-year period of media history—1950-2000—but unable to explore how the computerization of culture has impacted that history through TDM rooted in close viewing in Mediate.
6. Effective January 25, 2023, the NIH mandated that scientific data be made publicly available to facilitate the validation and replication of research findings. While, currently, data sets funded by the government are only required for NIH funding, many humanities and social science researchers are starting to discuss the potential implications and impact of similar mandates for other federal funding sources, such as the NEH, IMLS, or NSF. If these data mandates are expanded without an expansion of current copyright exemptions, humanities research will be further stymied. The current limitations mean we cannot make our corpora publicly available, but we will also be required to do so if we want to receive federal funding. That means humanities researchers will not be able to apply for federal funding if the NIH data mandates are expanded to other sectors because to replicate our research findings, researchers would need access to our corpora, which often includes material under copyright. Moreover, some aspects of projects such as ours might be fundable through the NIH, making this question not only a future but also a present day one.

The six points enumerated above constitute some of the reasons that not only already-existing exemptions should be renewed, but also the restrictions on corpora sharing urgently needs to be lifted for non-consumptive uses of the kinds that we pursue with Mediate. Scaling out from the specifics of each point, maintaining and expanding to exemptions DMCA § 1201 to TDM of in-copyright materials will have some wide-ranging benefits for teaching and research in the US that will allow our country's higher education institutions to remain the global leaders they are. The largest benefit,



especially of expanding the ability to share corpora, would be to allow researchers to pursue sustainable advances in the development of technological tools for humanistic inquiry and media literacy that are, at present, hampered by the requirement to endlessly replicate underlying data. This requirement obstructs new questions, new knowledge, and new tools emerging in the market of ideas.

Sincerely,

A handwritten signature in black ink, appearing to read 'Joel Burges'.

Joel Burges
Associate Professor
English | Digital Media Studies | Film and Media Studies | Visual and Cultural Studies
Director, Mediate
University of Rochester

A handwritten signature in black ink, appearing to read 'Emily Sherwood'.

Emily Sherwood
Director of Digital Scholarship and Studio X
Project Manager, Mediate
University of Rochester

Appendix F
Letter from Brandon Butler

December 13, 2023

To the Register of Copyrights:

I am the Director of Intellectual Property and Licensing at the University of Virginia library. I also advise creators, publishers, and memory institutions through my law firm, Jaszi Butler PLLC. I am writing today in my personal capacity in support of Authors Alliance's expansion of the 37 C.F.R. § 201.40(b)(4) and (b)(5)—the text and data mining (TDM) research exemptions to 17 U.S.C. § 1201's prohibition against circumvention of technological measures.

In my role as the Director of IP and Licensing, I serve as an expert consultant to UVA librarians, to groups and individuals within the University, and to national and international efforts focused on issues relevant to research and teaching at UVA. I provide guidance and expertise to the Library as it develops plans and strategies to address the challenges that it faces as a leading university research library. I have also written on the topic of TDM research and represented library and disability rights groups in administrative proceedings and in litigation as amici in landmark cases establishing the right to digitize and perform computational analysis on millions of in-copyright books in library collections.

Granting the current exemption has sparked great interest in using TDM methods in humanities research and enabled many projects. However, the inability to share corpora with unaffiliated researchers outside direct collaboration has created barriers for researchers to practically use the current exemption. In a forthcoming article I co-authored with Pat Aufderheide and Kimberly Anastacio, we conducted in-depth interviews with TDM researchers and outlined a range of obstacles encountered while conducting TDM research.¹ Due to these obstacles, researchers have been forced to change research design, delay research, and abandon research, and have been hampered in their ability to collaborate effectively.

Particularly, we observed that researchers widely believe that the challenges of conducting TDM research starts with copyright law. In a survey to 262 responses between March 2021 and November 2022, one fourth of the answers (24%, cited 73 times) attributed the problem to copyright law. The biggest single problem was with limits on the sharing of data (17%, cited 33 times), though there was also friction in other areas. Among specific problems mentioned, researchers cited:

- Inability to share raw data (which limits reproducibility);
- Unclear capacities to move the data should the researcher switch institutions;
- Digitizing print books and preparing them for research is expensive and not time-effective.

¹ Patricia Aufderheide, Brandon Butler, and Kimberly Anastacio, *The Chilling Effects of Obstacles to Accessing, Using, and Sharing In-copyright Data for Quantitative Research*, 59 *Information & Culture* (forthcoming 2024).

These concerns about the ability to share data across institutions have chilled valuable research and negatively affected research that is conducted. Out of 140 respondents and 155 answers, most of the answers said either that these problems had in some way impaired their research (43% total), or that they were unsure whether it had (17%). The most commonly reported problem was having to change the design of the research (23%). Fourteen percent of the answers reported avoiding taking on a project. Six percent of the answers mentioned having to abandon a project. Researchers reported that: “I no longer work on materials where copyright could be an issue” and “I pretty much shifted my whole area of research to avoid worrying about post-1922 issues.” One respondent was clear and direct, “I have stopped research on projects where copyright is confusing or otherwise impedes sharing of data.”

Additionally, because of the high costs associated with compiling corpora and the limitations on sharing, researchers tend to resort to suboptimal research design. One respondent noted that “[i]nstead of the TDM I had planned to do with a larger run of issues digitized by the vendor, I had to use a smaller range of issues that were ones the library originally contributed to the project.” Also, another respondent claimed that “instead of using the full content that was collected, I had to choose a very small part.”

These obstacles to effectively conducting TDM research negatively affect research quality. In particular, because of the limitation on corpora sharing, researchers are limited in their ability to test out different methods on the same corpus, and scholars without affiliation to the original project are unable to test published results. Having a shared corpus where researchers can test different methods and ask different questions would thus not only increase the efficiency of TDM research, but also reduce authority bias. One researcher highlighted that “[I] had to use older data of lower quality because newer material is copyrighted. This decreases the quality of research.” Further, instead of encouraging the use of standardized corpora such as Books2 and 3 (which contain in-copyright materials collected from self-described pirate sites), the proposed expansion would permit researchers to procure in-copyright works in shared corpora created lawfully and intended for scholarly fair use. Thus, the proposed expansion benefits both the researchers and the rightsholders.

The pedagogical value of the DMCA 1201 exemption is also hindered by the current exemption’s limitation on corpora sharing. Perceived copyright barriers to compiling corpora for specific projects have caused researchers to avoid using in-copyrighted works in their projects; this includes teaching. Almost all of the scholars we spoke with about TDM research are also teachers, and they worried that the inability to work with more contemporary materials in digital humanities courses was making it more difficult to cultivate students’ interest in these courses, and even in the humanities more generally. As one researcher observed, “[s]tudents would be so much more engaged if we could use more contemporary literature.” While general copyright fears were a factor, the prospect of assembling corpora from scratch for each new course, rather than building on the work of colleagues in the field, was another impediment.

Importantly, the inability to share corpora with unaffiliated researchers outside direct collaboration is particularly damaging to digital humanities research because digital humanities researchers generally have limited institutional support. 128 respondents gave 236 answers for where they get helpful information in making their TDM decisions. Nearly a third of the answers (29%, 68 times) were “colleagues/peers” and 18% (43 times) were “myself.” Thirteen percent (31 times) selected “librarian,” 11% (26 times) “lawyer,” and also 11% (26 times) “superior/boss.” “Friends” were the least cited source (7%, 17 times). Among the 11% answers (25 times) that selected “other,” respondents mentioned their department staff, university law experts, and the Internet through online communities. This pattern of a lack of professional support makes it even more important to expand the current exemption to remove barriers for researchers.

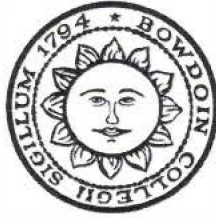
The proposed expansion would remove some of the barriers researchers face and enable TDM researchers to effectively carry out their projects, which have been determined to be fair use. These research projects are crucial to our understanding of our culture and history. I respectfully ask that you grant this expansion to the exemption.

Sincerely,

A handwritten signature in cursive script, appearing to read "Brandon Butler".

Brandon Butler
University of Virginia

Appendix G
Letter from Allison Cooper



22 November 2023

I am writing this letter in support of Authors Alliance's petition to the Copyright Office for an exemption to §1201. I am an associate professor of Romance languages and literatures and cinema studies at Bowdoin College, where I direct Kinolab, a digital humanities laboratory for the analysis of narrative film and television. Kinolab's research activities focus on how film language functions as a system of communication. Our work has been supported by the Mellon Foundation's Public Knowledge program and the Mozilla Foundation's Responsible Computing Challenge. Developing an at-scale, representative digital collection of narrative film and television clips is foundational to our work, since moving image clips, along with metadata added by student curators and researchers, form the basic data for our research.¹ My observations below reflect my own experience as director of a DH laboratory and a researcher with expertise in the analysis of narrative film and television, they are not made on behalf of any organization.

As principal investigator of "A Digital History of the Close-Up in Narrative Film and Television," a Mellon-funded, cross-institution collaboration with the University of Rochester, I am currently overseeing the creation and analysis of an at-scale dataset of moving image clips annotated to make the close-up and its representation of race, ethnicity, gender, and sexuality discoverable. Researchers at the University of Rochester's complementary Mediate project have been a vital part of my project team, contributing expertise in the curation of narrative film and television, data visualization, and the development of annotation schema related to diversity in media representation. Our research, which has thus far produced approximately 700 clips tagged for their distinctive representation of American identity via the close-up, depends upon the 2021 expansion of exemptions to the DMCA to permit the circumvention of TPM for data mining in service of scholarly research or teaching. Similarly, our partnership with the University of Rochester to form a 21-person team of faculty, information scientists, and student curators across our two institutions has been made possible by the same expansion of DMCA exemptions to allow corpus sharing with other researchers for the purposes of collaboration or research replication.

Unlike many digital humanities projects that work with digitized texts or preexisting datasets, at Kinolab we build our own corpora of annotated film and television clips

¹ See, for example: Allison Cooper (2021), "Collaborative Film Language Analysis in the Networked Information Era," *The Italianist Film Issue*, vol. 41, no. 2, pp. 175-184; Allison Cooper, Fernando Nascimento, and David Francis (2021), "Exploring Film Language with a Digital Analysis Tool: the Case of Kinolab," *Digital Humanities Quarterly*, vol. 15, no. 1; and Allison Cooper (2019), "Kinolab: A Digital Humanities Project for the Collection and Analysis of Film," *Italian Culture*, vol. 37, no. 2, pp. 137-143.

according to our research objectives. Of our \$100,000 grant from the Mellon Foundation for this project, which funded parallel curatorial campaigns at Bowdoin and the University of Rochester, only \$6,500 was spent on DVD acquisition and transcoding (breaking TPM) activities: the majority of the grant supported research activities on the part of project faculty, staff, and student curators to build a deliberate corpus of close-up clips. This process began with a wide-ranging review of existing scholarship and writing on the representation of race, ethnicity, gender, and sexuality in American film and television that included specialized texts like *The Celluloid Closet: Homosexuality in the Movies* (Vito Russo, 1987) and *Race in American Television: Voices and Visions That Shaped a Nation* (David J. Leonard and Stephanie Troutman Robbins, ed., 2021), as well as articles in the popular press like *Slate Magazine's* "The Black Film Canon: The 50 Greatest Movies by Black Directors" (Aisha Harris and Dan Kois, 2016) and trade books like *Colorization: One Hundred Years of Black Films in a White World* (Wil Haygood, 2021). Our research has focused on the second half of the twentieth century, in part because much twenty-first-century narrative film and television is relatively inaccessible to us since it is hosted exclusively on streaming platforms and unavailable on DVDs. Our work thus far has yielded curatorial lists of nearly 260 key films and 180 key television episodes. Joel Burges, my co-PI at the University of Rochester, and I each spent the equivalent of three weeks building our master curatorial lists for film and television, drawing on our own expertise as scholars of film and television and pre-existing research like the sources named above. Decisions about which films and TV episodes to prioritize for the project were then made in weekly curatorial meetings that included myself, Burges, and seven undergraduate and graduate student curators across our two institutions. Kinolab's largest expenditure for this kind of project is on training student curators to closely watch media and reliably identify and annotate aspects of film language like the close up, along with complex forms of identity representation, from queer coding and blackface in mid twentieth century film and television to the nonbinary or transgender characters that become increasingly visible onscreen in the late twentieth century. Our student curators transcode DVDs once they have been acquired, watch them carefully, take notes that are shared with the entire research team, pull clips of scenes that highlight the close-up's foregrounding of identity, annotate them with tags from specialized annotation schema developed by the project's lead researchers, and, finally, present their work for discussion and peer review in the aforementioned weekly curatorial meetings before adding brief text descriptions to each clip and uploading it to our database. In 2023, our student curators across both institutions spent a total of nearly 600 hours engaged in this work. Creating an annotation schema to make aspects of identity discoverable in moving image clips has also constituted a significant expenditure of time and money for the project, since my co-PI and I could find no pre-existing schema tailored to the representation of identity onscreen.

As the description of our current work suggests, the corpus that Kinolab is building is customized to our ongoing research on the close-up. Yet it would likely be valuable to other researchers if the existing TDM exemption were expanded to allow for corpora sharing beyond "collaboration or replication of the research." Our close-up corpus, with its emphasis on racial, ethnic, sexual, and gender diversity could, for example, serve as an especially useful set of training data to counter the kind of sample bias that has been well documented in machine learning. My understanding of the existing TDM exemption is that

sharing our work with information scientists or machine learning specialists beyond Bowdoin or the University of Rochester for a use other than our own research is disallowed. This is discouraging since Kinolab would benefit from an exploration of the ways in which AI might enable further research on film language. Given Bowdoin's small size and limited resources, we would likely only be able to do so in partnership with experts at larger, better resourced institutions, like UC Berkeley's School of Information, where David Baman's research on using machine learning for image recognition in film and television complements Kinolab's work.

Similarly, the existing TDM exemption is an impediment to Kinolab's objective of developing its relatively simple film language data model into a film language ontology, a complex data model with the potential to represent more than just the visible and/or audible technical practices and aesthetic techniques in narrative film and media. A film language ontology would be an expansive and detailed representation of our field's collective knowledge about film language that could represent broad concepts such as cinematic space and cinematic time, relationships such as that of the sequence shot to the long take, the affective attributes of the close up, and more. There is a project underway in Germany to develop just such an ontology, which is a collaboration between the Film Studies Department at Freie Universität Berlin and the Computational Sciences Department the Hasso Plattner Institute in Potsdam, called the Ada Filmontology. Kinolab's platform would provide an ideal test bed for the Ada Filmontology, but such a collaboration would likely require making our corpora available to its researchers – violating the parameters of the existing TDM exemption. If the Library of Congress expands the TDM exemption to allow for broader corpora sharing, Kinolab will pursue partnerships with researchers like Baman at other institutions of higher learning to explore the development of ethical AI-based tools for searching moving images, just as we will pursue the development of a film language ontology in collaboration with existing projects like the Ada Filmontology. Kinolab's research is simultaneously dependent upon and limited by existing exemptions to the DMCA. Our work could not legally take place without the exemptions permitting the circumvention of TPM for the purpose of criticism, comment, teaching, or scholarship and, more recently, the exemption permitting limited cross-institution collaborations for the purpose of TDM. Previous DMCA exemptions and expansions in 2018 and 2021 have directly enabled our research by minimizing legal barriers that otherwise would have prevented our work from keeping pace with new technologies and research methodologies. On behalf of Kinolab, I am grateful to the Library of Congress for its ongoing responsiveness to the needs of motion picture scholars and students and hope that the information I have shared in this letter will support Authors Alliance's petition for further expansion of existing exemptions to increase our ability to achieve the goals described above.

Sincerely,



Allison A. Cooper

Associate Professor of Romance Languages and Literatures and Cinema Studies

Appendix H
Letter from Hoyt Long

October 23, 2023

To the Register of Copyrights:

I am Professor of Japanese literature and digital studies at the University of Chicago. I'm also the Chair of the Department of East Asian Languages and Civilizations and Interim Director of the Japanese Language Program, as well as co-director of the Textual Optics Lab. I write today in my personal capacity in support of expanding the exemption allowing circumvention of technological protection measures to facilitate text and data mining ("TDM"). I wrote last cycle in support of the exemption when it was first proposed. For over a decade now, I have been involved in text and data mining research applied to literature in both English and Japanese. At the Textual Optics Lab, we use qualitative and computational methods to build large-scale collections of literary texts and to achieve scalable reading of textual works. These techniques allow observations to be made about large literary corpora while also facilitating close examination of details within a single text.¹ As a researcher in this space, I am thus deeply invested in efforts that make it easier for researchers to share resources and collaborate on efforts to enhance the field of text and data mining for cultural material.

In my own research, I apply computational methods to the study of literature and culture across different languages and written media. More specifically, I have used these methods to "scale up" more familiar humanistic approaches and investigate questions of how literary genres evolve, how literary style circulates within and across linguistic contexts, how patterns of racial discourse in society filter down into literary expression, and how online platforms are creating new spaces for the production and consumption of stories. I have authored and coauthored many essays that introduce computational methods like network analysis, natural language processing, and machine learning to the study of literary history in Japan, the US, and other parts of the world.²

¹ In addition, I serve on the board of the Journal of Cultural Analytics and have been involved with several large-scale and multi-institutional digital projects. This includes NovelTM, a multi-million dollar research initiative funded by the Social Science and Humanities Research Council of Canada; the ACLS funded History of Black Writing project at the University of Kansas, which aims to digitize a collection of over 1,000 African-American novels; the Mellon funded Scholar-Curated Worksets for Analysis, Reuse & Dissemination project; and the Japanese Text Mining initiative, a series of workshops introducing text mining methods to Japanese studies scholars.

² Some of my works include: Hoyt Long, Richard Jean So, Kaitlyn Todd, *#COVID, Crisis, and the Search for Story in the Platform Age*, 49 *Critical Inquiry* 530-56 (2023) <https://doi.org/10.1086/725059>; Richard

I have also published a monograph, *The Values in Numbers: Reading Japanese Literature in a Global Information Age* (Columbia University Press, 2021), which brings debates around computational literary history to the study of Japan. The book guides readers through increasingly complex techniques while making novel arguments about topics of fundamental concern, including the role of quantitative thinking in Japanese literary criticism; the canonization of modern literature in print and digital media; the rise of psychological fiction as a genre; the transnational circulation of modernist forms; and discourses of race under empire. Overall, the book models how computational methods can be applied outside English-language contexts and to languages written in non-Latin scripts, but also how these methods augment our understanding of the literary past.

All of these research projects have benefitted in some way from the current TDM exemption. The exemption has made it possible to address my research questions to the full range of literary and cultural output of the 20th century, and thus to demonstrate to scholars in the field how beneficial new computational methods can be for understanding the recent literary past. Without it, it would be impossible to advance computational literary or cultural studies in meaningful ways, potentially leading to the further marginalization of the humanistic sciences as other fields capitalize on data-centric methods. To give a recent example from my own research, the exemption allowed me to pursue experiments in the application of neural machine translation models to the analysis of newly digitized collections of contemporary literature in several languages.³ Such experiments will only become more vital as the research community tries to assess the capabilities of the newest machine learning models and their potential impact on literary and other forms of creativity. And given the resources and skillsets required to train and fine-tune such models, let alone to build datasets from scratch, collaboration will be essential to advancing such research. Expanding the TDM exemption to reflect this reality will further bolster innovative work in this area.

Of course, the bulk of TDM research is still conducted out using simpler computational techniques and machine learning models (e.g., topic-modeling, sentiment analysis, word embeddings). Even in such cases, however, the inability to share text collections except for direct collaboration and/or replication has prohibited TDM research agendas from reaching their full potential. When one set of researchers builds a large-scale text collection at their home institution, it is typically built to pursue a limited set of research questions. A single research team simply does not have the time

Jean So, Hoyt Long, and Yuancheng Zhu, *Race, Writing, and Computation: Racial Difference and the US Novel, 1880-2000*, 1 J. of Cultural Analytics (Jan. 12, 2019) <https://culturalanalytics.org/article/11057>; Hoyt Long & Richard Jean So, *Turbulent Flow: A Computational Model of World Literature*, 77 Mod. Language Q. 345 (2016) <https://doi.org/10.1215/00267929-3570656>; Hoyt Long & Richard Jean So, *Literary Pattern Recognition: Modernism between Close Reading and Machine Learning*, 42 Critical Inquiry 235 (2016) <https://doi.org/10.1086/684353>; and Richard Jean So & Hoyt Long, *Network Analysis and the Sociology of Modernism*, 40 boundary 2 147 (2013) <https://doi.org/10.1215/01903659-2151839>.

³ Hoyt Long, *Learning to Live with Machine Translation*, 53.4/54.1 New Literary History (Autumn 2022/Winter 2023) <https://muse.jhu.edu/article/898327>.

to pursue the full range of questions that a several-thousand volume collection of novels might open up. A significant amount of time and human labor are poured into creating these collections, usually at better-endowed private institutions. And yet the inability to share them with researchers who might have different research questions to explore leads to needless duplication of effort at other institutions or, as is more typical, a siloing effect that keeps the collections locked up at their home institutions. For example, here at UChicago, where our lab has developed a large collection of general US fiction and a corpus of novels by African-American writers, we constantly receive requests from other university faculty and graduate students who are hoping to pursue their own research projects. This includes, for instance, wanting to develop models for extracting characters from text and their narrative framing; measuring narrative coherence and its degree of correlation to reader preferences; exploring representations of climate; constructing sentiment and emotion arcs; studying how AAE is expressed in fiction by African-American writers; and investigating the construction of metaphorical language in the same body of fiction. These are not projects that members of our lab have the expertise to pursue or collaborate on, and yet there is no question that they are worthy of being pursued.

While the exemption has been immensely valuable as it currently exists, it can sometimes conflict with the actual practice of academic research in counterproductive ways. As should be clear from the example above, there is a demonstrated need for the requested expansion of the exemption. Other researchers have requested to work on corpora produced through the Textual Optics Lab yet, as a single research team, we do not have the capacity to directly collaborate with every researcher who wishes to work with these corpora. The field of digital humanities is diverse, and researchers will approach a corpus with a variety of unique perspectives and techniques that all contribute to the understanding of that corpus and the texts within it, as can be seen from the sample of requests above. Permitting a corpus to only be used with a single researcher's perspective leads to needlessly duplicated effort in preparing the same corpus and to growing disparities between institutions. An expansion of the exemption would go a long way to closing this equity gap and ensuring that TDM research can productively be carried out by researchers of diverse backgrounds and perspectives. It would instantly make possible the wide array of projects for which we have received requests, and which have come from researchers across North America and Europe.

Moreover, it would allow us to enter into new collaborations that are currently prohibited by the inability to share collections across institutions. For instance, I would like to be able to continue experimenting with neural-machine translation models and LLMs. But this means having to work with collaborators elsewhere who have an interest in literature and/or translation, or who have the necessary expertise to train, operate, and fine-tune the open-source versions of these models. This research will depend on being able to use our current digitized collections either as input or as a means to assess model output (e.g., comparing LLM produced literature against human produced fiction). None of this research can happen, of course, unless we are able to more broadly share the collections with collaborators at other institutions.

Last cycle, the Copyright Office's granting of an exemption sparked a flurry of valuable research at the intersection of humanistic inquiry and data science. I ask that you help ensure the long-term impact of these gains by granting a further expansion of the exemption, and thus contributing to a new round of research activity that can help to illuminate the world's literary and cultural heritage in response to new digital technologies and research methodologies.

Sincerely,

A handwritten signature in black ink, appearing to read 'Hoyt Long', with a stylized, flowing script.

Hoyt Long

Professor of Japanese Literature and Digital Studies
Chair, Department of East Asian Languages and Civilizations
Interim Director, Japanese Language Program
University of Chicago

Appendix I
Letter from Matthew Sag

To the Register of Copyrights:

Via electronic submission

Dear Register Perlmutter,

I am a Professor of Law in Artificial Intelligence, Machine Learning, and Data Science at Emory University, School of Law. I am also a member of the American Law Institute and the HathiTrust Research Center Advisory Board.

I write to you in my individual capacity in support of the petition to expand the current academic text and data mining (“TDM”) exemptions.¹

I am an expert on the legal issues relating to TDM research, particularly in relation to copyright law. My research in this area has been published in *Nature*, *Science*, the *Journal of the Copyright Society*, the *Northwestern Law Review*, and the *Berkeley Journal of Law and Technology*.² I was the lead author of the amicus briefs filed on behalf of “Digital Humanities and Legal Scholars” in the *HathiTrust* and *Google Books* cases that ultimately set the current favorable fair use precedent for text data mining.³ I have been a member of the HathiTrust Research Center Advisory Board Since 2016⁴ and I was one of the project team members for the Building Legal Literacies for Text Data Mining Institute (“Building LLTDM”), funded by the National Endowment for the Humanities. In July 2023 I testified to before the U.S. Senate Committee on the Judiciary Subcommittee on Intellectual Property about Copyright and AI.

¹ 37 C.F.R. § 201.40(b)(4)&(5).

² Matthew Sag, Copyright and Copy-Reliant Technology, 103 Nw. U. L. Rev. 1607 (2009); Matthew Sag, Orphan Works as Grist for the Data Mill, 27 Berkley Tech. L.J. 1503 (2012); Matthew Jockers, Matthew Sag & Jason Schultz, Digital Archives: Don’t Let Copyright Block Data Mining, 490 Nature 29-30 (Oct. 4, 2012); Matthew Sag, The New Legal Landscape for Text Mining and Machine Learning, 66 Journal of the Copyright Society of the U.S.A. 291–367 (2019); Sean M. Flynn, Matthew Sag, et al., and Jorge L. Contreras, *Legal reform to enhance global text and data mining research*, 378 SCIENCE 6623 (1 Dec 2022), 951-953 (<https://www.science.org/doi/10.1126/science.add6124>); *Copyright Safety for Generative AI*, Houston Law Review (forthcoming); *Fairness and Fair Use in Generative AI*, Fordham Law Review (forthcoming).

³ Brief of Digital Humanities and Law Scholars as Amici Curiae in Support of Defendants-Appellees and Affirmance, Brief of Digital Humanities and Law Scholars as Amici Curiae in Partial Support of Defendants’ Motion for Summary Judgment or in the Alternative Summary Adjudication, Authors Guild v. Google, Inc., 954 F. Supp. 2d 282 (S.D.N.Y. 2013) (No. 1:05-cv-08136), aff’d, 804 F.3d 202 (2d Cir. 2015) (https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2102542). Authors Guild v. HathiTrust, 755 F.3d 87 (2d Cir. 2014) (No. 12-04547), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2274832;

⁴ HathiTrust is a not-for-profit collaborative of academic and research libraries that maintains a corpus of over 17 million digitized items. The HathiTrust Research Center (HTRC) enables computational analysis of the HathiTrust corpus. The HTRC develops cutting-edge software tools and cyberinfrastructure to enable advanced computational access to the growing digital record of human knowledge.

I am well-versed with the objectives, methodologies, and organizational and legal challenges relating to TDM research from my work with the HathiTrust and my experience in the Building LLTDM Institute. I also have firsthand experience in academic text mining in my empirical work analyzing the transcripts of U.S. Supreme Court oral arguments. In this research I have used TDM techniques to draw empirical conclusions about litigation and judicial behavior.⁵ I also have significant experience in relation to the application of the fair use doctrine in analogous contexts: I was part of the legal advisory committee for the *Code of Best Practices in Fair Use of Copyrighted Materials for the Visual Arts*, and for the *Code of Best Practices in Fair Use Software Preservation*.

TDM is a non-expressive use that is strongly favored under the first fair use factor and is undoubtably fair use in the academic research context.

Text data mining is an umbrella term referring to computational processes for applying structure to unstructured electronic texts and employing statistical methods to discover new information and reveal patterns in the processed data. In other words, text data mining refers to any process using computers that creates metadata derived from something that was not initially conceived of as data. The text data mining relevant to this petition is used to produce statistics and facts about copyrightable works. These statistics and facts are not same as, or even substantially similar to, the original expression in the underlying works, but in combination they are interesting and useful for generating insights about the original expression.

United States courts have consistently held that technical acts of copying which do not communicate an author's original expression to a new audience are fair use. TDM is just one example of this broader category of non-expressive uses. The case law indicates that even though these "non-expressive uses" involved significant amounts of copying, they did not interfere with the interest in original expression that copyright is designed to protect.⁶ Deriving uncopyrightable information and insights from copyrighted expression is not just transformative, it is highly transformative.⁷

⁵ Matthew Sag, Predicting Fair Use, 73 Ohio St. L.J. 47 (2012); Tonja Jacobi & Matthew Sag, The New Oral Argument: Justices as Advocates, 94 Notre Dame L. Rev. 1161 (2019); Tonja Jacobi & Matthew Sag, Taking Laughter Seriously at the Supreme Court, 72 Vand. L. Rev. 1423–1496 (2019).

⁶ The terminology of "non-expressive use" originates with Matthew Sag, *Copyright and Copy-Reliant Technology*, 103 NW. U. L. Rev. 1607, 1610, 1682 (2009).

⁷ A.V. v. iParadigms Liab. Co., 544 F. Supp. 2d 473, 482 (E.D. Va. 2008): "This Court finds the "purpose and character" of iParadigms' use of Plaintiffs' written works to be *highly transformative*. Plaintiffs originally created and produced their works for the purpose of education and creative expression. iParadigms, through Turnitin, uses the papers for an entirely different purpose, namely, to prevent plagiarism and protect the students' written works from plagiarism. iParadigms achieves this by archiving the students' works as digital code and makes no use of any work's particular expressive or creative content beyond the limited use of comparison with other works." AV Ex Rel. Vanderhye v. iParadigms, LLC, 562 F. 3d 630, 640 (4th Cir. 2009): "The district court, in our view, correctly determined that the archiving of plaintiffs' papers was transformative and favored a finding of "fair use." *iParadigms' use of these works was completely unrelated to expressive content* and was instead aimed at detecting and discouraging plagiarism." Authors Guild, Inc. v. HathiTrust, 755 F. 3d 87, 97 (2nd Cir. 2014): "... we conclude that the creation of a full-text searchable database is a *quintessentially transformative use*."; Authors Guild, Inc. v. Google, Inc., 804 F.3d 202, 216-7 (2d Cir. 2015): "We have no difficulty concluding that Google's making of a digital copy of Plaintiffs' books for the purpose of enabling a search for identification of books containing a term of interest to the searcher involves a *highly transformative* purpose, in the sense intended by Campbell." Authors Guild, Inc. v. Google, Inc., 804 F.3d 202, 217 (2d Cir. 2015): "... through the ngrams tool, Google allows readers to learn the frequency of usage of selected words in the aggregate corpus of published books in different historical periods. *We have no doubt that the purpose of this copying is the sort of transformative purpose described in*

The non-expressive use cases make sense because TDM and analogous technological processes do not usurp the copyright owner's interest in communicating her original expression to the public because that expression is not communicated.

The Supreme Court's recent fair use decision in *Andy Warhol Foundation v. Goldsmith* ("AWF") does not call this into question, it actually reinforces the importance of focusing on the particular use made by the defendant and the prospect that the use might result in competitive substitution for the plaintiff's expressive work. By tying the availability of fair use defenses to the likelihood of expressive substitution, *AWF* helpfully clarifies the reason why transformative use has featured so prominently in the case law: the more transformative a use is, the less likely it is to substitute for the copyright owner's original expression. Using the author's work to reflect back on the original is an intrinsically different purpose; that difference in purpose makes expressive substitution less likely. In contrast, merely adding an overlay of new expression while leaving the original expression intact provides no such comfort. The majority in *AWF* rightly focuses our attention on how the defendant's use is likely to substitute for the author's original expression and makes that the measure of when the defendant's use is sufficiently transformative. This focus on expressive substitution makes it clear why non-expressive uses are strongly favored under the first fair use factor. By definition, non-expressive uses pose no threat of direct expressive substitution (in the language of transformative use, they are not just transformative, they are highly transformative.)

The Need to Expand the Current Exemptions

The academic text mining exemptions granted in 2021 have enabled some researchers to use digital methods to analyze e-books and DVDs without fear of liability under section 1201. Those existing exemptions struck a reasonable balance between maintaining the commercial utility of digital rights management and allowing genuine research that would qualify as fair use. The proposed expansion would merely allow researchers to collaborate more efficiently by sharing in DRM-free versions of works subject to the existing exemptions with other institutions who would independently qualify for the exemption.

The proposed expansion would not give any individual or institution access to a work that they did not already have. For example, if a researcher at University A has converted the contents of 100 DVDs into a format that allows for tagging and annotation, that researcher would be able to share the tagged and annotated works with a researcher at University B, if that University already has lawful access to all of those DVDs. Under the existing exemption the researcher at University B is already allowed to undertake this work separately, the expansion sought in the petition would simply allow her to do so without needlessly going through the process of removing the DRM herself. The point of the expansion is not just to save researchers the effort of reinventing the wheel, it will allow researchers to undertake more ambitious collaborative projects spanning across institutions.

The implications for the copyright owners, or rather, the absence of any implication for the copyright owners will remain the same.

Campbell as strongly favoring satisfaction of the first factor."

I ask that you grant the proposed very modest expansion exemption for text and data mining. Yours sincerely,

A handwritten signature in black ink, reading "Matthew Sag". The signature is fluid and cursive, with the first name "Matthew" and last name "Sag" clearly distinguishable.

Emory University School of Law
1301 Clifton Road, N.E.
Atlanta, GA 30322-2270
An equal opportunity, affirmative action university

Cell: (773)255-5856
Tel: (404)727-0535
Fax: (404)727-5685
msag@emory.edu

Appendix J
Letter from Rachael Samberg and Timothy Vollmer

December 1, 2023

To the Register of Copyrights:

We are copyright and scholarly publishing experts at University of California, Berkeley (UC Berkeley) writing in support of Authors Alliance's renewal and expansion of the 37 C.F.R. § 201.40(b)(4) and (b)(5)¹—the text and data mining (TDM) research exemptions to 17 U.S.C. § 1201's prohibition against circumvention of technological measures. We will refer to §§ 201.40(b)(4) and (b)(5) together as the "TDM Exemptions." TDM comprises a potent set of research methodologies advancing the progress of science and the useful arts, and researchers already rely on the TDM Exemptions to extract information from copyrighted works otherwise protected by technological protection measures. For the reasons set forth below, the TDM Exemptions should be renewed. Further, the TDM Exemptions should be expanded so that: scholars may securely share their decrypted corpora with regulation-compliant scholars at other institutions for purposes beyond collaboration or replication on a given research project.

Experience guiding scholars in TDM legal issues

Rachael Samberg is a lawyer and the Scholarly Communication Officer and Program Director of UC Berkeley Library's Office of Scholarly Communication Services (OSCS). Timothy Vollmer is Scholarly Communication and Copyright Librarian at OSCS. Through OSCS, we help scholars (including an ever-increasing number of TDM researchers) navigate the shifting publishing, intellectual property, and information policy landscapes in ways that promote research dissemination, accessibility, and impact.² We provide thousands of consultations to scholars each year, including regarding TDM law and policy issues.³ We have developed informational guides, workshops, and videos related to legal issues in TDM research,⁴ and speak and write about inconsistencies in TDM legal protections worldwide.⁵

Through all of our research and outreach with TDM researchers, we have developed a keen understanding of the legal and ethical challenges they face. In 2019, we obtained a National

¹ 37 CFR § 201.40(b)(4) pertains to motion pictures and (b)(5) to literary works.

² University of California, Berkeley Library. (n.d.). *Office of Scholarly Communication Services*. Available at <https://www.lib.berkeley.edu/research/scholarly-communication>

³ UC Berkeley Library Office of Scholarly Communication. (n.d.). *Annual Report FY22–23*. Retrieved September 23, 2023, from <https://docs.google.com/document/d/1WCVSVU6jNj8Kkt3zheY84LI8l5LS5fyFVXhycaQ7p0U/edit?usp=sharing>

⁴ For example, see UC Berkeley Library Office of Scholarly Communication. (n.d.). *Text Data Mining*. Retrieved September 23, 2023, from <https://www.lib.berkeley.edu/research/scholarly-communication/copyright?section=text-data-mining> and UC Berkeley Library Office of Scholarly Communication. (n.d.). *YouTube*. Retrieved September 23, 2023, from <https://www.youtube.com/channel/UCNUMwTyK0raTNNZVjhgB7KA>

⁵ Fiil-Flynn, S. M., Butler, B., Carroll, M., Cohen-Sasson, O., Craig, C., Guibault, L., Jaszi, P., Jütte, B. J., Katz, A., Quintais, J. P., Margoni, T., de Souza, A. R., Sag, M., Samberg, R., Schirru, L., Senftleben, M., Tur-Sinai, O., & Contreras, J. L. (2022). Legal reform to enhance global text and data mining research. *Science*, 378(6623), 951–953. <https://doi.org/10.1126/science.add6124>

Endowment for the Humanities (NEH) grant to host an institute to educate and empower digital humanities researchers and research-adjacent support staff (such as librarians and other professionals) to navigate the legal and ethical hurdles in the TDM research landscape.⁶ The *Building Legal Literacies for Text Data Mining* institute (“Building LLTDM Institute”), which we offered in June 2020, provided guidance and strategies on copyright, contracts, privacy, and ethics for TDM research in a U.S. context. To expand the community of TDM researchers able to navigate TDM law and policy issues, we subsequently published an openly licensed book covering institute topics.⁷ While our book’s chapter on technological protection measures predates the successful adoption of the TDM Exemptions, we have embarked on extensive outreach and education on how the TDM Exemptions operate, and offer workshops and presentations to researchers and the library community.⁸

In the Building LLTDM Institute’s instructional sessions and post-institute evaluations, participants identified cross-border and cross-institutional research collaborations as an ongoing challenge, noting that issues of collaboration pervaded their research.⁹ It became apparent that the U.S. digital humanities TDM practitioners lacked guidance on how to navigate these cross-border and cross-institutional concerns, including because of inconsistencies in copyright laws and regulations affecting the circumvention of technological protection measures for the purposes of conducting TDM. We secured a follow-on grant from the NEH to address these law and policy issues faced by U.S. digital humanities practitioners whose TDM research and practice intersects with foreign-held or -licensed content, or involves international research collaborations (see *Legal Literacies for Text Data Mining–Cross-Border* (“LLTDM-X”)¹⁰). We have now published a white paper¹¹ and pragmatic case study¹² offering analysis and guidance for U.S. TDM scholars working on cross-border TDM research.

⁶ *Building LLTDM Institute*. (n.d.). <https://buildinglltdm.org/institute/>

⁷ Althaus, S., Bamman, D., Benson, S., Butler, B., Cate, B., Courtney, K. K., Flynn, S., Gould, M., Hennesy, C., Koehl, E. D., Padilla, T., Reardon, S., Sag, M., Samberg, R., Schofield, B. L., Senseney, M., Vollmer, T., & Worthey, G. (2021). *Building Legal Literacies for Text Data Mining*. University of California, Berkeley. <https://doi.org/10.48451/S1159P>

⁸ See, e.g., Samberg, R., & Stallman, E. (2022, July 22). *DMCA & TDM: New Exemption for Text Mining DRM-Protected Materials*

<https://docs.google.com/presentation/d/1toAx6bEHc2BDcYyAFNOsE6NdGEzsp8IUY6srsLKK4CI> and Office of Scholarly Communication. (n.d.). *Special use case: Digital rights management*. Text Data Mining. Retrieved October 17, 2023, from <https://www.lib.berkeley.edu/research/scholarly-communication/copyright?section=text-data-mining>

⁹ Samberg, R., & Vollmer, T. (2021). *Building Legal Literacies for Text Data Mining: Institute White Paper*. <https://escholarship.org/uc/item/1db5350t>

¹⁰ *LLTDM-X*. (2022). <https://buildinglltdm.org/lltdmx/>

¹¹ Samberg, R., Vollmer, T., & Padilla, T. (2023). *Legal Literacies for Text Data Mining – Cross-Border (“LLTDM-X”): White Paper*. <https://escholarship.org/uc/item/5k91r1s1>

¹² Samberg, R., Vollmer, T., & Padilla, T. (2023). *Legal Literacies for Text Data Mining – Cross-Border (“LLTDM-X”): Case Study*. <https://escholarship.org/uc/item/1w03f9r2>

Scholars rely on the TDM Exemptions, but there are hurdles

Since the authorization of the TDM Exemptions, we have helped a number of TDM researchers rely on these regulations to advance global knowledge and science. By way of example:

- We have worked with UC Berkeley Professor David Bamman to obtain a grant to leverage TDM and artificial intelligence modeling to investigate the representation of race, gender, and place in both popular and prestige films and TV shows.¹³ Bamman is now answering questions such as: How are race and gender tied to the depiction of characters on screen, and how has this changed over the past 50 years? How much attention is given to cities vs. rural environments? And how might this kind of representation on screen shape the development of stereotypes in viewers? Because the motion pictures that Bamman is studying are protected by technological protection measures, this research would not have been possible absent the TDM Exemptions.
- Likewise, as part of our work supporting cross-border research in LLTDM-X, we provided guidance to a scholar¹⁴ who hosts a digital research platform with decrypted and annotated excerpts of films; these clips facilitate a wide range of media studies research by scholars worldwide. This project is made possible not only by the TDM exemptions, but also international copyright laws that, as required by the Berne Convention, have exceptions for private (individual) uses for educational, research, or analytical purposes.

While the rigors of the TDM Exemptions preserve the fair use and security of the copyrighted works whose technological protection measures are being circumvented, the TDM Exemptions as presently drafted do present some hurdles that could be addressed through regulatory expansion. One key issue that TDM researchers face in relying on the TDM Exemptions relates to the expense and time involved in having to undertake the circumvention. If researchers were able to securely share their decrypted corpora with researchers at other institutions who had already purchased or licensed the same works (yielding no market harm issues), this would remove research barriers—particularly for underfunded institutions and under-resourced disciplines that lack sufficient grant funding to cover decryption expenses.

To illustrate this point, the current grant-funded project that we are supporting aims to purchase and circumvent DRM on 2,500 films and 800 television seasons to use in the TDM research. A significant proportion of the requested grant funds needed to be allocated to hire student researchers to conduct the circumvention and quality-check the decryption. It is estimated that the resulting decryption pipeline has a current throughput of 75 DVDs per day, and could be increased with the purchase of additional computing and human resources.¹⁵ Even if a

¹³ UC Berkeley Library Communications. (2023, January 26). *Poetry, mining Hollywood, and digital books: Mellon Foundation grants will support groundbreaking work at the UC Berkeley Library and beyond*. <https://www.lib.berkeley.edu/about/news/mellon-grants>

¹⁴ Participants in LLTDM-X were assured anonymity (waivable at their own discretion) to encourage fulsome research discussion.

¹⁵ We also note that for many kinds of TDM research, including that predicated upon machine learning for which artificial intelligence must first be trained, researchers' current interpretation of the regulations necessitates breaking DRM twice on each DVD (carrying out the process of circumvention twice to make

corresponding scholar at another institution complied with the requirements of the TDM Exemptions and purchased the very same corpus works for study, the scholar would still have to pay thousands of additional dollars to set up a similar process to engage in what is ultimately duplicative circumvention and quality-checking. In many scholarly disciplines, these funds simply are not available.

Currently, under 37 CFR § 201.40(4)(i)(D) and (5)(i)(D), researchers or institutions may share the decrypted corpora only with similarly-compliant institutions or scholars for purposes of “collaboration or replication of the research.” A more permissive environment—specifically, one that extended sharing for the study of new or other research questions (i.e. beyond mere project collaboration or replication), and limited that sharing to scholars or institutions who have already purchased or licensed the same corpus materials—would create a more efficient research pipeline and speed up discovery and the advancement of knowledge. And because the scholars at other institutions have already licensed or purchased the same content in compliance with § 201.40(4)(i)(B) and (5)(i)(B), the sharing of the decrypted materials in a secure environment would not serve as a market substitute and would remain squarely within fair use.¹⁶

Digital humanities research in general, of which TDM methodologies form a growing part, is marked overall by collaborativeness across institutions and geographical boundaries.¹⁷ TDM scholars need to be able to share the research corpora with their colleagues to complete their research, conduct analysis, and ensure reproducibility—but also to conduct their work in an equitable environment that does not unfairly bias questions for study toward only those institutions or scholars who can afford to undertake duplicative decryption at scale.

Conclusion

Our office will continue to provide the requisite guidance for researchers to understand law and policy issues encompassing text data mining. By renewing the TDM Exemptions to 17 U.S.C. § 1201 for literary works and motion pictures, TDM researchers will be able to engage in critical research that advances the progress of science and the useful arts. By expanding the TDM Exemptions to enable broader sharing of corpora under controlled conditions, TDM scholars will be better equipped to work with collaborators in the U.S. and abroad, championing the fundamental greater freedom of inquiry and also aid research support staff in their quest to provide the most accurate information and education to the university community.

use of the affordances of two different exemptions for different research purposes, under 37 CFR 201.40 (b)(1) and then again under 37 CFR 201.40(b)(4)). This regulatory interpretation further exacerbates the burden of the decryption undertaking, and further underscores the benefit of avoiding duplicative decryption by researchers at other institutions who are otherwise compliant with the regulations.

¹⁶ *Campbell v. Acuff-Rose Music, Inc.*, 510 U.S. 569, 591 (1994); *Authors Guild v. Google, Inc.*, 804 F.3d 202 (2d Cir. 2015).

¹⁷ Su, F. (2020). Cross-national digital humanities research collaborations: structure, patterns and themes. *The Journal of documentation*. 76(6): 1295-1312; Nyhan, J., & Duke-Williams, O. (2014). Joint and multi-authored publication patterns in digital humanities. *Literary and Linguistic Computing*, 29(3), 387-399. <https://academic.oup.com/dsh/article/29/3/387/986317>; Kemman, M. (2019, August). Boundary practices of digital humanities collaborations. *Digital Humanities Benelux Journal*. 1-24. <https://journal.dhbenelux.org/journal/issues/001/Article-Kemman/kemman-main.tex.html>

Thank you for the opportunity to address this matter. We would be pleased to discuss it further as desired.

Regards,

Rachael Samberg

A handwritten signature in cursive script, reading "Rachael Samberg".

Timothy Vollmer

A handwritten signature in cursive script, reading "Timothy Vollmer".

Appendix K
Letter from Lauren Tilton and Taylor Arnold



To the Register of Copyrights,

Text and Data Mining (“TDM”) is key to our research at the Distant Viewing (DV) Lab. The DV Lab uses and develops computational techniques to analyze visual culture on a large scale. Our lab brings together scholars across fields including data science, digital humanities, and media studies to forge interdisciplinary approaches to studying sources such as film, TV, and photography. We develop tools, methods, and datasets that can be re-used by other researchers. Because visual sources are often subject to digital rights management software, the exemption is key to the success of our lab.

Projects include a new book titled *Distant Viewing: Computational Analysis of Digital Images* on The MIT Press, which is known for publishing cutting edge research related to technology. For example, the chapter on TV sit-coms shows how computer vision can intervene in the study of post-war TV and forge new methods for media studies. To study TV though often requires accessing the data through DVDs, which requires circumventing digital rights management. This is the only way to work with this important cultural data. The book also developed the distant viewing method, which offers a new approach to data science, a key growth area in university education and the US economy.

The DV Method outlines how to use computer vision to study digital images including TV and film. The innovation has led to significant funding, including a \$500,000 grant from the Mellon Foundation, the most prestigious private foundation in the humanities, to build the Distant Viewing Toolkit. Designed to support media and AI literacy, the toolkit is being designed to help a broader public understand how computer vision can help them analyze images as well as to show how these tools work. We are also incorporating these methods into our media studies, digital humanities, and data science classes. Students love to study the media they engage with regularly, including their favorite films and TV shows. They know that these films and shows have shaped their lives, from the ways they dress to the cultural values that they share. Yet, most of the 20th and 21st century media is under significant copyright and on data sources with DRM. Studying them is key to effective media literacy education and data science education. The toolkit and our curriculum rely on the exemption.

Being able to build a corpus is key to our work and our curriculum. Reducing the barriers to building a corpus is paramount. Researchers and students have a range of interests that aren’t served by packaged data sets, which often don’t even exist for visual culture data. There are significant funding limits in humanities subjects versus STEM in higher education. Add that only the most affluent institution such as Stanford have the resources to spend millions of dollars on one data set, which is creating a huge inequality in access to data. One change that would be greatly helpful is allowing corpora-sharing across institution. This would significantly reduce this funding barrier and increase quality of digital humanities and data science research. This would also increase the quality and quantity of digital humanities research, and contribute to our



UNIVERSITY OF RICHMOND

School of Arts and Sciences

University of Richmond
Rhetoric & Communication Studies
402-C Weinstein Hall
Richmond, VA 23173
T: 504 782-3485
E: ltilton@richmond.edu

understanding of history and culture. For example, we could work with TV data held at different institutions to better understand the history of TV over the last fifty years. Right now, we can only share metadata, but being able to see the actual image is key. To add, because of the different runs of DVDs over the decades and other data sources, it's hard to find the exact DVD that another group generated their data from. This produces a major replicability issue. Being able to see the actual metadata augmented with computer vision annotations such as people and objects from researchers alongside the actual media is key if we want to fully understand data, conduct larger scale research, and authenticate and reproduce results. For example, I would love to work with our colleagues at European universities who have access to TV shows such as sit-coms that aired in Europe. We'd love to study how American shows may have been changed or adjusted for European audiences. They can share with us, but we can't share our data with them. The inability to share data is harming our ability to collaborate, expand research questions, and scale up our research.

The exemption is key to the next three years of our research. The Mellon Grant involves working with film and TV, and we would love to expand to sources such as Hollywood film for we know this will be of great interest to users of the toolkit, particularly students. We would love to gain new research insights from other labs and groups across the US such as the Media Ecology Project at Dartmouth, but right now, we can't share much of our data from TV and film unless we are directly working on the same projects. They have their own corpus too. Imagine what we could do if we worked together! As well, there is a significant pedagogical payoff. One benefit is the messiness of humanities data such as TV and film. This data is tough to work with. As we think about a next generation of data scientists, the more they've seen complex, messy data, the better they will be trained for the future.

The exemption, and expanding to allow researchers to ask different research questions and apply different methods on a given topic, is key to the success of our lab and the study of digital humanities and data science. The inability to effectively use the exemption due to practical barriers would put American scholars at a competitive disadvantage to scholars in other parts of the world, specifically the European Union. National commitments such as the Netherlands's CLARIAH project and continental commitments such as the EU's DARIAH infrastructure are opening up extensive data for distant viewing, reading, and listening at institutions across the EU. These scholars are positioned to innovate in AI and machine learning while scholars in the United States would be barred from this kind of research if this expansion is not granted. Therefore, our appeal is not just about specific research areas, but a call to remove a barrier that prevents US scholars from being at the forefront of TDM with audiovisual data in the global community.

Sincerely,

Lauren Tilton
E. Claiborne Robins Professor of Liberal Arts and Digital Humanities



UNIVERSITY OF RICHMOND
School of Arts and Sciences

University of Richmond
Rhetoric & Communication Studies
402-C Weinstein Hall
Richmond, VA 23173
T: 504 782-3485
E: ltilton@richmond.edu

Department of Rhetoric & Communication Studies
Director, Distant Viewing Lab
University of Richmond

Taylor Arnold
Associate Professor of Data Science
Department of Mathematics & Statistics
Director, Distant Viewing Lab
University of Richmond

Appendix L
Letter from Henry Alexander Wermer-Colan

December 14th, 2023

I am writing in support of the expansion to the Exemption to the anti-circumvention provisions of the Digital Millennium Copyright Act to enable librarians and researchers to pursue their academic work in text mining and data analysis. As the Interim Academic Director and Digital Scholarship Coordinator at Temple University Libraries' Loretta C. Duckworth Scholars Studio, I regularly support students, librarians, and faculty in the development of research and teaching projects involving data curation and analysis. Restrictions imposed by copyright are one of the most frequent obstacles I encounter to impactful research on contemporary culture, limiting the majority of modern scholarship from fully accessing the subject of their analyses.

I write today to support and advocate for the expansion to the Exemption 37 C.F.R. § 201.40(b)(4) and (b)(5) to further enable this important field of scholarship. In my own work, I've conducted multiple analyses within the confines of the law to make copyrighted contemporary culture more accessible and understandable, and this Exemption has made possible research I would never have considered pursuing otherwise. In particular, I am working with Dr. Laura McGrath on a Mellon Foundation's Text and Data Mining "Demonstrating Fair Use" grant-funded project to study the current phenomenon of banning books in schools and libraries across the United States. Through this project, we explore the politics of representation in contemporary culture, bringing into relief what ideas of diversity might be considered taboo, expanding our understanding of the forces that suppress free speech, the dilemmas posed for the field of education, and the potential for the humanities to make a change in how the public thinks about problems of identity and liberty. Thanks to the Exemption provided by DMCA § 1201, we have been able to conduct large scale analyses of all the banned books within a short period of time. In comparison to physically scanning and transforming physical books, the ability to buy ebooks from Kobo Books and convert the files into text files we can use for non-consumptive research has increased the speed of our digitization process a hundred-fold, going from months to weeks. For further information about our project, its research scope, and the opportunities the Exemption has opened up for us to work with students on a pedagogical project, see the multiple news stories that have been published recently in [Temple Now](#) and [Temple News](#).

In my prior letter written to the Register of Copyrights in 2020, I addressed the broader scope of the obstacles copyright was presenting to research at scale in universities. Today I was only hoping to emphasize elements of the current Exemption which, while enabling new fields of research, are not sufficiently flexible to be useful at scale. The ability to share this corpus with other researchers outside of Temple University would be greatly beneficial. There are many socially valuable questions that can be asked of this corpus of banned books, and the perspectives of other academics are vitally important, even if we at Temple do not have the capacity to support direct collaboration. Scholars from a wide-range of disciplines and fields may have questions about the social and political context of banning books beyond the scope of Dr. McGrath's and my research project. It is nearly impossible to predict the types of cultural data they may wish to compare with our 'banned books' dataset, including data available from news coverage of the phenomenon on social media posts and websites. Scholars in the humanities working on other historical periods when censorship posed problems for

freedom of speech and education could find renewed interest in comparing their studies to this contemporary corpus. Enabling these wide-ranging research projects was always one of our aims of compiling and analyzing this corpus, and we hope the expansion to the DMCA Exemption can enable research on this timely subject to flourish in the coming years.

In order to ensure research we publish can remain relevant and provide a building block for future study, it's also important to ensure libraries can realistically meet the growing demands of scholars wanting to conduct computational research on contemporary culture. Outside of purchasing the relevant ebooks, building the corpus required significant amounts of time, labor, and costs (tens of thousands of dollars go into the data curation work of building valuable metadata about the books, and standardizing the wide-range of books into genre categories and data formats useful for analysis). In my capacity supporting this work as a librarian doing data curation, I would be happy to support other researchers who need to build their own datasets, but it would be costly and a waste of time for another scholar to rebuild the exact same dataset that already exists at Temple University Libraries. If a dozen Temple researchers all wanted to access a dataset, and each had to build it separately, it would require too much work for the Libraries to support, and we'd be stuck only supporting some researchers while others would be excluded.

I ask that you expand the Exemption to the DMCA to allow a more comprehensive study of our literary heritage and broader modes of sustainable research for text and data mining. Continued support in this field for researchers studying contemporary culture, especially by reducing the impositions to studying this material at scale, will have broad impacts on our understanding of the wide array of contemporary culture under copyright.

Sincerely,

A handwritten signature in cursive script that reads "Alex Wermer-Colan".

Henry Alexander Wermer-Colan
Interim Academic Director and Digital Scholarship Coordinator
Loretta C. Duckworth Scholars Studio
Temple University Libraries

Appendix M
Letter from the Mellon Foundation



Mellon Foundation
140 E. 62nd St.
New York, NY 10065

19 December 2023

To the U.S. Copyright Office:

This letter is submitted on behalf of the Mellon Foundation's Public Knowledge program in support of the Authors Alliance, Library Copyright Alliance, and American Association of University Professor's petition to expand the existing Section 1201 text and data mining exemptions currently found in 37 C.F.R. 37 C.F.R. § 201.40(b)(4) and 37 C.F.R. § 201.40(b)(5).

The Mellon Foundation is the largest private funder in the United States of the arts and the humanities, with more than \$530 million in grants made in 2023. By investing in higher education, the humanities, the arts, and the nation's cultural heritage, we support work that deepens our understanding of our shared humanity. With a commitment to social justice throughout all of our grantmaking, Mellon seeks to build just communities enriched by meaning and empowered by critical thinking, where ideas and imagination can thrive.

The Foundation has a strong interest in supporting advances in text and data mining ("TDM"). Its Public Knowledge program funds efforts in the creation and preservation of the cultural and scholarly record—vast and ever-expanding—that documents society's complex, intertwined humanity. The goal of the program is to increase equitable access to deep knowledge that helps to build an informed, heterogeneous, and civically engaged society. We aspire to cultivate networks and maintainable infrastructure, expand digital inclusion, and ensure that more authentic, reflective, and nuanced stories are revealed, preserved, and told.

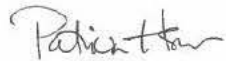
Through the Public Knowledge program, Mellon has funded a large number of TDM projects over the years to advance computational research tools, techniques, and corpora development. As the petitioners also explain, we have already seen the transformative potential of TDM research under the existing exemption, allowing researchers to expose a more nuanced understanding of history, culture, and society. Continued TDM of in-copyright content helps ensure that *contemporary* history, culture, and society are not omitted from the scholarly record. Importantly, because the exemption allows researchers to interrogate modern, culturally relevant in-copyright materials, it has allowed them to make their research more relevant and accessible for current social and civic concerns.

At the same time, we see the challenges researchers continue to face under the existing exemption. Because the current exemption limits the ways in which one research project can share access to their data with others, it has meant that subsequent research projects must start as if from a blank slate, effectively reinventing the wheel: each project must break independently the "digital locks" of technological protection measures ("TPMs"), process data, and build a corpus

in a form that is useful for research. As a funder of these efforts, the Mellon Foundation has seen first-hand how expensive and complicated it is to build a corpus—which requires technical staff and expertise, as well as computing resources and tools. These costs have meant that TDM research that engages works protected by TPMs has largely been limited to projects at institutions that have the resources to compensate and maintain technical staff and infrastructure, supplemented by grants like those we have supported. It does not benefit the “progress of science and the useful arts” when technical barriers mean that this type of research can be done only by researchers with ample resources. It is our belief that by expanding the number of institutions that can benefit from the technical work of breaking TPMs for TDM research, the proposed expansion of the exemption in the regulations would result in a more diverse and rich set of research projects.

Finally, we believe the expansion would catalyze the speed and quality of TDM research. The barrier to sharing fosters a siloed approach to TDM efforts and prohibits projects from benefiting from shared understandings and learnings, which can often lead to innovation. We have seen in a number of other grant areas the tremendous value of collaborative efforts to build, share, and innovate upon corpora. Often these efforts do not begin with specific or well-defined collaborative research questions, but the collective ability to develop corpora and make them available for new research questions has spurred innovative new lines of research. The current exemption imposes an artificial technical barrier to such collaborative sharing, while the proposed expansion would encourage it.

Sincerely,



Patricia Hswe
Program Director, Public Knowledge
ph@mellon.org