

This is a Word document that allows users to type into the spaces below. The comment may be single-spaced, but should be in at least 12-point type. The italicized instructions on this template may be deleted.

UNITED STATES COPYRIGHT OFFICE



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

Please submit a separate comment for each proposed class.

NOTE: This form must be used in all three rounds of comments by all commenters not submitting short-form comments directly through regulations.gov, whether the commenter is supporting, opposing, or merely providing pertinent information about a proposed exemption. When commenting on a proposed expansion to an existing exemption, you should focus your comments only on those issues relevant to the proposed expansion.

Check here if multimedia evidence is being provided in connection with this comment. Commenters can provide relevant multimedia evidence to support their arguments. Please note that such evidence must be separately submitted in conformity with the Office's instructions for submitting multimedia evidence, available on the Copyright Office website at <https://www.copyright.gov/1201/2021>.

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
UC Berkeley, School of Law
Catherine Crump, Director
Gabrielle Daley, Clinical Teaching Fellow
Elizabeth Fu, Jason Francis, Alistair McIntyre, and Erin Moore, Clinical Law Students
crcrump@clinical.law.berkeley.edu

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 Web site 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.

(2) **The American Association of University Professors (AAUP)** is a nonprofit membership association of over 44,000 faculty, librarians, graduate students, 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.

Risa Lieberwitz, AAUP General Counsel, rlieberwitz@aaup.org
Aaron Nisenson, AAUP Senior Counsel, anisenson@aaup.org
Nancy Long, AAUP Associate Counsel, nlong@aaup.org

(3) **The Library Copyright Alliance (LCA)** consists of three major library associations—the American Library Association (ALA), the Association of College and Research Libraries (ACRL), 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 —Motion Pictures and Literary Works—Text and Data Mining.....	4
Item C. Overview	4
1. Background on TDM	5
a. TDM is a valuable research technique.....	5
b. The section 1201 prohibition drives scholars’ research agendas away from focusing on in-copyright motion pictures and literary works.....	6
c. The section 1201 prohibition means TDM cannot be used to answer many important questions of contemporary relevance, and it reinforces biases in certain fields of scholarship.	9
d. The section 1201 prohibition is harmful to teaching.	10
e. Existing alternatives to circumventing TPMs are inadequate.....	10
i. Limits of existing digital libraries	11
ii. Limits on creating one’s own collection.....	13
f. Granting this exemption would allow valuable research to move forward.	16
2. Overview of Comment.....	17
Item D. Technological Protection Measures and Methods of Circumvention	18
1. Motion Pictures	18
2. Literary Works Distributed Electronically.....	21
Item E. Asserted Adverse Effects on Noninfringing Uses	21
1. The proposed classes include works protected by copyright.....	22
2. Creating collections of copyrighted works for TDM is a noninfringing use.....	22
a. The <i>Google Books</i> case	22
b. The proposed exemption is a fair use for the reasons articulated in <i>Google Books</i>	25
3. TDM researchers and teachers are adversely affected in their ability to make noninfringing uses.....	28
a. Granting the exemption would increase the availability for use of copyrighted works. .	29
b. Granting the exemption would increase the availability for use of works for nonprofit archival, preservation, and educational purposes.....	31
c. The prohibition on the circumvention of technological measures applied to copyrighted works adversely impacts criticism, comment, news reporting, teaching, scholarship, and research.	31

d. Circumvention of technological measures to enable TDM research is unlikely to adversely affect the market for or value of copyrighted motion pictures and electronically distributed literary works.	32
e. Granting the exemption would promote equity.	33
4. The statutory prohibition on circumventing access controls is the cause of adverse effects on TDM researchers.	33
Documentary Evidence.....	35

Item B. Proposed Class Addressed —Motion Pictures and Literary Works—Text and Data Mining

The Proposed Classes include:

Proposed Class 7(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 7(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.

Item C. Overview

Proponents seek an exemption from 17 U.S.C. § 1201’s prohibition on circumventing technological protection measures (“TPMs”) to facilitate text and data mining (“TDM”) of lawfully accessed motion pictures and lawfully accessed literary works distributed electronically.¹ Specifically, we seek an exemption to allow copying of these classes of works to create a collection of works on which to conduct TDM.

An exemption is necessary because TDM requires creating a dataset of works of interest, which “typically involves digitizing or downloading (i.e. reproducing) potentially copyrighted works in order to perform algorithmic extractions” on them.² Courts have found that creating these sorts

¹ The umbrella term “TDM” is used internationally to refer to the use of copyrighted work in computational research. Michael W. Carroll, *Copyright and the Progress of Science: Why Text and Data Mining is Lawful*, 53 U.C. Davis L. Rev. 893, 899 n.19 (2019). TDM of literary works is sometimes called computational text analysis. See Rachael G. Samberg and Cody Hennesy, *Law and Literacy in Non-Consumptive Text Mining: Guiding Researchers Through the Landscape of Computational Text Analysis*, in *Copyright Conversations: Rights Literacy in a Digital World* 289, 289 (ACRL, 2019), <https://escholarship.org/uc/item/55j0h74g#main>. Because this comment focuses on both motion pictures and literary works distributed electronically, it uses the term TDM.

² Letter from Rachael Samberg and Timothy Vollmer, Appendix K, at 2.

of collections is fair use.³ However, many in-copyright motion pictures and electronically distributed literary works are protected by TPMs, making an exemption necessary to prevent researchers from being adversely affected in their ability to make non-infringing uses of these classes of works. It will also allow researchers to teach this valuable technique to others, another function substantially impeded by section 1201.

1. Background on TDM

TDM makes it possible to glean insights from large volumes of material, more than a single researcher could ever examine unaided. Yet because of section 1201's prohibition on bypassing TPMs, researchers are being deterred from applying the technique to virtually all in-copyright motion pictures and literary works distributed electronically, as these are generally protected by TPMs. Because of these obstacles, little TDM research is done on motion pictures at all, and what work is done on literary works almost exclusively focuses on works at least a century old. For this reason, the technique cannot be used to focus on issues that are of even remote contemporary relevance, from the rise of the Harlem Renaissance to the shifting portrayals of Muslim Americans after the 9/11 terrorist attacks. It also distorts whose work is analyzed using TDM by limiting its application to the demographic groups most published nearly 100 years ago. In addition, the section 1201 prohibition impedes teaching by making it more difficult to pass the techniques along to the next generation of scholars. Many academics have written letters in support of this comment describing these problems, as well as putting forth a dazzling array of valuable research and teaching projects they will be able to pursue if the exemption is granted.

a. TDM is a valuable research technique.

TDM makes it possible to sift through substantial amounts of information to draw groundbreaking conclusions. This is true across disciplines. In medical science, TDM has been used to perform an overview of a mass of coronavirus literature.⁴ Researchers have also begun to explore the technique's promise for extracting clinically actionable information from biomedical publications and clinical notes.⁵ Others have assessed its promise for drawing insights from the masses of medical images and associated reports that hospitals accumulate.⁶

³ See *Authors Guild, Inc. v. HathiTrust*, 755 F.3d 87 (2d Cir. 2014) [hereinafter *HathiTrust*]; *Authors Guild v. Google, Inc.*, 804 F.3d 202 (2d Cir. 2015) [hereinafter *Google Books*].

⁴ Xian Cheng et al., *An Overview of Literature on COVID-19, MERS and SARS: Using Text Mining and Latent Dirichlet Allocation*, *J. Information Sci.* (Aug. 31, 2020), <https://doi.org/10.1177/0165551520954674>.

⁵ Michael Simmons et al., *Text Mining for Precision Medicine: Bringing Structure to EHRs and Biomedical Literature to Understand Genes and Health*, 939 *Translational Biomedical Informatics* 139 (2016), https://link.springer.com/chapter/10.1007%2F978-981-10-1503-8_7.

⁶ Yifan Peng et al., *Text Mining and Deep Learning for Disease Classification*, in *Handbook Med. Imaging Computing and Comput. Assisted Intervention*, 109–35 (Academic Press 2020), <https://www.sciencedirect.com/science/article/pii/B9780128161760000107?via%3Dihub>.

In social science, studies have used TDM to analyze job advertisements to identify direct discrimination during the hiring process.⁷ It has also been used to study police officer body-worn camera footage, uncovering that police officers speak less respectfully to Black than to white community members even under similar circumstances.⁸

TDM also shows great promise for drawing insights from literary works and motion pictures. Regarding literature, some 221,597 fiction books were printed in English in 2015 alone, more than a single scholar could read in a lifetime.⁹ TDM allows researchers to “‘scale up’ more familiar humanistic approaches and investigate questions of how literary genres evolve, how literary style circulates within and across linguistic contexts, and how patterns of racial discourse in society at large filter down into literary expression.”¹⁰ TDM has been used to “observe trends such as the marked decline in fiction written from a first-person point of view that took place from the mid-late 1700s to the early-mid 1800s, the weakening of gender stereotypes, and the staying power of literary standards over time.”¹¹ Those who apply TDM to motion pictures view the technique as every bit as promising for their field. Researchers believe the technique will provide insight into the politics of representation in the Network era of American television,¹² into what elements make a movie a Hollywood blockbuster,¹³ and into whether it is possible to identify the components that make up a director’s unique visual style.¹⁴

- b. The section 1201 prohibition drives scholars’ research agendas away from focusing on in-copyright motion pictures and literary works.

Although TDM shows great promise for analyzing in-copyright motion pictures and literary works, section 1201 impedes the application of TDM techniques to motion pictures and literary

⁷ Panggih Kusuma Ningrum et al., *Text Mining of Online Job Advertisements to Identify Direct Discrimination During Job Hunting Process: A Case Study in Indonesia*, 15 PLoS ONE 6 (2020), <https://doi.org/10.1371/journal.pone.0233746>.

⁸ Rob Voigt et al., *Language from Police Body Camera Footage Shows Racial Disparities in Officer Respect*, 114 Proc. Nat’l Acad. Sci. 25 (June 2017), <https://www.pnas.org/content/114/25/6521>.

⁹ Erik Fredner, *How Many Novels Have Been Published in English? (An Attempt)*, Stan. Literary Lab (Mar. 14, 2017), <https://litlab.stanford.edu/how-many-novels-have-been-published-in-english-an-attempt/>.

¹⁰ Letter from Hoyt Long, Appendix G, at 1.

¹¹ Letter from Ted Underwood, Appendix N, at 1. For a list of questions that can be answered through large scale TDM, see Letter from Matthew Jockers, Appendix F, at 1–2.

¹² Letter from Lauren Tilton and Taylor Arnold, Appendix M, at 1.

¹³ *Id.*

¹⁴ Letter from David Bamman, Appendix B, at 2.

works by prohibiting researchers from circumventing the TPMs that guard most in-copyright motion pictures and literary works distributed electronically. This means that researchers cannot access these works to create the collections they need to conduct TDM, even though (as we discuss in greater detail in Item E) this type of research constitutes a fair use. Put simply, “[t]he absence of a § 1201 exemption specifically for TDM research constrains scholarly inquiries into important sociopolitical and humanistic trends and, as a result, inhibits the advancement of knowledge.”¹⁵

To understand the problem, compare two researchers: one who seeks to conduct TDM by relying on in-copyright works that are *not* protected by TPMs and a second who seeks to conduct TDM by relying on in-copyright works that *are* protected by TPMs. Researchers in the first group are free to pursue their research because it constitutes a fair use. Researchers in the second group, however, must abandon their efforts, because circumventing TPMs is unlawful.¹⁶

Many scholars and other researchers wrote letters of support for this comment that describe the ways section 1201 is distorting research into motion pictures and literature.¹⁷ For example, Professor Matthew Sag, an expert on legal issues related to TDM and a member of the HathiTrust Research Center Advisory Board, concludes:

Researchers, who are rightfully concerned about § 1201 liability, frequently curtail their studies because these otherwise useful digital works are protected by TPMs. They instead focus their efforts on low-risk data, such as works in the public domain. Because of the chilling effect of § 1201, the world is deprived of significant research into contemporary cultural works including literature, movies, and tv shows.¹⁸

¹⁵ Letter from Rachael Samberg and Timothy Vollmer, Appendix K, at 2.

¹⁶ *Id.*

¹⁷ *See, e.g.*, Letter from The Association for Computers and the Humanities (ACH), Appendix A, at 2 (explaining that section 1201 has “a profound impact on the career trajectories of scholars” because these scholars are forced to focus on public domain works merely to avoid section 1201 liability); Letter from Matthew Jockers, Appendix F, at 3 (“researchers like me are forced to choose what literature to analyze, not based on what offers the most interesting or important research questions, but instead based on what literature we can lawfully access.”); Letter from Hoyt Long, Appendix G, at 3 (“Scholars, knowing the limitations imposed by § 1201, often discard nascent inquiries before a research question can be properly defined. In effect, § 1201 is warping the development of digital humanities as a discipline.”); Letter from Rachael Samberg and Timothy Vollmer, Appendix K, at 3 (concluding that “some digital humanities researchers have gravitated to low-friction research questions and texts” to avoid decisions about rights-protected data).

¹⁸ Letter from Matthew Sag, Appendix J, at 2.

In addition to these general observations, several academics provided letters of support describing the ways in which section 1201 has had an adverse impact on their own research. As to literary works:

- James Clawson, Professor of English at Grambling State University in Louisiana, explains that he has had to shift his focus away from the post-World War II novels he studied during his PhD in order to pursue TDM—a shift he reports is directly attributable to section 1201.¹⁹
- Dr. Henry Alexander Wermer-Colan, Postdoctoral Fellow and coordinator of digital scholarship at Temple University Libraries’ Loretta C. Duckworth Scholars Studio, writes about the way the prohibition on circumventing TPMs blocks him from pursuing valuable projects. He writes, “I was unable to include sophisticated data analysis methods in my dissertation because there was no way, through my university, available library databases, or other means to build a representative dataset of post-WWII literature”—a problem he attributes directly to section 1201.²⁰
- Melanie Walsh, a Postdoctoral Associate in Information Science at Cornell University, explains how the section 1201 prohibition has forced her to analyze tweets rather than books: “For example, I have studied how #BlackLivesMatter tweets quote the novelist and civil rights activist James Baldwin, using computational methods to archive and analyze these tweets. But I have not attempted a computational analysis of Baldwin’s literary corpus itself (1949–1985) because these texts are so difficult to access due to Section 1201.”²¹

The problem is even more severe for motion pictures.²² The Association for Computers and the Humanities (ACH) is the professional organization for the digital humanities in the United States. ACH notes, although film and television are among the “most powerful cultural forms of the late 20th and early 21st century,” it is illegal to bypass TPMs to apply TDM to these materials.²³ Professors Lauren Tilton and Taylor Arnold, who direct the Distant Viewing Lab²⁴ at the University of Richmond, report that “the scope of our research is incredibly restricted due to DMCA § 1201. Essentially, we cannot study the 20th and 21st century visual culture, which is only available through media formats such as DVDs and protected digital files.”²⁵ While those studying literary works can at least rely on mostly pre-1925 public domain works, when it comes

¹⁹ Letter from James Clawson, Appendix C, at 1.

²⁰ Letter from Henry Alexander Wermer-Colan, Appendix P, at 1.

²¹ Letter from Melanie Walsh, Appendix O, at 1.

²² Letter from ACH, Appendix A, at 3 (describing the “equal or greater” negative effect of section 1201 on scholarship that uses image and sound data).

²³ *Id.* at 4.

²⁴ The Distant Viewing Lab “uses and develops computational techniques to analyze visual culture on a large scale.” Letter from Lauren Tilton and Taylor Arnold, Appendix M, at 1.

²⁵ *Id.*

to motion pictures, “§ 1201 is a major barrier because it eliminates entire areas of study, including much of the available materials in 20th and 21st century film and television.”²⁶

- c. The section 1201 prohibition means TDM cannot be used to answer many important questions of contemporary relevance, and it reinforces biases in certain fields of scholarship.

Section 1201 imposes real and substantial costs on TDM scholarship of both motion pictures and literary works. With literary works, because of section 1201, a substantial amount of TDM research has focused on texts in the public domain—which largely consists of works created before 1925.²⁷ This “represents works nearly a century removed from the present day, limiting its ability to answer research questions that are relevant to a contemporary audience.”²⁸ For motion pictures, virtually all materials are in copyright.

There is real value in studying modern works. As members of the Data-Sitters Club, a research group under the Stanford Literary Lab, explain, “the critical analysis of modern, popular texts is a vital part of humanities research; it helps us to understand how books both mirror and shape people’s understanding of the world and the major issues of our time.”²⁹

Section 1201 means there are many questions that TDM cannot be used to answer. As Professor Bamman of the School of Information at UC Berkeley explains, “[t]hese include not only 21st-century questions on the influence of the internet and social media on literary forms and reading behavior—but even much older questions including the rise of the Harlem Renaissance in the 1920s and 30s.”³⁰ Jes Lopez, a PhD candidate in the English department at Michigan State University, studies novels in which characters are described as persons with autism or Asperger syndrome, and hopes to expand her dissertation research by focusing on a more substantial collection of works.³¹ Because the American Psychiatric Association added Asperger syndrome to the Diagnostic and Statistical Manual of Mental Disorders only in 1994, there are no out of copyright works she can study.³²

In addition, the inability to conduct TDM on most post-1925 texts “contributes to already existing racial and gender biases within the field” of digital humanities.³³ This is because “the

²⁶ *Id.*

²⁷ Letter from David Bamman, Appendix B, at 1.

²⁸ *Id.*

²⁹ Letter from Data-Sitters Club, Appendix D, at 1.

³⁰ Letter from David Bamman, Appendix B, at 1.

³¹ Letter from Jes Lopez, Appendix H, at 1–2.

³² *Id.* at 2.

³³ Letter from Melanie Walsh, Appendix O, at 2; *see also* Letter from ACH, Appendix A, at 3.

majority of available, digitized, pre-1925 texts are authored by white men.”³⁴ As one researcher explains, “public domain works on Project Gutenberg systematically overrepresent white, male authors.”³⁵ Thus, “the research questions it is able to answer again privilege that social group over others.”³⁶ Focusing TDM on pre-1925 texts “further reinscribes white men as the center of the field and further marginalizes women and people of color.”³⁷

For motion pictures, the situation is even more dire. As Professors Tilton and Arnold explain, virtually all of these works are in-copyright and protected by DRM.³⁸

d. The section 1201 prohibition is harmful to teaching.

In addition to its negative impact on research, the section 1201 prohibition harms teaching related to TDM. First, students are deterred from pursuing TDM because they cannot apply it to contemporary motion pictures and texts. As the Data-Sitters Club explains, incorporating computational analysis of modern literary works into curricula would serve the purpose of “enhancing students’ awareness of the possibility and limitations of digital methods, using material that is more familiar and resonant than the public domain.”³⁹ The inability to draw on more contemporary motion pictures and literary works also prevents those who teach TDM from making their syllabi more diverse and inclusive, as they can only include works that are reflective of authors published before 1925.⁴⁰ Second, the section 1201 prohibition deters students who want to work with TDM from pursuing fields where they would need to work with in-copyright works. As Dan Sinykin, Assistant Professor of English at Emory University, describes, “students turn to older periods where literature is not under copyright or to text they can acquire from the internet. So long as students do not pursue TDM in the field [of contemporary literature], the field will be stunted.”⁴¹

e. Existing alternatives to circumventing TPMs are inadequate.

There are no adequate alternatives to circumventing TPMs, for either motion pictures or literary works distributed electronically. Existing digital libraries are challenging to use, do not provide

³⁴ Letter from Melanie Walsh, Appendix O, at 2.

³⁵ Letter from David Bamman, Appendix B, at 1–2. Project Gutenberg is a free online collection of public domain books. Frequently Asked Questions About Project Gutenberg, <https://www.gutenberg.org/help/faq.html> (last visited Dec. 14, 2020).

³⁶ *Id.*

³⁷ Letter from Melanie Walsh, Appendix O, at 2.

³⁸ Letter from Lauren Tilton and Taylor Arnold, Appendix M, at 1.

³⁹ Letter from Data-Sitters Club, Appendix D, at 3.

⁴⁰ Letter from Melanie Walsh, Appendix O, at 2.

⁴¹ Letter from Dan Sinykin, Appendix L, at 2.

sufficient analytical tools, and have incomplete collections. Researchers seeking to build their own collection of works can rely on optical character recognition (“OCR”) (for text) or screen capture (for motion pictures), but these tools are too slow and expensive to be of practical use, and the data they generate is of insufficient quality.

i. Limits of existing digital libraries

First, some researchers wishing to conduct TDM on literary works can use pre-built collections of works.⁴² However, none of the researchers who submitted letters in support of this exemption request viewed any of these as a realistic option for TDM other than HathiTrust.⁴³ This is because it is unclear what these other collections contain, they tend not to have contemporary works, and researchers can only use algorithms provided by those who built the collection.⁴⁴

HathiTrust is a non-profit collaborative of academic and research libraries that preserves over 17 million items.⁴⁵ While HathiTrust is a valuable resource, “it is not a solution for all research questions.”⁴⁶ There are two major problems with HathiTrust: it is difficult to use, and there are substantial gaps in its collection.

HathiTrust requires researchers to use a secure “data capsule,” and all computing is carried out on servers at HathiTrust. However, this “limits computational processing to the capacity of the HathiTrust’s resources, which is occasionally outside the demands of contemporary state-of-the-art models” in natural language processing.⁴⁷ Beyond this, even experienced researchers report that the HathiTrust data capsule method is technically difficult to negotiate.⁴⁸ For example, one researcher reports that “the code I use to analyze works can only be refined outside the capsule, but can only be tested inside the capsule (with the ‘door’ to the outside world firmly shut). This means I often have to go in and out of the capsule as many as a hundred times before I have effective analytical tools.”⁴⁹

⁴² Letter from Andrew Piper, Appendix I, at 1.

⁴³ *Id.* at 1.

⁴⁴ Letter from Andrew Piper, Appendix I, at 1–2.

⁴⁵ HathiTrust Digital Library, *Welcome to HathiTrust*, <https://www.hathitrust.org/about> (last visited Dec. 14, 2020).

⁴⁶ Letter from David Bamman, Appendix B, at 2.

⁴⁷ *Id.*

⁴⁸ *See, e.g.*, Letter from Ted Underwood, Appendix N, at 2; Letter from Andrew Piper, Appendix I, at 2.

⁴⁹ *Id.*; *see also* Letter from Dan Sinykn, Appendix L, at 1 (describing HathiTrust as “cumbersome”).

Also, there are significant gaps in the HathiTrust collection. For example, “[t]he materials in the HathiTrust originate in university libraries, and so are necessarily biased toward academic monographs and away from, for example, the mass-market romances and science fiction more commonly found in city public libraries.”⁵⁰ Indeed, Dr. Wermer-Colan explains that, when he was in the process of ingesting the Temple Library collection into HathiTrust, he “identified that the vast majority of materials held in Temple Libraries’ Special Collection Research Center’s Paskow Science Fiction Collection are not contained in HathiTrust.”⁵¹ Also, for Ms. Lopez’s research into characters who are autistic or diagnosed with Asperger syndrome, “HathiTrust is unworkable because it doesn’t have an extensive collection of books from the young adult market, which a large portion of books with autistic characters targets.”⁵² Similarly, Ted Underwood, a professor in the School of Information Sciences at the University of Illinois at Urbana-Champaign, writes that he “would like to pursue questions about romance fiction, but the lack of such works in HathiTrust’s collection makes this impossible.”⁵³

Nor is it a solution to piece together collections across databases. As Dr. Wermer-Colan explains:

Databases are also siloed repositories, and research access for text mining to one database does not enable researchers to conduct research in a holistic process across multiple databases. As these databases develop proprietary corpora and text mining tools for their corpora, they continue to further silo researchers into limited available datasets. As a result, these balkanized databases both lack datasets representing the diversity of cultural production in the twentieth century by underrepresented groups and restrict researchers to relatively random sets of that data in each database.⁵⁴

Finally, HathiTrust is only available to a small subset of those individuals who would like to pursue TDM on in-copyright works. Professor Clawson is on the faculty of Grambling State University. While well-resourced institutions may have access to HathiTrust, “historically black

⁵⁰ Letter from David Bamman, Appendix B, at 2.

⁵¹ Letter from Henry Alexander Wermer-Colan, Appendix P, at 2.

⁵² Letter from Jes Lopez, Appendix H, at 2.

⁵³ Letter from Ted Underwood, Appendix N, at 1. Given how HathiTrust is structured, there is no way for researchers to correct the gaps in its dataset. Letter from Hoyt Long, Appendix G, at 2–3.

⁵⁴ Letter from Henry Alexander Wermer-Colan, Appendix P, at 2; *see also* Letter from Matthew Jockers, Appendix F, at 3 (explaining that the metadata used to differentiate sections within a book is inconsistent across different databases).

colleges like Grambling are unlikely to participate in this kind of partnership” because of the expense involved.⁵⁵

For motion pictures, there are no large-scale digital libraries available to researchers who want to conduct TDM on these sorts of works.⁵⁶ These researchers have no choice but to create their own collections of works.

ii. Limits on creating one’s own collection

Another potential option is for researchers to create their own collections of works to analyze. However, the two available methodologies—OCR for texts and screen capture for motion pictures—are inadequate for TDM.

a) Limits of OCR

For researchers who work with text, one method to create a collection of works is to scan the works and use OCR to create machine-readable text. Many scholars report that, for TDM research, OCR is not a feasible path forward because of the time and labor required to scan books and correct OCR errors for the number of books that TDM requires.⁵⁷

ACH surveyed its members about how access to in-copyright materials has affected their work in general, particularly the impact of section 1201. One hundred percent of survey respondents agreed with the statement that it was “financially unfeasible for them to pay someone to scan and perform optical character recognition (OCR) in order to transform all of the books that they would like to use into digital copies.”⁵⁸

Dr. Wermer-Colan describes using OCR to create a collection for TDM as a slow and labor-intensive process:

During my postdoctoral fellowship, I have worked in Temple Libraries to digitize twentieth-century literary texts. This process involved collaborating with the

⁵⁵ Letter from James Clawson, Appendix C, at 2; *cf.* Letter from Dan Sinykin, Appendix L, at 1 (stating he can access HathiTrust “because Emory [University] is a partner institution”).

⁵⁶ Letter from David Bamman, Appendix B, at 2.

⁵⁷ Letter from ACH, Appendix A, at 2 (explaining that one survey respondent wrote: “I am running a three-year project that is only possible because I started digitizing the corpus more than a year before the project start date.”); Letter from David Bamman, Appendix B, at 2 (describing laborious and “errorful” process of OCRing 500 in-copyright books); Letter from James Clawson, Appendix C, at 1 (describing challenges of OCR); Letter from Hoyt Long, Appendix G, at 3 (describing spending several years digitizing about 1,000 novels to support the Black Writing Project); Letter from Ted Underwood, Appendix N, at 2.

⁵⁸ Letter from ACH, Appendix A, at 1–2.

Digitization and Metadata Services department to purchase twentieth-century canonical novels, de-bind the books, scan them through a sheet-feed scanner, convert the scanned images through Optical Character Recognition software into machine-readable text, and then, using automated and manual methods, fix errors produced during the conversion into machine-readable text as well as clean those texts of paratextual information such as copyright statements and other front matter. This process is so laborious that even at a R1 university (doctoral degree-granting institutions with very high research activity) like Temple, with multiple library departments and half-dozen staff and student workers contributing to various stages of the project, we have only been able to digitize approximately 500 books over a three-year period.⁵⁹

For certain kinds of texts, OCR is a particularly poor fit. OCR is, for example, often a poor fit for scholars who study certain non-English texts.⁶⁰ As Hoyt Long, Associate Professor of Japanese Literature and Director of Graduate studies in the East Asian Languages and Civilizations Department at the University of Chicago, explains, “[i]deographic languages, such as Japanese, are particularly difficult for OCR software.”⁶¹ This is because “[i]deographic languages tend to have many more characters and greater geometric complexity and variation than written English; there are several thousand characters currently in use in the Japanese language, and this number increases to tens-of-thousands [sic] as one goes further back in time.”⁶² OCR also faces challenges with “handwriting-style fonts, which OCR handles very poorly and need to be transcribed manually.”⁶³

OCR is additionally simply out of reach for many early-career TDM scholars, who lack their own research budgets and may operate under compressed timeframes. As Ms. Lopez describes, “scanning printed books, running OCR, and correcting the output is a time-consuming process, and doing so beyond a handful of books would be untenable, especially for a graduate student like myself who needs to complete research on a short time scale and with limited resources.”⁶⁴

Even small error rates with OCR are problematic. As Professor Long explains, “[g]iven that the ultimate goal of TDM is to analyze the text files generated through OCR, often at the level of

⁵⁹ Letter from Henry Alexander Wermer-Colan, Appendix P, at 1.

⁶⁰ Letter from Hoyt Long, Appendix G, at 2.

⁶¹ *Id.*

⁶² *Id.*

⁶³ Letter from Data-Sitters Club, Appendix D, at 2.

⁶⁴ Letter from Jes Lopez, Appendix H, at 2; *see also* Letter from Melanie Walsh, Appendix O, at 2 (“[A]s an early career researcher, I do not have my own research funds, and I do not have my own undergraduate or graduate student advisees. The fact that OCR is more expensive and time-consuming than breaking DRM is not merely an inconvenience. It actively prevents me from researching and teaching about literary culture after 1925.”).

individual words, researchers diligently ensure that their text files are error-free. Otherwise, analysis will reflect errors introduced by OCR, preventing accurate insight into the work.”⁶⁵

Further, OCR does not work at all when the goal of a project is to study data that is “embedded in a digitized file (such as provenance or information about the devices used to create the work).”⁶⁶ For example, ACH described one researcher’s desire to “study the provenance of files where it appeared that multiple versions of an ebook were for sale from the same bookseller. This is difficult to do/impossible without insight into the code of the file.”⁶⁷

b) Limits of screen capture

Researchers examining motion pictures have no choice but to create their own collections to examine. Screen capture is the only possibility known to those writing declarations in support of this petition, but it is very slow and degrades the quality of the images.

Having to rely on screen capture has significantly impeded use of TDM on motion pictures. For example, Professor Bamman has sought to create a dataset of motion pictures to determine whether it is possible to measure directorial style in movies.⁶⁸ In other words, are there quantifiable and observable aspects of films that can be used to identify who directed the film? To pursue the project, Professor Bamman’s goal was to create a dataset of 10,000 movies.⁶⁹ While it would have been possible to buy DVDs and circumvent the TPMs, because of section 1201’s prohibition on doing so, he instead bought the movies, played them on a computer, and used screen-capture tools to record the screen while the movie was playing in real time.⁷⁰ There are several significant drawbacks to this approach. First, screen capture “is an imperfect process that necessarily loses important information—the subsequent data is of lower resolution than the original medium, and important structural markers like chapter boundaries are lost.”⁷¹ Second, it would have required too much time: “If a human operator were present for the duration of the screen capture for each movie with an average run time of 2 hours (and worked 8 hours a day, 5 days a week, 50 weeks per year) it would take 10 years to complete the act of digitization alone.”⁷²

⁶⁵ Letter from Hoyt Long, Appendix G, at 2.

⁶⁶ Letter from ACH, Appendix A, at 2.

⁶⁷ *Id.*

⁶⁸ Letter from David Bamman, Appendix B, at 2–3.

⁶⁹ *Id.* at 2.

⁷⁰ *Id.* at 3.

⁷¹ *Id.*

⁷² *Id.*

f. Granting this exemption would allow valuable research to move forward.

This comment is supported by numerous declarations describing research projects in both classes that will be able to go forward if this exemption is granted. For example:

- Professor Bamman writes that if an exemption were granted, he would take up his research into directorial style. In addition, he would “begin examining other ways in which movies can function as an object of study in their own right—questions worth examining include historical trends in film over the past century (has the depiction of violence within movies become more or less predominant?), imitation and influence (can we trace which directors have had the greatest impact on the visual style of directors after them?) and reception (which specific aspects of film do audiences, critics, and the box office most respond to?).”⁷³
- Professor Clawson writes, “If an exemption is granted, . . . I would direct part of Grambling’s upcoming course in text analytics around a single, large corpus of African American writers, many of whose works are often underrepresented in available archives.”⁷⁴
- Eric Hoyt, Professor of Media Production in the Department of Communication Arts at the University of Wisconsin-Madison, writes that if the exemption were granted, he would be interested in building a collection of high-quality television and film materials to examine the evolution of televisual and filmic styles, looking at questions such as how uses of film color, editing patterns, and the duration of close-ups have changed over time.⁷⁵
- Matthew Jockers, Dean of the College of Arts and Sciences and Professor of English and Data Analytics at Washington State University, has provided a list of 17 research questions he would like to investigate if this exemption is granted. For example, Dean Jockers writes he would research “whether literary trends correlate with historical events.”⁷⁶
- Professor Sinykin is working on a book on the conglomeration of the United States publishing industry. Whereas in 1950 almost every publisher was independent, by the year 2000 only six multinational media conglomerates controlled most of publishing. His research project seeks to understand how conglomeration changed fiction. Detecting patterns of change across thousands of novels requires application of TDM. Professor Sinykin writes, “If I were exempt from 1201, I would be able to write a better, truer book about [publishing] conglomeration.”⁷⁷ But he adds that “[m]aybe more profoundly, I

⁷³ Letter from David Bamman, Appendix B, at 3.

⁷⁴ Letter from James Clawson, Appendix C, at 2.

⁷⁵ Letter from Eric Hoyt, Appendix E, at 1–2.

⁷⁶ Letter from Matthew Jockers, Appendix F, at 2.

⁷⁷ Letter from Dan Sinykin, Appendix L, at 2.

would be able to teach TDM to the next generation of scholars who would transform our field of study.”⁷⁸

- Andrew Piper, Professor in the Department of Languages, Literatures, and Cultures at McGill University, writes “our number one priority is to build a global collection of literature across a broad array of national and linguistic cultures . . . Our dream project is to begin to understand how these different cultures tell stories—where are the fault lines that make one narrative world different from another and where are the lines of commonality, that illustrate for us a common human approach to storytelling?”⁷⁹
- Dr. Walsh writes: “I am eager to compare collections of novels published by authors who graduated from different MFA writing programs.”⁸⁰

There is good reason to believe that many researchers other than those who submitted letters will take advantage of the exemption to produce research if the exemption is granted. Both the ACH and scholarly publishing experts submitted letters demonstrating that a broad range of researchers are interested in pursuing research that is currently blocked by section 1201.⁸¹

2. Overview of Comment

The proposed exemption should be granted because researchers are currently experiencing substantial adverse effects on their ability to make non-infringing uses of motion pictures and literary works distributed electronically.

Section 1201 seriously inhibits TDM research. Courts have recognized that copying motion pictures and literary works distributed electronically to create a collection on which to conduct TDM is a highly transformative fair use and is therefore noninfringing. Section 1201’s prohibition on circumventing access controls adversely affects TDM researchers. The prohibition deters researchers from applying the technique to in-copyright motion pictures and literary works distributed electronically, despite the fact that doing so would be a fair use. As a result, researchers are rarely able to use this technique to understand contemporary motion pictures or literary works. This deprives the public of the knowledge that could be derived about motion pictures and literary works created during most of the 20th century and all of the 21st century. In addition, it replicates biases because it limits the study of these classes of works to a time when authorship was far less diverse.

Section 1201’s statutory factors weigh in favor of an exemption. First, an exemption for TDM research will increase the availability of copyrighted works because it will increase the production of articles and research based on insights revealed using TDM techniques. At the

⁷⁸ *Id.*

⁷⁹ Letter from Andrew Piper, Appendix I, at 4.

⁸⁰ Letter from Melanie Walsh, Appendix O, at 1.

⁸¹ Letter from ACH, Appendix A, at 1; Letter from Rachael Samberg and Timothy Vollmer, Appendix K, at 2.

same time, an exemption limited to circumventing TPMs to facilitate TDM research is unlikely to deter creation of copyrighted works, just as other narrow exemptions have not deterred creation of copyrighted works. Second, an exemption for TDM will lead to greater availability of works for nonprofit archival, preservation, and educational purposes. In particular, the development of educational materials about TDM will greatly increase, and a broader range of researchers will be interested in learning the technique because they will be able to apply it to works relevant to their fields. Third, the removal of the threat of liability will lead to more research and scholarship, specifically, to work focusing on in-copyright motion pictures and literary works. Currently, the prohibition on circumvention impedes research and scholarship, limiting the conclusions that can be drawn and the development of the discipline. Fourth, allowing an exemption for deployment of TDM will not harm the market for copyrighted works or their value. No one would read an article presenting the results of TDM of a collection of motion pictures or literary works in lieu of watching or reading the works themselves. Fifth, an additional factor favors the exemption. Creating an exemption for TDM would promote equity by allowing the technique to be performed on a more diverse and inclusive set of works.

Finally, it is the prohibition on circumvention that is the source of the adverse effects TDM researchers are experiencing, rather than other causes. Alternatives, such as existing digital libraries, OCR, and screen capture, are insufficient.

Granting the exemption commenters request would allow intellectual curiosity, rather than copyright infringement liability, to guide TDM researchers, which would in turn “promote the progress of science and the useful arts” by enabling researchers to discover and share new insights and contribute to public knowledge.⁸²

Item D. Technological Protection Measures and Methods of Circumvention

The technological protection measures and methods of circumvention at issue for this proposed exemption include those measures and methods applicable to motion pictures and literary works distributed electronically.

1. Motion Pictures

Researchers require access to high-quality motion picture files to conduct TDM research, but access controls block this critical access. Therefore, these TPMs restrict the ability of TDM researchers to conduct their analysis on motion pictures.

⁸² U.S. Const. art. I, § 8.

Motion pictures continue to be distributed to consumers on DVDs.⁸³ CSS is used to encrypt DVDs, limiting their access and use.⁸⁴ As the Register has previously explained, “[t]he vast majority of DVDs use the Content Scramble System (“CSS”) to encrypt audiovisual works on DVDs using a fixed set of decryption keys.”⁸⁵ The Register and courts have found that “CSS is an ‘access control’ within the meaning of section 1201(a)(1).”⁸⁶ CSS was first decoded in 1999 and it is now possible to circumvent CSS protections using a variety of readily available software.⁸⁷

In addition to DVDs, motion pictures are also commonly distributed on Blu-ray discs.⁸⁸ The Register has recognized that “Blu-ray discs are protected primarily by the Advanced Access Content System (“AACS”), which allows vendors to revoke compromised keys and distribute new keys.”⁸⁹ In 2012, the Register recognized AACS as a TPM subject section 1201’s anti-circumvention provisions.⁹⁰ A limited number of keys necessary to decrypt Blu-ray discs have been discovered, enabling AACS to be bypassed for many Blu-ray discs, and various software programs exist that can circumvent AACS protection on Blu-ray discs.⁹¹

⁸³ See, e.g., Stephane Prange, *NPD: DVD and Blu-ray Player Sales Jump in Pandemic*, Media Play News (Apr. 30, 2020), <https://www.mediaplaynews.com/npd-dvd-and-blu-ray-player-sales-jump-in-pandemic/>.

⁸⁴ Frequently Asked Questions and Answers, DVD Copy Control Ass’n, <http://www.dvdcca.org/home/faq> (last visited Dec. 10, 2020) (explaining that DVDs are encrypted, and a CSS license is needed to enable consumers to access DVD content).

⁸⁵ U.S. Copyright Off., Recommendation of the Reg. of Copyrights on Section 1201 Rulemaking: Sixth Triennial Proc. to Determine Exemptions to the Prohibition on Circumvention (Oct. 8, 2015), <https://cdn.loc.gov/copyright/1201/2015/register-recommendation.pdf>, at 29 [hereinafter 2015 Recommendation].

⁸⁶ 2015 Recommendation at 29 (citing *DVD Copy Control Ass’n, Inc. v. Bunner*, 116 Cal. App. 4th 241, 255 (2004)).

⁸⁷ *Id.* (“The CSS key was decoded in 1999, and decryption software is now available on the internet, including the programs MactheRipper, DVDDecrypter, and Handbrake.”).

⁸⁸ See, e.g., *Sales Report for Week Ended 10-24-20*, Media Play News, <https://www.mediaplaynews.com/research/sales-report-for-week-ended-10-24-20/> (last visited Dec. 10, 2020) (showing that Blu-ray disc sales are comparable to sales of DVDs).

⁸⁹ 2015 Recommendation at 29.

⁹⁰ U.S. Copyright Off., Recommendation of the Reg. of Copyrights on Section 1201 Rulemaking: Fifth Triennial Proc. to Determine Exemptions to the Prohibition on Circumvention (Oct. 12, 2012), https://cdn.loc.gov/copyright/1201/2012/Section_1201_Rulemaking_2012_Recommendation.pdf, at 126 [hereinafter 2012 Recommendation].

⁹¹ Fred von Lohmann, *09 f9: A Legal Primer*, Elec. Frontier Found. (May 2, 2007), <https://www.eff.org/deeplinks/2007/05/09-f9-legal-primer>.

Some Blu-ray discs, specifically those with Ultra HD content, are also protected by AACS2.⁹² AACS and AACS2 are sufficiently similar that the justifications for permitting circumvention of AACS also apply to AACS2.⁹³ Some motion pictures now appear in Ultra HD,⁹⁴ and AACS2 does for Ultra HD content what AACS did for standard Blu-ray content.⁹⁵

Finally, motion pictures are also commonly distributed through digital transmissions. Much like CSS and AACS, the protection measures found on digitally transmitted video seek to control access through encryption and other mechanisms, and thus qualify as a TPM within the meaning of section 1201(a)(3) by requiring the “application of information”—namely, encryption keys—in order to gain access to the work.⁹⁶ The Register reached the same conclusion in her 2015 recommendation, determining that a “significant number of platforms that offer digitally transmitted motion pictures, both for digital downloads and for streaming, constitute technological measures controlling access to those works under § 1201(a)(1).”⁹⁷ The same is true

⁹² U.S. Copyright Off., Comment of Advanced Content Sys. Licensing Admin’r on a Proposed Exemption Under 17 U.S.C. § 1201 (Feb. 12, 2018) https://cdn.loc.gov/copyright/1201/2018/comments-021218/class1/Class_01_Opp'n_AACS_LA.pdf, at 2. While the Register has previously declined to create an exemption that allows circumvention of AACS2, that appears to be because proponents did not identify AACS2 as a TPM requiring circumvention to eliminate the adverse effects of the section 1201 prohibition, and because they did not argue that AACS and AACS2 were sufficiently similar that the exemption for AACS already encompassed AACS2. U.S. Copyright Off., Recommendation of the Acting Reg. of Copyrights on Section 1201 Rulemaking: Seventh Triennial Proc. to Determine Exemptions to the Prohibition on Circumvention (Oct. 26, 2018) https://cdn.loc.gov/copyright/1201/2018/2018_Section_1201_Acting_Registers_Recommendation.pdf, at 40 [hereinafter 2018 Recommendation]. This comment therefore explicitly identifies Ultra HD content as a type of content necessary to access to avoid adverse effects, and AACS2 as a TPM that blocks access to this content.

⁹³ U.S. Copyright Off., Reply Comments of Joint Educators et al. (Mar. 14, 2018) https://cdn.loc.gov/copyright/1201/2018/comments-031418/class1/Class_01_Reply_Joint_Educators.pdf, at 9 [hereinafter Reply Comments of Joint Educators] (“Compared to other TPMs, AACS2 is different in form but it is fundamentally the same in function. It is an access control that is currently employed or is likely to be employed by copyright owners in the next three years to prevent users from accessing a copyrighted work.”).

⁹⁴ See, e.g., Ty Pendlebury, *Best 4K Blue-rays*, CNET (Nov. 26, 2020 7:11 AM), <https://www.cnet.com/pictures/best-4k-blu-rays/>.

⁹⁵ Reply Comments of Joint Educators at 9.

⁹⁶ See 2015 Recommendation at 8.

⁹⁷ *Id.* at 69.

today.⁹⁸ The Register has previously recognized that decryption tools are “widely available” for motion pictures distributed through digital transmissions.⁹⁹

2. Literary Works Distributed Electronically

Similar to researchers who study motion pictures, TDM researchers studying literary works need access to high quality text files, and the existence of TPMs on e-books restricts their ability to use these files to conduct research on literary works. Literary works continue to be distributed as e-books.¹⁰⁰ The Register has previously recognized that “many e-books are protected by TPMs that interfere with the proper operation of”¹⁰¹ proposed uses, for example the proper operation of assistive technologies for persons with disabilities.¹⁰² Therefore the Librarian has previously adopted exemptions in previous rulemaking to allow circumvention of these TPMs.¹⁰³

Item E. Asserted Adverse Effects on Noninfringing Uses

Individuals who wish to apply TDM to motion pictures and literary works distributed electronically are adversely affected by the prohibition on circumvention in their ability to make noninfringing uses of these works. As required by the Notice of Proposed Rulemaking, this comment demonstrates that: (1) the proposed classes include at least some copyrighted works, (2) copying motion pictures and literary works distributed electronically to create a collection on which to conduct TDM is a noninfringing use, (3) researchers are adversely affected in their ability to make such noninfringing uses, and (4) the statutory prohibition on circumventing access controls is the cause of the adverse effects.¹⁰⁴

⁹⁸ See, e.g., U.S. Copyright Off., Comment of Authors All. et al. app. A Memorandum from Alex Podobas (Dec. 18, 2017) <https://cdn.loc.gov/copyright/1201/2018/comments-121817/class1/class-01-initialcomments-authors-alliance-et-al.pdf>, at app. A (describing the technological protection measures applied to digital content distribution services).

⁹⁹ 2015 Recommendation at 93.

¹⁰⁰ See Michael Kozlowski, *EBook Sales Are Undergoing a Revival in 2020*, Good eReader (May 9, 2020), <https://goodereader.com/blog/e-book-news/ebook-sales-are-undergoing-a-revival-in-2020>.

¹⁰¹ 2015 Recommendation at 128.

¹⁰² *Id.*

¹⁰³ *Id.*

¹⁰⁴ Notice of Proposed Rule Making, 85 Fed. Reg. 65293, 65294 (Oct. 15, 2020) <https://www.govinfo.gov/content/pkg/FR-2020-10-15/pdf/2020-22893.pdf> (citation omitted) [hereinafter 2020 NPRM].

1. The proposed classes include works protected by copyright.

The proposed classes include at least some works protected by copyright. The Register has already recognized that many literary works distributed electronically¹⁰⁵ and motion pictures¹⁰⁶ are protected by copyright. Many TDM research projects depend on the ability to access works currently in copyright. For researchers interested in studying television shows or movies, nearly all works of interest are under copyright.¹⁰⁷ This is also true for researchers interested in contemporary literature.¹⁰⁸

2. Creating collections of copyrighted works for TDM is a noninfringing use.

The second element is whether the use at issue is noninfringing under title 17.¹⁰⁹ The Copyright Act specifies that the “fair use” of copyrighted works, including for such purposes as “criticism, comment, news reporting, teaching . . . scholarship, or research, is not an infringement of copyright.”¹¹⁰ In *Authors Guild v. Google (Google Books)*, the Second Circuit concluded that creating digital copies of books to build a full-text searchable database of those books was a fair use, as was creating a tool known as “ngrams,” which allowed members of the public to determine the frequency of usage of particular words over time.¹¹¹ The Second Circuit reached this conclusion because the database and the ngram tool made “available information *about* Plaintiffs’ books without providing the public with a substantial substitute for matter protected by the Plaintiffs’ copyright interests in the original works or derivatives of them.”¹¹² Creating collections of motion pictures and literary works for the purpose of allowing TDM is a fair use for the same reason: it allows researchers to derive information *about* books without providing the public with a substitute for them.

a. The *Google Books* case

In *Google Books*, the Second Circuit addressed whether various aspects of Google’s Library Project were noninfringing fair uses.¹¹³ Through the Library Project, major research libraries submitted books from their collections to Google, which “ma[de] a digital scan of each book, extract[ed] a machine-readable text, and create[d] an index of the machine-readable text of each

¹⁰⁵ 37 C.F.R. § 201.40(b)(1)(i)(C).

¹⁰⁶ *Id.* at (ii).

¹⁰⁷ Letter from Lauren Tilton and Taylor Arnold, Appendix M, at 1.

¹⁰⁸ Letter from Melanie Walsh, Appendix O, at 1.

¹⁰⁹ 2020 NPRM at 65294.

¹¹⁰ 17 U.S.C. § 107.

¹¹¹ *Google Books*, 804 F.3d at 207–08.

¹¹² *Id.* at 207 (emphasis in original).

¹¹³ *Id.* at 208.

book.”¹¹⁴ The digital collection Google created served as the basis for the Google Books search engine.¹¹⁵ The Google Books search engine allowed members of the public “to identify those books, out of millions, that do, as well as those that do not, use the terms selected by the researcher.”¹¹⁶ The “identifying information instantaneously supplied would otherwise not be obtainable in lifetimes of searching.”¹¹⁷

In addition to the Google Books search engine, Google also made available its ngrams research tool, which drew on the same digital collection to provide researchers with statistical information about the frequency of word and phrase usage over time.¹¹⁸ The tool “permits users to discern fluctuations of interest in a particular subject over time and space by showing increases and decreases in the frequency of reference and usage in different periods and different linguistic regions.”¹¹⁹ Further, it “allows researchers to comb over the tens of millions of books Google has scanned in order to examine ‘word frequencies, syntactic patterns, and thematic markers’ and to derive information on how nomenclature, linguistic usage, and literary style have changed over time.”¹²⁰

Authors of published works that had been scanned and included in the digital collection sued Google, which defended on fair use grounds.¹²¹ The Second Circuit carefully considered each of the four statutory fair use factors and concluded that all challenged aspects of Google’s Library Project were fair use.

The court first considered the “purpose and character” of the secondary work.¹²² It approached this inquiry by asking whether Google’s use was “transformative,” which, relying on *Campbell*

¹¹⁴ *Id.*

¹¹⁵ *Id.*

¹¹⁶ *Id.* at 209.

¹¹⁷ *Id.*

¹¹⁸ *Id.*

¹¹⁹ *Id.*

¹²⁰ *Id.* at 209 (quoting *Authors Guild, Inc. v. Google, Inc.*, 954 F. Supp. 2d 282, 287 (S.D.N.Y. 2013)). In addition, the Google Books search function allowed limited viewing of text, specifically, the display of up to three “snippets” from a single book containing terms a user had queried. *Id.* at 209-10. Because the use covered by the exemption does not include displaying snippets of text to the general public, this comment does not discuss this part of the *Google Books* opinion in detail. Although TDM researchers may use copyrighted works in their publications and other research outputs in a manner that is consistent with fair use, their ability to exercise this right is not dependent upon this exemption.

¹²¹ *Id.* at 211.

¹²² *Id.* at 214.

v. Acuff-Rose, it defined as a work that “adds something new, with a further purpose.”¹²³ The court relied heavily on its earlier decision in *Authors Guild, Inc. v. HathiTrust (HathiTrust)*, in which authors brought copyright claims against HathiTrust, an entity created by libraries participating in Google Books that pooled together the digital copies of their books made for them by Google.¹²⁴ Among other uses, authors whose books had been scanned challenged HathiTrust’s creation of a database that permitted full-text searches of the collection of works, allowing users to locate books in which particular words or phrases appeared.¹²⁵ The *HathiTrust* court had concluded that “the creation of a full-text searchable database is a quintessentially transformative use . . . [as] the result of a word search is different in purpose, character, expression, meaning, and message from the page (and the book) from which it is drawn.”¹²⁶ In a similar vein, the court in *Google Books* reasoned:

As with *HathiTrust* [], the purpose of Google’s copying of the original copyrighted books is to make available significant information *about those books*, permitting a searcher to identify those that contain a word or term of interest, as well as those that do not include reference to it. In addition, 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 . . . strongly favoring satisfaction of the first factor.¹²⁷

Thus, creating a collection of works to permit identification of information about those works is a “purpose and character” that favors fair use. It did not matter that Google had a commercial motivation. The court explained that while “in some circumstances, a commercial motivation on the part of the secondary user will weigh against her,” this would primarily be the case when “a persuasive transformative purpose is lacking.”¹²⁸ Given Google’s “highly convincing transformative purpose,” the court did not count Google’s commercial motivation against it.¹²⁹

Turning to the second statutory factor, the court considered the “nature of the copyrighted work.”¹³⁰ The court noted that “[t]he second factor has rarely played a significant role in the determination of fair use disputes.”¹³¹ The court did not believe it was relevant whether the

¹²³ *Id.* at 215 (quoting *Campbell v. Acuff-Rose*, 510 U.S. 569, 578–79 (1994)).

¹²⁴ *Id.* at 217 (relying on *HathiTrust*, 755 F.3d at 98).

¹²⁵ *Id.*

¹²⁶ *Id.* at 217 (quoting *HathiTrust*, 755 F.3d at 97).

¹²⁷ *Id.* at 217 (emphasis in original).

¹²⁸ *Id.* at 219.

¹²⁹ *Id.*

¹³⁰ *Id.* at 220.

¹³¹ *Id.*

books that had been copied were factual or creative, and concluded instead that “the second factor favors fair use not because Plaintiffs’ works are factual, but because the secondary use transformatively provides valuable information about the original, rather than replicating protected expression in a manner that provides a meaningful substitute for the original.”¹³²

As for the third statutory factor, “the amount and substantiality of the portion used in relation to the copyrighted work as a whole,” the court explained that copying an entire work “has repeatedly been found justified as fair use when the copying was reasonably appropriate to achieve the copier’s transformative purpose and was done in such a manner that it did not offer a competing substitute for the original.”¹³³ In *Google Books*, the court concluded that copying the whole of a work was permissible because if “Google copied less than the totality of the originals, its search function could not advise searchers reliably whether their searched term appears in a book (or how many times).”¹³⁴ The court also found it significant that “[w]hile Google *makes* an unauthorized digital copy of the entire book, it does not reveal that digital copy to the public.”¹³⁵

Finally, as for the fourth statutory factor, the effect on the potential market for or value of the copyrighted work, the court explained that, in general, “the more the copying is done to achieve a purpose that differs from the purpose of the original, the less likely it is that copy will serve as a satisfactory substitute for the original.”¹³⁶ The court drew a contrast between *HathiTrust*, where there was no substitution effect at all because its search tool did not display any text from works, and the Google Books search engine, which did display some short passages (“snippets”) of the text of works. However, it concluded that the snippets were too insubstantial to cause market harm.¹³⁷

Thus, weighing the factors together, the court concluded that “Google’s making of a complete digital copy of Plaintiffs’ works for the purpose of providing the public with its search and snipped view functions . . . is a fair use.”¹³⁸

b. The proposed exemption is a fair use for the reasons articulated in *Google Books*

Proponents seek an exemption to allow copying of motion pictures and literary works distributed electronically to create a collection of works on which to conduct TDM. This use is a fair use for the same reasons the uses considered in *Google Books* were fair uses: the collections will make “available information *about* . . . books [and motion pictures] without providing the public with a

¹³² *Id.* (quoting 17 U.S.C. § 107(3)).

¹³³ *Id.* at 221.

¹³⁴ *Id.*

¹³⁵ *Id.*

¹³⁶ *Id.* at 223.

¹³⁷ *Id.* at 224–25.

¹³⁸ *Id.* at 225.

substantial substitute for matter protected by the . . . copyright interests in the original works or derivatives of them.”¹³⁹

The first fair use factor is the “purpose and character” of the use.¹⁴⁰ Here, as in *Google Books*, the proposed use is highly transformative. Proponents seek to create collections of works for TDM analysis. This is “something new, with a further purpose.”¹⁴¹ Much like the ngrams tool in *Google Books*, TDM is used to provide information “about” works, rather than providing the works themselves. For example, TDM has been used to “observe trends such as the marked decline in fiction written from a first-person point of view that took place from the mid-late 1700s to the early-mid 1800s, the weakening of gender stereotypes, and the staying power of literary standards over time.”¹⁴² A study of the decline in fiction written from a first-person point of view has a different “purpose, character, expression, meaning, and message”¹⁴³ than the book (or books) from which it is drawn. Moreover, although proponents’ exemption would apply to commercial and non-commercial uses, this does not weigh against proponents because they are pursuing a “highly convincing transformative purpose,”¹⁴⁴ specifically, research helping us to better understand motion pictures, literary works, and their bearing on society.

The second fair use factor is the “nature of the copyrighted work.”¹⁴⁵ Here, as in *Google Books*, this factor is not very significant.¹⁴⁶ That said, the factor may favor fair use because the proposed exemption “transformatively provides valuable information about the original, rather than replicating protected expression in a manner that provides a meaningful substitute for the original.”¹⁴⁷ For example, Dr. Walsh’s proposed study comparing collections of novels published by authors who graduated from different MFA writing programs would provide information about the influence of MFA writing programs, rather than serving as a substitute for reading one of the novels published by authors who graduated from these programs.¹⁴⁸

¹³⁹ *Id.* at 207 (emphasis in original).

¹⁴⁰ 17 U.S.C. § 107(1).

¹⁴¹ *Campbell*, 510 U.S. at 578–79.

¹⁴² Letter from Ted Underwood, Appendix N, at 1.

¹⁴³ *HathiTrust*, 755 F.3d at 97.

¹⁴⁴ *Google Books*, 804 F.3d at 219; see also *A.V. v. iParadigms Liab. Co.*, 544 F. Supp. 2d 473, 482 (E.D. Va. 2008) (finding a use to be transformative where “Plaintiffs originally created and produced their works for the purpose of education and creative expression. iParadigms . . . uses the papers for an entirely different purpose, namely, to prevent plagiarism and protect the students’ written works from plagiarism.”).

¹⁴⁵ 17 U.S.C. § 107(2).

¹⁴⁶ *Google Books*, 804 F.3d at 220.

¹⁴⁷ *Id.*

¹⁴⁸ Letter from Melanie Walsh, Appendix O, at 1.

The third fair use factor is “the amount and substantiality of the portion used in relation to the copyrighted work as a whole.”¹⁴⁹ As in *Google Books*, the fact that TDM requires copying entire works is consistent with fair use. Just as Google had to copy the totality of works to “advise searchers reliably whether their searched term appears in a book (or how many times),”¹⁵⁰ so too must TDM collections include entire works if researchers are to draw accurate conclusions. For example, Ms. Lopez’s study of novels in which characters are described as persons with autism or Asperger syndrome would not be accurate if only partial works were available. Additionally, just as in *Google Books*, although TDM requires copying entire works, those digital copies are not revealed to the public.

The fourth fair use factor assesses the use’s impact on “the potential market for or value of the copyrighted work.”¹⁵¹ Here, as in *Google Books*, the creation of a collection of works to draw conclusions about those works does not diminish the market for or value of the copyrighted works.¹⁵² The creation of a collection of works for TDM will not cause market harm to the original works contained in the collection. No one seeking to watch a motion picture would instead read a computational analysis of that motion picture. No one seeking to read an e-book would instead read a TDM research paper that rests its conclusions in part on computational analysis of the text of that novel.

To be sure, some major publishers license collections for TDM purposes.¹⁵³ However, those offerings do not constitute evidence of cognizable market harm because the use publishers are seeking to license is a plainly transformative fair use. A “copyright holder cannot prevent others from entering fair use markets merely ‘by developing or licensing a market for parody, news reporting, educational, or other transformative uses of its own creative work.’”¹⁵⁴ Similarly, the relevant question is whether works resulting from TDM compete in the marketplace for the *original* work, not whether allowing others to compete in the marketplace of TDM corpora may result in competition for publishers also offering such services.

Rights holders may argue that creating an exemption for the creation of collections for TDM poses the risk that unauthorized users might gain access and make the copyrighted works in the

¹⁴⁹ 17 U.S.C. § 107(3).

¹⁵⁰ *Google Books*, 804 F.3d at 221; *see also A.V. v. iParadigms Liab. Co.*, 544 F. Supp. 2d 473, 483 (E.D. Va. 2008) (“[I]t is clear that iParadigms uses the entirety of the original works. In order to be successful in its plagiarism detection services, it must.”).

¹⁵¹ *HathiTrust*, 755 F.3d at 99 (quoting 17 U.S.C. § 107(4)).

¹⁵² *Google Books*, 804 F.3d at 224–25.

¹⁵³ *See, e.g.*, Gale Primary Sources Platform, <https://www.gale.com/primary-sources/platform> (last visited Dec. 14, 2020).

¹⁵⁴ *Bill Graham Archives v. Dorling Kindersley Ltd.*, 448 F.3d 605, 614–15 (quoting *Castle Rock Entm’t Inc. v. Carol Publ’g Grp.*, 150 F.3d 132, 145 n.11 (2d Cir. 1998)).

collection widely available. But this concern is unfounded. Higher education institutions and other research institutes are extraordinarily well equipped to help researchers secure their data. This makes sense, as these institutions deal with sensitive data all the time. Medical centers collect patient data, social scientists collect human subjects data, and researchers from many disciplines generate data on sensitive research topics that must be safeguarded. Academic researchers license data from private sources routinely. University libraries negotiate licenses for access to collections of publications, and these publications must be secured. Higher education institutions thus already have the technological infrastructure in place to protect TDM collections. They also have the institutional infrastructure. They commonly offer research data management programs specifically designed to help researchers develop data management plans and keep their data secure.¹⁵⁵ Just as the court found in *HathiTrust*, there is “no basis . . . on which to conclude that a security breach is likely to occur, much less one that would result in the public release of the specific copyrighted works” in the collection.¹⁵⁶

Finally, TDM research may actually increase demand in the source work. For example, Professor Bamman noted he has purchased DVDs for his research project.¹⁵⁷ Likewise, the Data-Sitters Club was willing to pay for copies of e-books for their TDM research instead of buying second-hand print copies for the laborious OCR process.¹⁵⁸

In summary, creating collections of motion pictures and literary works for the purpose of allowing TDM is a noninfringing fair use. It will allow researchers to gather information *about* the classes of works they collect but will not provide a meaningful substitute for the works themselves.

3. TDM researchers and teachers are adversely affected in their ability to make noninfringing uses.

The third element in assessing the need for an exemption to section 1201 liability is whether researchers “are adversely affected in their ability to make such noninfringing uses.”¹⁵⁹ The adverse effect must be more than “*de minimis*.”¹⁶⁰ The third element “is analyzed in reference to Section 1201(a)(1)(C)’s five statutory factors:”¹⁶¹

¹⁵⁵ See, e.g., Berkeley Research Data Management, <https://researchdata.berkeley.edu/> (last visited Dec. 13, 2020).

¹⁵⁶ *HathiTrust*, 755 F.3d at 100–101.

¹⁵⁷ Letter from David Bamman, Appendix B, at 3.

¹⁵⁸ Letter from Data-Sitters Club, Appendix D, at 3.

¹⁵⁹ 2020 NPRM at 65294.

¹⁶⁰ U.S. Copyright Off., Reg. of Copyrights, *Section 1201 of Title 17*, 28 (June 2017) <https://www.copyright.gov/policy/1201/section-1201-full-report.pdf> (quotations marks omitted) [hereinafter 2017 Report].

¹⁶¹ 2020 NPRM at 65294.

- (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.¹⁶²

In applying and weighing these factors, the Register “evaluates whether there are adverse effects on noninfringing uses based on the totality of the evidence, including market alternatives to circumvention that enable noninfringing uses.”¹⁶³

The statutory factors demonstrate that an exemption is warranted here. Granting the exemption will increase the availability of copyrighted works because researchers will generate more scholarly work as a result, specifically that which presents the results of TDM. Section 1201 creates adverse effects on teaching TDM because the inability to apply the technique to contemporary works deters students from learning it. It also deters students interested in using TDM from concentrating on certain fields of study where the section 1201 prohibition makes TDM prohibitively difficult. Section 1201 creates adverse effects on research by eliminating in-copyright motion pictures and literary works as objects of study. Granting the exemption will not harm the market for or value of the motion pictures and literary works researchers seek to access. Moreover, another factor supports the exemption: it will help promote equity by allowing researchers and teachers to focus their work on a more diverse and inclusive set of works.

- a. Granting the exemption would increase the availability for use of copyrighted works.

Under the first factor, the Register considers “whether there will be greater availability of copyrighted works in general if an exemption is granted.”¹⁶⁴ In doing so, the Register considers “both the positive and adverse effects of the prohibition [on circumvention] on the availability of copyrighted materials.”¹⁶⁵ In this case, granting the exemption will increase the availability of copyrighted works by allowing the production of materials presenting the results of TDM research. Prior exemptions allowing circumvention of TPMs on the materials encompassed by

¹⁶² 17 U.S.C. § 1201(a)(C)(i)–(v).

¹⁶³ 2017 Report at 28–29 (quotation marks omitted).

¹⁶⁴ 2015 Recommendation at 310. For example, in the previous triennial rulemaking the Register recommended expanding the current security research exemption because it would increase the availability of copyrighted works such as articles, presentations, and computer programs aimed at rectifying security flaws. *Id.*

¹⁶⁵ 2017 Report at 122.

the proposed exemption have not created market harm, and the exemption is sufficiently narrow and specialized that granting it will not discourage people from making copyrighted works available for use.

The Register has previously found that the first factor favors granting an exemption when it “would increase the availability of copyrighted works in the form of articles, presentations,” and other scholarly works.¹⁶⁶ Granting the exemption will incentivize the creation of new copyrightable works based on TDM research on both motion pictures and literary works distributed electronically; the new works will include scholarly articles, presentations, and books discussing findings resulting from TDM. Item C describes some of the research projects that are planned to be undertaken if the exemption is granted.¹⁶⁷

At the same time, granting an exemption is unlikely to reduce the availability for use of copyrighted works. The Register has previously recognized that a limited exemption that allows circumvention of TPMs on DVDs, Blu-ray discs, and digitally transmitted motion pictures would not have a material impact on the availability of motion pictures.¹⁶⁸ The Register has long granted certain exemptions for literary works distributed electronically,¹⁶⁹ and has in the past concluded that the exemptions have not undermined the value of or market for e-books.¹⁷⁰ Although the Register has not previously addressed this question as to Blu-rays with Ultra HD content, there is no logical reason to believe the outcome would be different. Moreover, the market for digitally transmitted video, e-books, and DVDs appears to be robust.¹⁷¹ While the same cannot be said of Blu-ray discs,¹⁷² there is no reason to think that is because of circumvention of TPMs rather than consumer preference. It is unlikely that a narrow exemption

¹⁶⁶ 2018 Recommendation at 312.

¹⁶⁷ *Supra* Item C.1.f.

¹⁶⁸ 2015 Recommendation at 93–94.

¹⁶⁹ *See, e.g.*, U.S. Copyright Off., Recommendation of the Reg. of Copyrights in RM 2002-4 on Rulemaking on Exemptions from Prohibition on Circumvention of Copyright Protection Sys. for Access Control Techs. (Oct. 27, 2003) <https://cdn.loc.gov/copyright/1201/docs/register-recommendation.pdf>, at 66.

¹⁷⁰ 2015 Recommendation at 136.

¹⁷¹ Michael Balderston, *More Than 25% of Consumers Added at Least One Streaming Service During Pandemic*, TVTechnology (Aug. 11, 2020), <https://www.tvtechnology.com/news/more-than-25-of-consumers-added-at-least-one-streaming-service-during-pandemic> (“40% of consumers said they were watching more streaming TV because of the pandemic”); John Latchem, *DVD Format Getting Pandemic Boost*, Media Play News (May 20, 2020), <https://www.mediaplaynews.com/dvd-format-getting-pandemic-boost/>; Adam Rowe, *Ebook App Readership Is Up 30% Amid Pandemic Lockdowns*, Forbes (Mar. 28, 2020), <https://www.forbes.com/sites/adamrowe1/2020/03/28/ebook-app-readership-is-up-30-amid-pandemic-lockdowns/?sh=71d2212a721d>.

¹⁷² *See* Latchem, *supra* note 171.

for creating collections of works on which to conduct TDM will reduce the availability of motion pictures or literary works distributed electronically.

- b. Granting the exemption would increase the availability for use of works for nonprofit archival, preservation, and educational purposes.

Under the second factor, the Register considers whether an exemption is likely to increase the availability of works for nonprofit archival, preservation, and educational purposes.¹⁷³

The Register has previously found that granting an exemption to enable good-faith security research would likely increase the use of copyrighted works in educational settings because “the current prohibition plays a negative role in universities’ willingness to engage in and fund security research, and may limit student involvement in academic research projects.”¹⁷⁴

Similarly, here, the prohibition adversely impacts teaching. As described above, more students will be interested in learning about TDM if they can apply it to contemporary works, instructors will be able to design syllabi that draw on a more diverse and inclusive set of works, and students will not be deterred from pursuing particular fields because of the inability to apply TDM to those fields.¹⁷⁵

Thus, this statutory factor weighs in favor of granting the exemption.

- c. The prohibition on the circumvention of technological measures applied to copyrighted works adversely impacts criticism, comment, news reporting, teaching, scholarship, and research.

When a proposed exemption seeks to enable uses of copyrighted works that advance the core purposes of copyright—criticism, comment, news reporting, teaching, scholarship, and research—this factor “weighs strongly in favor of properly crafted exemptions to foster such uses.”¹⁷⁶ Where “research is at the core of the proposed exemption . . . adopting such an exemption would thus serve to promote research.”¹⁷⁷

An exemption to enable TDM research on motion pictures and literary works will promote research. As described in detail above, TDM is a valuable research technique across disciplines.¹⁷⁸ While it has been used to great effect in areas such as medical science and social

¹⁷³ 2018 Recommendation at 76.

¹⁷⁴ 2015 Recommendation at 310.

¹⁷⁵ *Supra* Item C.1.d.

¹⁷⁶ 2015 Recommendation at 94.

¹⁷⁷ *Id.* at 311.

¹⁷⁸ *Supra* Item C.1.a.

science, its application to in-copyright motion pictures and literary works distributed electronically has been substantially hindered by section 1201.¹⁷⁹ As a result, there has been little work applying TDM to motion pictures, and most TDM on literary works focuses on works in the public domain.¹⁸⁰ The consequence is that the technique cannot be applied to understand many important questions.¹⁸¹ Additionally, TDM can only be applied to works of authorship at a time when many who were not white and male were excluded from opportunities to create motion pictures and literary works.¹⁸² Finally, the fourteen academic researchers who have submitted letters in support of this exemption have presented a vast array of specific research projects they hope to pursue if the exemption is granted, and the letters from ACH and the scholarly publication experts attests to the high level of interest in pursuing TDM research on the proposed classes if the exemption is granted.¹⁸³ Thus, this factor weighs greatly in favor of granting an exemption.

- d. Circumvention of technological measures to enable TDM research is unlikely to adversely affect the market for or value of copyrighted motion pictures and electronically distributed literary works.

Under the fourth statutory factor, the Register considers “the effect of circumvention of technological measures on the market for or value of copyrighted works.”¹⁸⁴ The language of this factor is nearly identical to that of the fourth fair use factor,¹⁸⁵ and the Register has applied the same analysis to both factors.¹⁸⁶ Thus, this factor weighs in favor of the proposed exemption for the same reason as the fourth fair use factor.¹⁸⁷ The creation of a collection of works to draw conclusions about those works does not diminish the market for or value of the copyrighted works. No one seeking to watch a motion picture or read a literary work would instead read a TDM research paper analyzing such works. While some publishers do license collections of works for TDM, creation of alternative collections do not create markets because these

¹⁷⁹ *Supra* Item C.1.b.

¹⁸⁰ *Supra* Item C.1.c.

¹⁸¹ *Id.*

¹⁸² *Id.*

¹⁸³ *Supra* Item C.1.f.

¹⁸⁴ 17 U.S.C. § 1201(a)(1)(C)(iv).

¹⁸⁵ *Compare id.* (“the effect of circumvention of technological measures on the *market for or value of copyrighted works*”) (emphasis added), *with id.* § 107(4) (“the effect of the use upon the potential *market for or value of the copyrighted work*”) (emphasis added).

¹⁸⁶ 2018 Recommendation at 312 (“As to the fourth factor, the Acting Register finds that granting the expansion, while retaining the other limitations discussed, is unlikely to adversely affect the market for or value of copyrighted computer programs, for the same reasons stated above in reference to the fourth fair use factor.”).

¹⁸⁷ *Supra* Item E.2.b.

collections are plainly transformative. There are no legitimate security concerns with allowing researchers to create collections for TDM. Finally, granting the exemption may even increase demand for motion pictures and literary works as researchers seek to purchase them to build collections for TDM.

e. Granting the exemption would promote equity.

Finally, the statute allows consideration of “such other factors as the Librarian considers appropriate.”¹⁸⁸ This “catchall” provision is wide-ranging and not limited to copyright considerations.¹⁸⁹ An additional reason to grant the exemption is that it would promote equity by allowing TDM to be performed on a more diverse and inclusive set of works. As discussed previously, public domain works are less diverse in terms of authorship than in-copyright contemporary works.¹⁹⁰ As a result, research questions TDM can be used to answer once again center white men. In addition, those who teach TDM have a more difficult time persuading today’s increasingly diverse student bodies to take up the technique given the limits of the material they can analyze using it. Thus, granting the exemption would promote equity.

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

The fourth element in the section 1201 rulemaking analysis is whether “[t]he statutory prohibition on circumventing access controls is the cause of the adverse effect.”¹⁹¹ Proponents must show that “but for the [section 1201] prohibition, users likely could gain lawful access” to the copyrighted work for their particular purpose.¹⁹² Adverse impacts “that flow from other sources, or that are not clearly attributable to implementation of a technological protection measure, are outside the scope of the rulemaking.”¹⁹³

The TDM researchers have offered ample evidence that the statutory prohibition on bypassing access controls—and nothing else—is causing the adverse effects they are experiencing.¹⁹⁴ They have documented a range of adverse effects, not only demonstrating that TPMs have prevented them from studying in-copyright motion pictures and literary works, but also further detailing the distorting impact this has had on teaching and research in the field.¹⁹⁵

¹⁸⁸ 17 U.S.C. § 1201(a)(1)(C)(v).

¹⁸⁹ 2017 Report at 124.

¹⁹⁰ *Supra* Item C.1.c.

¹⁹¹ 2020 NPRM at 65294.

¹⁹² 2018 Recommendation at 67.

¹⁹³ 2018 Recommendation at 15–16.

¹⁹⁴ *Supra* Item C.1.b.

¹⁹⁵ *Supra* Item C.1.c. & d.

Beyond this, proponents have demonstrated that there are no alternatives to circumventing access controls available to them.¹⁹⁶ As discussed above, while some digital libraries do exist, they are too difficult to use, lack adequate analytical tools, and their collections are too incomplete.¹⁹⁷ Moreover, no such libraries exist for motion pictures.

Thus, researchers focused on literary works and motion pictures have no choice but to build their own collections. For literary works, this means relying on OCR. However, this option too is not feasible. OCR requires too much time and labor, and introduces too many errors, for it to be a workable solution.¹⁹⁸

For those who work with motion pictures, the available option is screen capture. The Register has previously observed that “screen capture technology produces lower-quality images than those that are available by circumvention.”¹⁹⁹ The Register has also concluded screen-capture is not an adequate alternative to circumvention where high-quality content is required.²⁰⁰ That is the case for TDM, which is harmed by the lower resolution footage that screen capture generates and the fact that it does not capture structural markers.²⁰¹ In addition, the process is far too slow to build the collections researchers need to apply TDM to motion pictures.²⁰²

The statutory prohibition on circumvention is the cause of the adverse effects on TDM research because existing alternatives to circumvention are unworkable. Therefore, the final element of the section 1201 rulemaking analysis is satisfied.

As the court explained in *Google Books*, “[t]he ultimate goal of copyright is to expand public knowledge and understanding.”²⁰³ An exemption would advance this goal by making it possible for researchers to apply TDM, a groundbreaking technique, to in-copyright motion pictures and literary works distributed electronically. It would also make it easier for teachers to pass this technique on to a new generation of students. The proposed exemption meets all of the criteria for a new exemption and should be granted.

¹⁹⁶ *Supra* Item C.1.f.

¹⁹⁷ *Supra* Item C.1.e.i.

¹⁹⁸ *Supra* Item C.1.e.ii.

¹⁹⁹ 2012 Recommendation at 133.

²⁰⁰ 2018 Recommendation at 36.

²⁰¹ *Supra* C.1.e.ii.

²⁰² *Id.*

²⁰³ 804 F.3d at 212.

Documentary Evidence

Commenters are encouraged to submit documentary evidence to support their arguments or illustrate pertinent points concerning the proposed exemption. Any such documentary evidence should be attached to this form and uploaded as one document through [regulations.gov](https://www.regulations.gov).

APPENDICES

APPENDIX A: LETTER FROM THE ASSOCIATION FOR COMPUTERS AND THE HUMANITIES

APPENDIX B: LETTER FROM DAVID BAMMAN

APPENDIX C: LETTER FROM JAMES CLAWSON

APPENDIX D: LETTER FROM THE DATA-SITTERS CLUB

APPENDIX E: LETTER FROM ERIC HOYT

APPENDIX F: LETTER FROM MATTHEW JOCKERS

APPENDIX G: LETTER FROM HOYT LONG

APPENDIX H: LETTER FROM JES LOPEZ

APPENDIX I: LETTER FROM ANDREW PIPER

APPENDIX J: LETTER FROM MATTHEW SAG

APPENDIX K: LETTER FROM RACHAEL SAMBERG AND TIMOTHY VOLLMER

APPENDIX L: LETTER FROM DAN SINYKIN

APPENDIX M: LETTER FROM LAUREN TILTON AND TAYLOR ARNOLD

APPENDIX N: LETTER FROM TED UNDERWOOD

APPENDIX O: LETTER FROM MELANIE WALSH

APPENDIX P: LETTER FROM HENRY ALEXANDER WERMER-COLAN

Appendix A
Letter from the Association for Computers and the Humanities



The Association for
Computers and the
Humanities

December 3, 2020

To the Librarian of Congress:

On behalf of the members of the Association for Computers and the Humanities (ACH), we support an exemption to DMCA § 1201 for non-consumptive text and data mining (TDM) of in-copyright materials. As we demonstrate below, the law adversely affects current research and the development of the field of digital humanities (DH), with negative impacts for students, scholars and the public at large.

The Association for Computers and the Humanities (ACH) is the professional organization for the digital humanities (DH) in the United States. ACH was founded in 1978, and co-founded the international Alliance of Digital Humanities Organizations (ADHO) in 2005. Since then, ADHO has grown to include ten member organizations with collective global reach. Our peer-reviewed open-access journal, *Digital Humanities Quarterly*, publishes four issues per year, is now on volume 14, and is among the most widely-cited publications in the field. ACH has over 350 members, and our most recent conference in 2019 was attended by 422 individuals from across the country and world. The annual ADHO conference, which we support, most recently had over a thousand attendees in 2019. We take an active role in advocating on behalf of our members, particularly with regard to legal issues that impede or restrict digital scholarship, and have previously submitted letters to support related cases, including *Authors Guild v. Google*. To inform this letter, we sent out a survey to our members, asking how access to in-copyright materials has affected their work in general, particularly the impact of DMCA § 1201. Some of our members have submitted their own letters of support for this petition, but ACH wants to ensure that a broader range of voices are represented in the discussion, including those who wished to comment anonymously.

As the ACH President Kathleen Fitzpatrick writes, DH is “a nexus of fields within which scholars use computing technologies to investigate the kinds of questions that are traditional to the humanities, or, as is more true of my own work, ask traditional kinds of humanities-oriented questions about computing technologies.” As an interdisciplinary field, DH brings together research areas across the humanities, social sciences, and sciences such as data science, literary studies, and computer science to engage with and forge new methods for text and data mining (TDM). As a result, access to data including still and/or moving images, sound, and text is critical to the work of scholars in this field. Because of DMCA § 1201, research in our field is curtailed. We will now turn to three main areas: text analysis, image analysis, and sound analysis.

Computational methods for text analysis have been available to scholars for over three decades, which has led text analysis to be one of the longstanding methodological pillars of the field. In our survey of ACH members, 100% of the respondents agreed with the statement that circumventing technical protection measures (TPM) on legally purchased ebooks would be useful for their research. Furthermore, 100% of respondents agreed with the statement that it was financially unfeasible for them to pay someone to scan and perform optical character recognition



(OCR) in order to transform all of the books that they would like to use into digital copies. One scholar noted that, even in cases where scholars do have access to the funding necessary for scanning and OCRing books, it has an impact on the timeline required for projects that involve in-copyright texts: “I am running a three-year project that is only possible because I started digitizing the corpus more than a year before the project start date.” For some research questions in our field, scanning and OCR is not enough to get the necessary data, which may be embedded in the digitized file (such as provenance or information about the devices used to create the work). A scholar commented that they “wanted to study the provenance of files where it appeared that multiple versions of an ebook were for sale from the same bookseller. This is difficult to do/impossible without insight into the code of the file.” As a result of DMCA § 1201, scholarship in fields as varied as literary studies, media studies, and STS (science, technology, and society) are inhibited.

DMCA § 1201 has had a profound impact on the career trajectories of scholars in our field. In response to our survey, multiple scholars noted that they have made deliberate choices about what topics to pursue in response to copyright law. As one scholar stated, “I oriented my entire career around literature in the public domain in order to avoid having to deal with copyright.” One scholar noted that in-copyright materials are typically more resonant for broader publics beyond the academy, and given an exemption for circumventing TPM for research, “I would be able to refocus my research on work that is especially relevant for the public, and not just Victorian novels!” Another response to the survey described how DMCA § 1201 warps the process of identifying a research question. “It changes the projects I work on--we end up starting from a place of ‘what can we do?’ instead of ‘what would be best for this research?’” The respondent added that “it's also dramatically slowed my progress to dissertation--it has taken me so long to compile things from a variety of sources--and it has increased the cost of my dissertation in software, purchases, and time.” The impact of DMCA § 1201 is adversely shaping the lives, projects, and career trajectories of scholars, and unless an exemption is permitted, it will continue to do so for years to come.

When asked about what text analysis work they would undertake in the next three years if an exemption to DMCA § 1201 were granted, our members envisioned a wide range of possibilities. “I end up doing the DH for my students in certain classroom settings because I don't want to risk getting them in trouble. This changes how and what I can teach and has a gatekeeping effect--I'm the one with the methods and the texts, and even if I take steps to make it more transparent (such as running computational text analysis code in front of them), at the end of the day, they didn't do the work and will have harder time replicating it if they want to,” explained one scholar. An exemption would put these methods directly in the hands of students at a moment when computational analysis and machine learning are a global research priority. Another scholar also imagined a positive impact on students' ability to do research that matters to them: “As just one small example, my undergraduate course asks students to do an experiment with type-token ratios around a research question of their own choosing; 90% of students pose absolutely fascinating research questions about contemporary literature that they cannot pursue due to ebook encryption, and glumly accept our public-domain substitutions. These students would have an unambiguously more effective learning experience if able to pursue questions that matter to them with texts they already care about.”

DMCA § 1201 substantially restricts the data available for TDM not just by time, but other features such as gender, race, and class. Another response noted that “that the bias toward pre-1925 texts prevents my digital humanities classes from including more women authors, non-binary authors, and authors of color, as digitized and available pre-1925 texts are mostly written by white men”. In this way, an exemption to DMCA § 1201 would expand the representation in the textual corpora that scholars can meaningfully access.

There is also work that some scholars are currently doing that cannot be published due to concerns about text provenance, when the scholar obtained those texts in some way other than scanning and doing OCR. One respondent remarked that “The ability to use texts for text analysis that are currently under copyright would completely open up my ability to publish the scholarship I've been working on for over 10 years.” When scholars do OCR texts, they may choose to not proofread for errors (which can be numerous, based on factors including the quality of the scan, and the complexity of the fonts used on the page) in order to try to keep digitization costs down. A scholar noted that reliably having texts without those errors would allow them to do analyses with greater precision than currently possible. Given the increase in born-digital books that are inherently free of the errors introduced by OCR, an exemption to DMCA § 1201 would unlock research questions that are not feasible to pursue under the current system.

Multiple scholars had specific projects in mind that they would undertake during the exemption’s three-year timeframe, including a “project exploring the evolution of novel genres in response to sales over the long 20th century” that previously had to be discarded due to insufficient data. Another hopes to build on work that has already been done on texts in the public domain: “Speaking just to the English language context, there is a tremendous amount of work to be done on 20th century literature. A lot of innovative analysis and model building has been done for 19th century fiction, and it would be very generative to bring these up to date with 20th century examples.” Other projects include the application of distant reading methods to global literature, genre, and translation across time and geography.

The restrictions imposed by DMCA § 1201 have an equal or greater negative effect on scholarship that uses image and sound data. As Thomas Smits and Melvin Weavers argue, DH is undergoing a visual turn while scholars such as Mary Caton Lingold have highlighted how DH is also undergoing a sound turn. The growth of these areas within the field has been demonstrated in areas such as the field’s major publications and international conferences and facilitated by recent technological advancements, particularly the expanded capabilities of machine learning and neural networks. Scholars engaging with this work are asking question such as:

- Which visual cultures have film, TV, and photography produced across the 20th and 21st century? (Visual Culture and Media Studies)
- How can we use computer vision to analyze film and TV style at scale? (Film and TV Studies)
- How have genres of music developed sonically over the 20th century? (Sound Studies)
- How has the form of podcasts changed in the past decade?
- What are the major topics covered in radio across the world?



The Association for
Computers and the
Humanities

- Are there gender and dialect patterns for audiobook narrators (by publisher, genre, year, etc)?
- Which new algorithms and methods can we develop for the large scale analysis of human culture?

As a result, scholars are also making methodological interventions. Like distant reading, there are calls for distant viewing and distant listening. Yet, the analytical possibilities of these methods are massively limited by DMCA § 1201. The law eliminates much of the available materials in the 20th and 21st century. Much of film, music, and TV - among the most powerful cultural forms of the late 20th and early 21st century - is illegal to TDM because of their file formats. Large scale TDM requires access to sound files such as AAX, moving image files such as DVDs, and born digital streaming files such as Silverlight. Scholars report having been denied funding because of concerns about copyright as well as the cost of access to copyrighted materials. The result is that scholars either do not engage with these areas of research or shape their project around data that isn't subject to DRM.

The damage is further elucidated when we compare our context to scholars in areas such as Europe. Large national and EU infrastructure such as DARIAH (EU) and MediaSuite (Netherlands) has resulted in large scale commitments to making accessible audio and visual data for TDM. With access to the data, scholars have received tens of millions of euros of funding for research projects. They are now pioneering new approaches and methods, leaving US scholars at a disadvantage. The ability to use materials with DRM for TDM would open up entire areas of scholarship in DH such as large scale analysis of moving images and sound animated by questions from fields such as Communications, Film Studies, and Media Studies. It would also facilitate the development of methods such as distant listening and distant viewing.

Vast amounts of cultural materials -- in text, image, audio, and moving image media -- are effectively cut off from computational analysis as the result of DMCA § 1201. Whether the issue lies in the unfeasibly high cost of alternate forms of digitization (i.e. scanning and OCR) or file formats that are inaccessible through TPM, the effect is the same: DMCA § 1201 is constraining the research questions that our members can pursue, and shaping their careers to steer them away from topics of significant public interest and relevance. Even in the span of a three-year exemption, our members have ideas and projects that they would like to undertake to intervene in this situation, with positive effects on scholarly disciplines, students, and the public at large.

Sincerely,

A handwritten signature in black ink, appearing to read 'Kathleen Fitzpatrick', written in a cursive style.

Kathleen Fitzpatrick
President, Association for Computers and the Humanities
On behalf of the ACH officers and executive council

Appendix B
Letter from David Bamman



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

November 18, 2020

I am writing this letter in support of the Authors Alliance’s petition to the Copyright Office for an exemption to §1201. I am an assistant 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.

At the core of all work in text data mining is access to data; the ability to access data shapes the research questions we are able to ask, the methods that we select to answer them, and the ways in which our findings are disseminated to the broader public. For a long time, work in cultural analytics was focused on texts in the public domain, such as those accessible through open resources like Project Gutenberg; public domain texts provide a proving ground for analytical methods in text data mining and facilitate the important scientific goal of reproducibility: by providing a stable source of data that *everyone* can access, it enables researchers to verify claims made by others, thereby strengthening trust in the scientific process and encouraging innovation. Part of my research group’s work over the past few years has focused on improving the state of the art in NLP for literary texts; in order to create a benchmark that others can use to evaluate their own systems, we purposely selected 100 public domain works from Project Gutenberg (Bamman et al. (2019), “An Annotated Dataset of Literary Entities,” <https://github.com/dbamman/litbank>).

At the same time, however, public domain resources are necessarily limited. At the time of writing, the public domain in the United States largely spans materials created before 1925. While this body of material includes important works in the 19th-century literary canon (such as Jane Austen’s *Pride and Prejudice* and Mark Twain’s *Tom Sawyer*), it still represents works nearly a century removed from the present day, limiting its ability to answer research questions that are relevant to a contemporary audience. These include not only 21st-century questions on the influence of the internet and social media on literary forms and reading behavior—but even much older questions including the rise of the Harlem Renaissance in the 1920s and 30s. Indeed, public domain works published on Project Gutenberg systematically overrepresent white, male

¹See, for example: 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; and Lara McCounaughy, Jennifer Dai and David Bamman (2017), “The Labeled Segmentation of Printed Books,” EMNLP 2017.

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*.

authors, and so the research questions it is able to answer again privilege that social group over others.

There are two primary ways that researchers carry out work on in-copyright texts. The first is through the use of large-scale digital libraries like the HathiTrust, which enable non-consumptive research access to a vast collection of in-copyright materials (17.4 million total works at the time of writing). The HathiTrust Digital Library is a trailblazer in facilitating transformative research by hundreds of researchers by providing access to in-copyright materials, enabling researchers to answer questions at the scale of thousands of texts that simply could not be answered otherwise, but it is not a solution for all research questions. In order to carry out research in a secure environment, all computing is carried out on servers at the HathiTrust through the use of a secure “data capsule” which allows researchers to computationally process texts without being able to directly transfer any material outside of the secure environment. This limits computational processing to the capacity of the HathiTrust’s resources, which is occasionally outside the demands of contemporary state-of-the-art models in NLP—which, for example, may require the use of graphics processing units (GPUs) common in NLP research labs, but not in large-scale conventional compute clusters.

While this mismatch between computing demands and available resources can of course be alleviated as GPUs make their way into compute clusters, one issue that also arises in the use of digital collections compiled by a third party is the presence of gaps in the collection needed to answer a specific research question. The materials in the HathiTrust originate in university libraries, and so are necessarily biased toward academic monographs and away from, for example, mass-market romances and science fiction more commonly found in city public libraries. This gap is a common impetus for the second way that researchers carry out work on in-copyright texts: by digitizing a collection themselves. In my group’s own work on creating annotated resources to improve the state of the art for NLP, we did just that: we bought 500 in-copyright books, scanned them, and carried out OCR on those page scans to recognize the text printed on them. OCR is an errorful process; on a sample of books we scanned, we measured the word error rate to be 0.305% (i.e., roughly one incorrectly recognized word every page); and this process of scanning each one of 500 books is also very labor intensive, consuming the better part of four months. A much faster and more accurate way that we could have selected would have been to buy digital versions of those texts as eBooks; but our concern over violating §1201 dissuaded us from that route, committing our efforts to the slower, more error-prone process and consuming research time that could have been more productively applied elsewhere.

While text certainly has the longest history as the subject of research in data mining and cultural analytics, the rise of computer vision and video processing techniques have also enabled film to arise as a meaningful object of computational inquiry. However, while the existence of public-domain datasets of texts (such as Project Gutenberg) and in-copyright secure environments (like the HathiTrust data capsule) allow researchers to explore text data mining methods without risk of implicating §1201, no such pre-existing resource exists for movies or television. Researchers need to create such datasets themselves.

In early 2018, I decided to create such a dataset in order to explore several questions around film: can we measure directorial *style* in movies? What is it that allows us to immediately recognize that a movie is directed by Wes Anderson and another by David Lynch? While computational methods have shed light on the field of authorship attribution—predicting the author of text either in order to deanonymize them (such as the authorship of the *Federalist Papers*) or to simply characterize what makes them distinct—no such work exists for using computational methods to characterize the visual properties of a movie that uniquely make it recognizable as the style of a particular director. We might hypothesize that “style” in this case can be decomposed into a number of aspects that *could* be measured—including pacing variables such as average shot length, proportions of shot types (close-up vs. long shot), and the color palette used over the course of

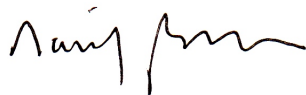
a movie. I decided to begin assembling a dataset to examine these questions, ideally creating a dataset of ca. 10,000 movies; a dataset this large would not only be sufficient to answer questions of directorial style, but would revolutionize the computational study of film by putting it on the same scale as work in text analysis.

The barrier, however, was in creating such a dataset. One fast and accurate way for a researcher to create such a dataset would be to buy DVDs and use software to rip the film from that medium. Given my own low tolerance for §1201 risk, however, I decided to digitize movies in a way that accords with fair use: buying DVDs of the movies, playing them on a computer, and using screen-capture tools to record the screen while the movie is playing in real time (note the time required to digitize a movie using this method is exactly the original runtime of the movie itself). While this digitization method allows for movie data to be used in computational analysis, it is an imperfect process that necessarily loses important information—the subsequent data is of lower resolution than the original medium, and important structural markers like chapter boundaries are lost. However, after digitizing roughly 200 movies in this way, it became clear that this was an infeasible path forward. If a human operator were present for the duration of the screen capture for each movie with an average runtime of 2 hours (and worked 8 hours a day, 5 days a week, 50 weeks per year) it would take 10 years to complete the act of digitization alone.

Two years later, I have still not taken up this original line of research, which I expect will be transformative once it is able to be carried out. If an exemption to §1201 were granted, I would certainly pick up this line of research and begin examining other ways in which movies can function as an object of study in their own right—questions worth examining include historical trends in film over the past century (has the depiction of violence within movies become more or less predominant?), imitation and influence (can we trace which directors have had the greatest impact on the visual style of directors after them?) and reception (which specific aspects of film do audiences, critics, and the box office most respond to?).

All of these questions have been examined within text data mining given the existence of large digitized text collections, but so far remain outside the limits of our knowledge for film.

Sincerely,

A handwritten signature in black ink, appearing to read "David Bamman". The signature is fluid and cursive, with a prominent initial "D" and a long, sweeping tail.

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

Appendix C
Letter from James Clawson



October 30, 2020

To the Register of Copyrights,

I am the Ann Petry Endowed Professor of English at Grambling State University in Louisiana, where I teach courses in literature and in data analytics. Part of my role at Grambling also includes conducting research, for which I combine literary approaches with methods of text data mining to find new understandings of literature written since 1900. Because the digital copyright protection systems covered by § 1201 limit both my research and my teaching, I support the petition for an exception for text and data mining. I am writing in my personal capacity.

The burdens from these protections have restricted my area of scholarship. The research I did as part of my Ph.D. focused primarily on novels written after World War II—works that are still in copyright. Since then, I have learned techniques of data analysis and text and data mining that I apply to my research; because these techniques require machine-readable texts, I've had to shift my scrutiny to works that are available digitally online—commonly literature written before the 1920s. This limitation is frustrating because it means I seldom combine my doctoral expertise with newer analytical methods. It is doubly frustrating when, even as I own books in ebook editions, protection by digital rights management (DRM) software keeps me from using these clean digital editions of books in non-consumptive text data mining research.

Without direct access to texts protected by DRM, literary researchers like me have limited options. To use methods of large-scale text data mining, I turn to online libraries like the Internet Archive and Project Gutenberg. Unfortunately, these online libraries are also limited. While they provide a useful starting point to contextualize many analyses of out-of-copyright works, they do not offer an exhaustive selection of out-of-copyright material, and they understandably do not provide any material that is still in copyright. Without the ability to bypass DRM, a researcher interested in studying other material must commit to an arduous process: first scanning a book, then using optical character recognition software to turn images into words, cleaning up the output to remove bad transcriptions, correcting line-break hyphens, stripping page numbers, and excising running titles. The final result is a series of simple text files, not unlike those already embedded within ebooks. This work typically takes dozens of hours to prepare a single novel, and some methods of text data mining require hundreds of novels, making it impossible to pursue some questions. Even for out-of-copyright works, the DRM on ebooks presents an insurmountable barrier, the removal of which would theoretically cut dozens

of hours down to almost zero; allowing the removal of DRM protections would thereby also allow for more research to be done on literatures poorly represented in online libraries; and it would open the door for more text data mining research to be done on contemporary writing.

In addition to all the above, the limitations of DRM software make existing disparities more pronounced. At larger institutions with bigger budgets and libraries, students and researchers often have access to databases like the HathiTrust, a compendium of digitized texts that far surpasses anything available publicly. Students and faculty at member institutions gain access to millions of volumes for text and data mining research, including material that is still in copyright. But underfunded public institutions and historically black colleges like Grambling are unlikely to participate in this kind of partnership, which requires a university to share library holdings and to pay a high membership fee; at schools like ours, investments to significantly expand holdings (in order to meet a contribution threshold) or to pay new membership fees necessarily fall behind priorities like keeping tuition costs low. HathiTrust's list of member institutions shows a pattern of well-resourced universities further enjoying the benefit of access, and only a small number of historically black colleges are fortunate enough to be included. Because we lack options for sourcing material for text and data mining research, the pinch is felt first by faculty serving historically disadvantaged populations; it is felt most by our students.

If an exemption is granted, I would pursue projects that are otherwise out of reach. In the classroom, I would direct part of Grambling's upcoming course in text analytics around a single, large corpus of African American writers, many of whose works are often underrepresented in available archives. And in my ongoing research on pseudonymity and style in twentieth-century writers, I would widen the scope of study for a comprehensive understanding of the topic, considering authors and works who would otherwise not make the cut. Except on a small scale and with a big investment of time, the limitations of DRM restrict these kinds of text and data mining projects to institutions able to pay for access to private databases.

I ask that you grant an exemption to Section 1201 for text and data mining. Allowing researchers to remove DRM would enable the criticism and research of contemporary texts, and it would lower the barrier to studying overlooked texts from under-represented literatures. Additionally, allowing for the removal of DRM protections for text and data mining purposes would make it possible for faculty and students at under-resourced universities to apply these techniques to studying a wider variety of literature, narrowing the gap of access between students at elite schools and those everywhere else.

Sincerely,



James Clawson, Ph.D.
Grambling State University

Appendix D
Letter from the Data-Sitters Club



Stanford University

October 16, 2020

Dear Librarian of Congress,

We are writing on behalf of the Data-Sitters Club, a research group under the Stanford Literary Lab, in support of an exemption to the anti-circumvention provisions of the Copyright Act to allow researchers like us, and the students we teach, to access ebooks for fair use research purposes relating to computational text analysis.

Computational text analysis methods allow literary scholars to ask and answer questions that previously would have taken decades of painstaking research, if they were possible at all. This analysis has value whether it is about the works of William Shakespeare or more recent authors outside the traditional literary canon. The critical analysis of modern, popular texts is a vital part of humanities research; it helps us to understand how books both mirror and shape people's understanding of the world and the major issues of our time.

In the 1980's and 1990's, Ann M. Martin and a team of ghostwriters wrote a total of over 200 children's books, known collectively as the *Baby-Sitters Club* series. It is an iconic depiction of girlhood in the upper-middle-class American suburbs of the time, and was tremendously popular with elementary- and middle-school-age girls at the time. Its distinctive characters personally resonated with many girls; the 2020 documentary *The Claudia Kishi Club* focuses on the impact of a character who was one of the few broadly popular Asian-American role models during those decades. There's been relatively little scholarship written on the series, and what has been published focuses on the close reading of specific, individual texts. Applying the tools and methods of text and data mining to a corpus like the *Baby-Sitters Club* can make it possible to address a different set of questions. It allows researchers to draw upon all the books at once in order to gain an understanding of the totality of this series and how it builds its fictional world.

The Data-Sitters Club has begun to explore a broad agenda of research questions in relation to the *Baby-Sitters Club* series. Each novel is written in the voice of one (or multiple) characters, by Ann M. Martin herself or one of numerous acknowledged ghostwriters. Using computational methods, we are interested in whether each character has a distinct voice, and whether that voice is different across writers. We are interested in whether non-narrating characters themselves have distinct voices expressed through their dialogue, or if they just form classes of character types like "generic mother" or "generic classmate". We would like to find out how the characters' "written" language (shown through the portions of the text in the characters' "handwriting") differs from their implicitly spoken text through the first-person narration. The *Baby-Sitters Club* is (in)famous for its use of tropes, such as Claudia Kishi's "almond-shaped eyes", or

“Mal is white and Jessi is black”. We are interested in what else can we find out about how and where explicit text reuse happened in the most formulaic parts of the book, where the premise and characters are described in order to orient new readers. We are interested in how these books treat religion, race, adoption, divorce, and disability. The instructive role of children’s literature and the popularity of this series make it a particularly valuable one to study as a step towards understanding the worldview of American women currently in their 30’s and 40’s.

Finally, we are interested in adaptations into new media formats: what material was included (and what was removed or significantly transformed) in the creation of a recent graphic novel series, and a Netflix series, based on the original books.

The Data-Sitters Club also has pedagogical aims: we write up our process -- the decision-making and interpersonal aspects of our work, along with the technical steps -- and publish them as “books” on our website. Our goal is for anyone to be able to apply the same methods to texts and questions that interest them, and these “books” have already been incorporated into course syllabi by professors at Emory and Northeastern Universities. There remains one significant barrier for other people to do this same kind of work: access to texts.

Computational text analysis is not possible without text files, whether they come from ebooks, or are digitized from scans of printed books. While a vast amount of literature (including the entire *Baby-Sitters Club* corpus) is available for purchase as ebooks, which could be trivially easily converted to the plain text format used in computational research, most ebooks are protected by a technological protection measure (TPM). Although TPMs were intended to prevent piracy, for us they are often a roadblock to lawful and socially valuable research. To obtain the text in the necessary format without risking liability under the anti-circumvention provisions, scholars must go to great lengths. Typically, this involves scanning a book, and processing those scanned images using Optical Character Recognition (OCR) software, which generates usable text corresponding to the words that appear in the image. OCR is imperfect, and frequently makes mistakes, particularly if words are distorted near the edge of the page. Scanning a 130-page book (like one of the books in the *Baby-Sitters Club* series) can take 15-20 minutes, OCR can take another 10, and double-checking and correcting the OCR can take anywhere from 10-40 minutes, depending on the number of errors. The OCR error rate is particularly problematic in the sections of the *Baby-Sitters Club* books written in handwriting-style fonts, which OCR very poorly and need to be transcribed manually. These numbers increase when working with longer books, or books with complex formatting like tables. While scholars affiliated with a well-resourced institution such as Stanford may be able to bear the costs associated with paying someone to do this work, the costs are prohibitive for scholars at the vast majority of institutions in the US, including smaller public institutions and community colleges.

While computational methods can allow scholars to ask questions about thousands or even millions of books, the feasibility of doing that work plummets when that requires thousands or millions of hours of scanning and OCRing, even for a version of the text that contains errors. Converting an ebook, in contrast,

takes less than five minutes, and does not introduce any errors in the resulting text file. We purchased books we scanned for the project for a couple dollars, as used copies or library cast-offs. Even books that are generally in poor physical shape are fine for scanning and OCR. But if we were able to circumvent TPM without risking legal liability in order to build a corpus using ebook files, we would be happy to purchase ebook versions from the publisher. Circumventing TPM rather than scanning and OCRing books would enable scholars to spend more time pursuing research questions, allowing them to pursue projects with a more ambitious scope. Were it not for the legal uncertainty created by Section 1201, we could imagine in the next three years expanding the scope of our project to contextualize the Baby-Sitters Club within series books for girls, or even children's literature more broadly. Furthermore, it would become feasible for all of us — regardless of institution — to incorporate computational analysis of modern texts into the curriculum, enhancing students' awareness of the possibility and limitations of digital methods, using material that is more familiar and resonant than the public domain.

We urge you to consider adopting the proposed exception to the anti-circumvention law both to make computationally-supported research feasible without the extreme costs of needless digitization when digitized copies already exist as ebooks, and to support copyright holders in securing the ebook purchases of scholars with an interest in legally building research corpora.

Sincerely,

Lee Skallerup Bessette, Georgetown University
Katherine Bowers, University of British Columbia
Maria Sachiko Cecire, Bard College
Quinn Dombrowski, Stanford University
Anouk Lang, The University of Edinburgh
Roopika Risam, Salem State University

Appendix E
Letter from Eric Hoyt



WISCONSIN

UNIVERSITY OF WISCONSIN-MADISON

December 5, 2020

Dear Librarian of Congress:

I am respectfully writing to you in my individual capacity to support an exemption to DMCA Section 1201's anti-circumvention provisions because of their negative effect upon my research and teaching, along the wider fields of Cinema Studies, Media Studies, and the Digital Humanities.

By way of background, I am the Kahl Family Professor of Media Production in the Department of Communication Arts at the University of Wisconsin-Madison. I am also the Director of the Wisconsin Center for Film and Theater Research. My research and teaching takes place at the intersection of media history and digital technology. I am the author of the book, *Hollywood Vault: Film Libraries before Home Video* (2014), and co-editor of the anthologies *Hollywood and the Law* (2015) and the *Arclight Guidebook to Media History and the Digital Humanities* (2016). In working on these books, along with digital projects involving movies, magazines, podcasts, and screenplays, I have confronted research obstacles to data mining and the production of new knowledge as a result of DRM. In this letter, I specifically want to call attention to the problems caused by DRM restrictions to the large-scale analytics of moving image collections.

Over the past decade, one of the central methods I have tried to pursue in my research and teaching can be described as cognitive to computer comparison. Simply put, I look at a group of movies one way, a computer looks at them differently, what new knowledge can emerge from comparing those perspectives? I was able to generate some very partial answers to these questions through early work involving digitized out-of-copyright trade papers. However, I have felt continually limited and frustrated in my abilities to apply even a very basic form of this sort of this analysis to the media forms that interest me the most: movies and television. One of the major reasons for this is that it has been so difficult to build a meaningful corpus of high quality movies and television works for analysis. An exemption to DMCA Section 1201's anti-circumvention provisions would help enormously in this regard.

To provide a more specific example, I am one of many scholars who are keenly interested in the historical development of filmic and televisual styles. I have long wanted to run queries over a large and meaningful sample of films and TV programs to look for patterns in camera framing, shot length, color, brightness, and contrast. Such computational visual analysis is only

Department of Communication Arts

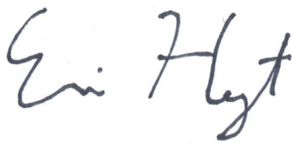
possible, though, if the digitized images are consistently pulled and assessed at a high quality. An exemption to DMCA Section 1201's anti-circumvention provisions would enable the building of a collection of films and television shows for research use from Blu-ray discs and other high-resolution sources. How have uses of film color, editing patterns, and the duration of close-ups changed over time? I have my own hypotheses based on viewing, taking notes, and reading the work of other film historians. But I would love to see what the computational algorithms notice (and what they don't).

I also know that this exemption would benefit my students. This semester, I am teaching a course called Digital Media Production for Graduate Students. This hands-on course explores ways that digital technologies can be leveraged toward research and teaching in Cinema and Media Studies. I once had a past student in this class express interest in data mining film collections and who asked me to recommend resources. This put me in a difficult position. As a professor, I never want to advise a student to take an action that might involve breaking the law. Ultimately, the student wound up picking a less ambitious and less consequential project to pursue. The requested exemption would help avoid this situation in the future.

I will conclude by saying: This is the right moment for such an exemption. For years, the data storage and computational processing requirements for the computational analysis of audiovisual media collections created technological barriers to this line of research inquiry. Compared to literature and text, for example, audiovisual media is far more resource-intensive to try to data mine. Fortunately, the technology has finally caught up. Universities can now better accommodate the storage and computational needs. If this exemption is granted, the research can finally catch up, too.

Thank you for your consideration.

Sincerely,

A handwritten signature in black ink that reads "Eric Hoyt". The signature is written in a cursive, slightly slanted style.

Eric Hoyt
Kahl Family Professor of Media Production
Department of Communication Arts
University of Wisconsin-Madison
ehoyt@wisc.edu

Director, Wisconsin Center for Film and Theater Research
<http://wcftr.commarts.wisc.edu/>

Appendix F
Letter from Matthew Jockers

October 5, 2020

To the Register of Copyrights:

I am dean of the College of Arts and Sciences and professor of English and Data Analytics at Washington State University. In addition to my academic career, I have founded and directed a non-profit organization, directed research and development at a technology startup company, worked as principal research scientist and software development engineer in iBooks Engineering at Apple, and cofounded a text mining company. I am writing in my individual capacity to support the creation of an exemption to enable non-consumptive text and data mining ("**TDM**") of in-copyright materials. Such an exemption will significantly aid my work in this field in the next three years and beyond if it is granted.

My academic research focuses on text and data mining. As an example, one of my projects analyzed stylistic changes in novels over the course of 200 years, examining thousands of different works. For another project, I created an algorithm designed to identify linguistic hallmarks of bestselling fiction. I have authored and co-authored numerous papers on text mining as well as three books: *Macroanalysis: Digital Methods and Literary History* (2013); *Text Analysis with R for Students of Literature* (2014); and *The Bestseller Code: Anatomy of the Blockbuster Novel* (2016).

Large scale TDM approaches allow us to answer questions that were previously unthought of or impossible to even ask due to their size and scope. "Strictly speaking," wrote Russian formalist Juri Tynjanov in 1927, "one cannot study literary phenomena outside of their interrelationships" Unfortunately for Tynjanov, the multitude of interrelationships far exceeded his ability to study them, especially with close and careful reading as his primary tools. With computational methods, we can now go beyond what Tynjanov could have ever even imagined. TDM approaches offer to provide specific insights into literary historical questions, including insights into:

- the historical place of individual texts, authors, and genres in relation to a larger literary context
- literary production in terms of growth and decline over time or within regions or within demographic groups
- literary patterns and lexicons employed over time, across periods, within regions, or within demographic groups
- the cultural and societal forces that impact literary style and the evolution of style
- the cultural, historical, and societal linkages that bind or do not bind individual authors, texts, and genres into an aggregate literary culture

- the waxing and waning of literary themes
- the tastes and preferences of the literary establishment and whether those preferences correspond to general tastes and preferences

Moreover, TDM provides a practical and tractable way of approaching questions such as:

- whether there are stylistic patterns inherent to particular genres
- whether style is nationally determined
- whether and how trends in one nation's literature affect those of another
- the extent to which subgenres reflect the larger genres of which they are a subset
- whether literary trends correlate with historical events
- whether the literature that a nation or region produces is a function of demographics, time, population, degrees of relative freedom, degrees of relative education, and so on
- whether literature is evolutionary
- whether successful works of literature inspire schools or traditions
- whether there are differences between canonical authors and those who have been traditionally marginalized
- whether factors such as gender, ethnicity, and nationality directly influence style and content in literature

My ability to use a wide variety of literary work is important because the research projects I pursue require substantial amounts of data in the form of text to draw conclusions about the central tendencies of works.

However, DMCA § 1201 is a major barrier to TDM: the statute makes it much more difficult for researchers to aggregate creative works into a dataset. But it is crucial that researchers are allowed to build their own corpora (collections of text) without fear of § 1201. The current alternatives limit the scope of potential research projects in three important ways.

First, existing corpora are often not adequate for a given TDM research project. Pre-assembled corpora may be unrepresentative and therefore would result in biased or distorted analysis if used. For instance, even the Hathitrust Digital Library corpus, which has proven to be a useful tool for TDM, does not have broad coverage of more modern, popular literature. Any project that attempted to use this corpus to draw conclusions about late 20th or 21st century literature risks missing vital inputs. This is a problem for researchers because it precludes any project that would focus on that kind of literature.

Additionally, the quality of optical character recognition ("**OCR**") across existing databases is often variable, tending to be lower-quality for older books. These inaccuracies may confound subsequent TDM analysis, especially when the errors are inconsistent from book to book. Error-riddled input data can undermine otherwise carefully conducted research leading to errors in analysis and conclusions.

Finally, usage of tags, the metadata used to differentiate sections within a book, is also inconsistent across existing databases. When examining texts in a database, researchers are unable to correct these tags, which can propagate errors in subsequent analysis. To illustrate, one of my projects tracks the shape of a novel's narrative structure, based on where particular words appear within the novel. Many serialized books contain a bonus chapter which previews the next book in the series. If the bonus chapter is not properly tagged as being outside the scope of the novel's narrative, my TDM software treats it as though it is the novel's final chapter. Having access to the text file, rather than working with a pre-built database, would allow me to tag the work however I need and thus avoid the problem of mislabeled or unlabeled sections introducing error into the analysis.

Enabling researchers to build their own corpora would avoid the barriers of under-representative corpora, error-riddled OCR, and missing metadata described above.

Without the restrictions imposed by § 1201, the easiest way to build a corpus would be to convert eBooks into text files. While researchers can scan print books and then use OCR, it is only a viable tool for creating smaller, more limited corpora due to the time and effort required per book. Scanning takes time and the fidelity of the resulting text file is seldom perfect, requiring a researcher to hand correct the errors. On the other hand, grabbing a text file from an eBook is comparatively quick and requires no error correction—enabling a researcher to build large, comprehensive corpora that would be impossible using OCR or by relying on existing databases. The only barrier to doing so is that circumventing the eBook's technological protection measure opens the researcher to legal liability because of § 1201.

Because of § 1201, I, and many other TDM researchers, have directed our research efforts towards literature in the public domain or to poor quality works that have been scanned and made partially available (i.e. snippets, word counts, etc.) through, for example, Google Books and the HathiTrust. This is not because public domain works or snippets are inherently more valuable subjects of scholarly examination than complete in-copyright literature, but because corpora containing exclusively public domain works or partial works carry fewer legal concerns. In other words, researchers like me are forced to choose what literature to analyze, not based on what offers the most interesting or important research questions, but instead based on what literature we can lawfully access. But this means that in-copyright work is vastly under-studied. As the law exists currently, the threat of § 1201 chills research on in-copyright works. As a field, we're missing insight into recent, culturally relevant literature because of this chilling effect.

If the exception Authors Alliance seeks is granted, I would likely want to immediately pursue many of the items that I provided in the bulleted list above. I have done a great deal of work on these questions using pre-1923 public domain texts; the opportunity to continue these analyses into the 20th and 21st century is incredibly exciting, especially so since it promises to tell us something about who we are today and not just in the now distant past.

TDM is an increasingly important method for researching and understanding literary works. An exemption for TDM would facilitate valuable research into modern in-copyright works—research which § 1201 prevents me and many others from currently performing.

Sincerely,

A handwritten signature in black ink, appearing to read 'M. Jockers', with a stylized flourish at the end.

Matthew L. Jockers
Dean, College of Arts and Sciences
Professor of English and Data Analytics
Washington State University
PO Box 642630 | Pullman, WA 99164-2630
509-335-5540 | matthew.jockers@wsu.edu

Appendix G
Letter from Hoyt Long

November 13, 2020

To the Register of Copyrights:

I am an Associate Professor of Japanese Literature and Director of Graduate Studies in the East Asian Languages and Civilizations Department at the University of Chicago. I also co-direct the Textual Optics Lab, which uses qualitative and computational methods to build large-scale text collections 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 de-tails within a single text.¹ I write this letter in my personal capacity in support of an exemption allowing circumvention of technological protection measures to facilitate text and data mining (“TDM”).

In my own research, I apply computational methods to the study of literature and culture. 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, and how patterns of racial discourse in society at large filter down into literary expression. 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: Richard 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>.

Most recently I have completed a monograph titled *The Values in Numbers: Reading Japanese Literature in a Global Information Age* (Columbia University Press, 2021). This book brings debates around computational literary history to the study of Japan, guiding 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. 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 can augment our understanding of the literary past.

To conduct computational analysis of literary works, one begins with a dataset of text files for the works of interest. Many TDM researchers, including myself, have devoted much time and effort to building corpora of works that are sufficiently comprehensive for such analysis. Often, because of the DMCA's prohibition on circumvention, we build corpora using optical character recognition ("OCR") on scans of print books. But OCR is a deeply flawed method, particularly for the examination of non-English texts. OCR software will examine scanned pages to create a text file that reflects the words it reads on those pages. However, this process also requires human input. Because OCR software is not perfectly accurate when translating print characters into digital text, a human needs to apply manual corrections.

In my experience, OCR accuracy is even worse for non-English texts. Ideographic languages, such as Japanese, are particularly difficult for OCR software. Ideographic languages tend to have many more characters and greater geometric complexity and variation than written English; there are several thousand characters currently in use in the Japanese language, and this number increases to tens-of thousands as one goes further back in time. Print quality also decreases with materials published before the Pacific War, thus further hindering accuracy. These factors make it more difficult for OCR software to correctly identify a given character. Although one can achieve accuracy levels of 90 to 95% with more contemporary publications, this degrades sharply with older materials. Even at 95%, the possibility of introducing systematic errors is still problematic, regardless of language. Given that the ultimate goal of TDM is to analyze the text files generated through OCR, often at the level of individual words, researchers diligently ensure that their text files are error-free. Otherwise, analysis will reflect errors introduced by OCR, preventing accurate insight into the work.

Licensing an existing database for access to a corpus also presents problems, even if it minimizes the issue of error-correction. As is often the case for much of my own research, there is not always a relevant corpus available; in these instances, I am forced to build my own dataset. These platforms also limit the computational methods available. Many commercially and non-commercially available databases only allow use of the platform's proprietary algorithms, with little flexibility. Direct access to the works in the corpus is also limited. For instance, it is impossible to add a particularly relevant work to that licensed corpus. If a database is biased in

some way, I am powerless to correct that bias by adding or excluding works. But cutting-edge TDM research requires that researchers exercise more control. I need to be able to change the computational analysis to ask follow-up questions in response to initial results. Having direct access to the corpus, so that I can use my own algorithms, facilitates this flexibility.

To build one's own corpus, usage of digital literary works would be a superior alternative to conducting OCR on printed books. For instance, e-books need no correction to be useful in analysis aside from removing publisher information. Using e-books, researchers could build corpora much more quickly and painlessly. The expediency of not needing to correct errors would allow for larger, more comprehensive datasets. An individual researcher relying on OCR may feasibly be able to build a corpus of a hundred books within a typical project timeframe, because it can take hours to correct a single scanned book. With a larger team of research assistants, such as the one I've managed for the History of Black Writing project, it has taken us several years to digitize about 1,000 novels, in part owing to the need for manual correction of OCR. With access to e-books, we could have built this collection many times faster and with greater efficiency, allowing us to dedicate more time to materials not available in e-book form. In general, researchers with access to e-books could also build corpora many times this size in the same length of time. Such a drastic increase in the size of the dataset allows a researcher to pull more robust and representative samples of a particular time period or population of writers, but also to make stronger inferences about large scale shifts in literary phenomena over time. Moreover, it allows for results to be evaluated against broader "control" corpora and for finer comparisons to be made between different styles and populations of writers.

Unfortunately, using e-books is often not possible. Many digital text sources are protected by technological measures. DMCA § 1201's prohibition on breaking these protections effectively excludes these digital literary works from analysis by any law-abiding researcher.

Section 1201 has broad implications in the field of digital humanities. Ideally, research questions should dictate the design of a project and the composition of the needed corpus. In reality, this relationship is reversed: ease of access to literary works shapes and limits the questions that are asked. Scholars, knowing the limitations imposed by § 1201, often discard nascent inquiries before a research question can be properly defined. In effect, § 1201 is warping the development of digital humanities as a discipline.

If this exemption were granted, I would be able to make rapid progress in the next few years with a collaborative research project that aims to study trends in contemporary literature across the globe, including works in English, Japanese, Chinese, German, Spanish, Russian, and Portuguese. In coordination with scholars at UC Berkeley, McGill University, and potentially other research institutions, this project will not only develop methods of comparative analysis for the study of world literatures but will also enhance NLP tools for the extraction of linguistic data from texts in each of these languages (e.g., named-entities, dependencies, parts-of-speech). A portion of this work is funded by a National Endowment for the Humanities grant. Essential to

the success of the project will be the efficient creation of digital text collections of several hundred novels for each language. With the DMCA exemption, this project can proceed to the analysis stage much more quickly and put us in a stronger position to fulfill the terms of the NEH award in a timely manner. Given that the tools developed for the grant will be made openly available, a more rapid progress also ensures that other researchers can begin utilizing these tools sooner, strengthening the computational study of non-English literatures and thereby contributing to the growth of this exciting new field.

In sum, I will be able to greatly expand my research to improve our understandings of literature and culture if I were able to use digital texts without fear of liability. I ask that you grant an exemption to § 1201 for text and data mining.

Sincerely,

A handwritten signature in black ink, appearing to read 'Hoyt Long', written in a cursive style.

Hoyt Long
Associate Professor of Japanese Literature
Director of Graduate Studies
East Asian Languages and Civilizations Department
University of Chicago

Appendix H
Letter from Jes Lopez

To the Register of Copyrights:

I am a PhD candidate in the English Department and a Digital Humanities Certificate Graduate from Michigan State University. In my research, I use a combination of text data mining and sentiment analysis to study how authors use sentiment in literature featuring autistic characters. I critically analyze the results of sentiment analysis through a convergence of literary cognitive studies and disability studies. And through these methods and major fields, I look at the visualizations of the research to find how they represent diverse ways of re-reading that can open up textually focused narratives to audiences that prefer more visually focused narratives. In my individual capacity, I am writing in support of an exemption to § 1201 of the DMCA for text and data mining (“TDM”) research.

For my dissertation, I studied novels that feature neuroatypical narrators and neurodiverse narrators, investigating the characters the author has described as autistic or having Asperger’s syndrome. I first analyze these autistic characters through the more traditional method of close reading. Next, I analyze these characters through my method of “scaled reading” which uses sentiment analysis by comparing the text of novels against the “bing” sentiment lexicon of positive and negative words in order to identify parts of the text that have greater or fewer uses of sentiment. I named this method using the word “scaled” in order to differentiate it from other machine learning methods such as distant reading. Whereas distant reading attends to large amounts of novels in direct comparison to find patterns, scaled reading looks to the patterns within a single novel which adds more direct context to traditional close reading. Also, the word “scaled” alludes to being able to see a full novel on a visual graph, in other words to see the shape of the narrative arc. Thus, scaled reading diverges from previously established methods, engaging with sentiment analysis to critically analyze literary texts.

I have conducted my scaled reading analysis on four novels: two which feature a neuroatypical (autistic) narrator, one with multiple narrators where one of the narrators is neuroatypical and the rest are neurotypical, and one with a neuroatypical character and multiple neurotypical characters who are followed through third-person omniscient narration. These novels span multiple genres, including romance, coming of age, science fiction, and mystery. My research indicates that the neuroatypical narrators provide insight into how sentiment is used in a neurotypical and ableist focused society. That autistic characters tend to use more words that are charged with sentiment in order to translate their experiences from a neuroatypical way of being into a form that is understood by a neurotypical majority. And that sometimes their experiences resist translation. Furthermore, in listening to and understanding what is already in society and what we desire in the future, that we should think back to how our relationship with technology has permanently altered our engagements with the world. The machines and algorithms that we interact with constantly throughout our day has changed our neurological constructions in how we seek information and how that influences our perspectives and values. And more closely thinking about how the machine interprets the codes of the text in novels inevitably brings up further conversations about the machine and its place as a “reader.” One might argue that scaled reading is less human reading and more machine reading. Yet the machine is an extension of the human mind in that the codes it “executes” to provide answers to the questions we ask of scaled reading

are all framed around the instructions a human has provided. Accordingly, the machine delivers a precise reply which unerringly picks up everything based upon those instructions given. Thus, to critique a machine for a result is to call out the problematic structures provided by the humans who generated the many different aspects of the instructional code and the society that generates these humans who hold uniquely biased beliefs.

Section 1201 presents significant barriers to my research. Because I am studying novels in which characters are explicitly identified as autistic or having Asperger's syndrome, I am necessarily limited to works published after 1994, when the American Psychiatric Association added Asperger's syndrome to the Diagnostic and Statistical Manual of Mental Disorders in the DSM-IV. This means there are no out of copyright works I can study. The books I am interested in are available either on paper or in the form of DRM-restricted eBooks.

I cannot rely on collections created by other organizations. In particular, HathiTrust is unworkable because it doesn't have an extensive collection of books from the young adult market, which a large portion of books with autistic characters targets. It also lacks works from the genres I have relied on for my analysis, such as romance, coming of age, science fiction, and mystery. Moreover, even if the necessary novels were present, this would still not be a suitable substitute for building my own corpora because I must have access to the full text in order to determine context through the combination of close reading and scaled reading.

Purchasing printed copies of books and then conducting optical character recognition is also unsuitable for this work. My analysis, which looks at the way characters are described, is especially intolerant of errors in the text file. Additionally, scanning printed books, running OCR, and correcting the output is a time-consuming process, and doing so beyond a handful of books would be untenable, especially for a graduate student like myself who needs to complete research on a short time scale and with limited resources.

My dissertation research is in part to build confidence in my analytic technique. Over the next few years, I would like to expand this work beyond four novels to a collection of about 25 post-1994 works with autistic characters. This would allow me to look for similarities and differences in the use of sentiment across genres and characters. I would also like to make comparisons with books that do not feature autistic characters to explore further differences in the use of sentiment. For example, my research thus far suggests that authors increase their use of sentiment when discussing autistic characters in order to translate their neuroatypical way of being into language that can be understood by a neurotypical majority audience. Expanding my research to a larger set of works would allow me to validate this hypothesis.

Without an exemption to § 1201, continuing this research with an expanded set of works will be impossible. As someone who is beginning my career in academia, I lack the resources to scan, OCR, and correct a large number of works. An exemption to § 1201 would enable me to quickly extract the necessary text from purchased eBooks, enabling me to move my research forward.

I urge you to grant this exemption so that I can continue this important research that helps us understand how we as a society think and talk about autism. Without an exemption, my research simply cannot continue.

Sincerely,

A handwritten signature in black ink, appearing to read 'Jes Lopez', with a stylized, overlapping structure.

Jes Lopez

Appendix I
Letter from Andrew Piper



November 17, 2020

To the Register of Copyrights,

I am a Professor in the Department of Languages, Literatures, and Cultures at McGill University. I direct .txtLAB, a laboratory for cultural analytics at McGill, where we explore the use of computational and quantitative approaches for the study of literature and culture. Our aim is to use the tools of data science, network analysis and machine learning to promote a more inclusive understanding of culture and creativity. I am also the editor for the *Journal of Cultural Analytics*, an open-access journal dedicated to the computational study of culture.

Over the past seven years, I have directed a partnership grant, "Text Mining the Novel: Establishing the Foundations of a New Discipline," funded by the Social Sciences and Humanities Research Council of Canada, that brings together over 20 academic and non-academic partners across North America in the humanities, computer science, and industry to facilitate the first large-scale quantitative and cross-cultural study of the novel. The goal of this grant is to establish the foundations of a new discipline that brings together computational approaches to the study of documents alongside the over 2,000-year-old tradition of textual interpretation. Our aim is to train the next generation of scholars to participate in larger debates about data mining and the place of information technology within society.

I write to you now in my personal capacity in support of a §1201 exemption for text and data mining ("TDM"). During my leadership of this research partnership, I have authored numerous articles and books that use TDM to better understand our literary heritage and why literature is such an important aspect of society.¹ In my recent book, *Enumerations: Data and Literary Study*, for example, I draw on a data set of over 20,000 works of fiction and non-fiction from the nineteenth century to try to answer the fundamental question, "What is fiction for?" Rather than read a few great works, by examining a massive amount of writing from the past we can begin to better understand what role fiction plays within society. As I show in that book, one of the principal concerns of the rise of the novel during that period depended not on being more "realistic," but in immersing readers in the sensory, perceptual experiences of characters' lives.

¹ A few examples: Andrew Piper, *Can We Be Wrong? The Problem of Textual Evidence in a Time of Data* (Cambridge 2020); Andrew Piper, *Enumerations: Data and Literary Study* (Chicago 2018); Andrew Piper, "Novel Devotions: Conversional Reading, Computational Modeling, and the Modern Novel," *New Literary History* 46.1 (2015): 63-98; and Sherif Abuelwafa et al. "Detecting Footnotes in 32 million pages of Eighteenth-Century Collections Online," *Journal of Cultural Analytics* (2018).



Learning how to see through the minds of others is what fiction teaches us, an insight which is now empirically grounded not in the selective hand-chosen example, but based on a large, representative swath of the written past.

If TDM can begin to tell us new insights about the longstanding nature of categories like fiction, it can also help us identify widespread biases within the culture industry today. Debates about the whiteness of Hollywood or the centrality of men within contemporary publishing have been hotly debated. In our lab, we have been able to utilize small collections of digitized books or freely available screenplays on the web to study these biases in more empirical detail. For example, in a recent article my student Eve Kraicer was able to estimate the gender distribution of over 26,000 characters in contemporary novels using new techniques in natural language processing. She found that for main characters, which she could count by hand, the ratio of men to women was almost exactly 50:50. When she looked at all characters using TDM, the percentage of women characters was estimated to be 37.8%. Her work highlights how when we take a broader view of culture, social biases re-emerge and in the process reinforce longstanding social hierarchies. These insights would not be possible without the emerging techniques of TDM.

Though computational methods allow me to analyze larger volumes of literary work, building a digitized corpus of works to analyze is still very difficult in the current legal environment. Optical character recognition (“OCR”) of scanned, printed texts is one way to do this. But, due to the imperfect fidelity of OCR software, it introduces errors into the dataset. A researcher relying on OCR faces an uncomfortable decision: to proceed with an erroneous dataset that might prevent meaningful computational analysis, or spend precious time and effort to correct the errors. Notably, the time required to correct errors means that the researcher can’t build as large a corpus as he would if error correction were not an issue. The reduction in corpus scope precludes the kinds of large datasets I need to support the broad questions I address in *Enumerations*. For researchers with fewer resources, this phenomenon blocks entire avenues of research entirely. Too much time is being spent building small corpora when more time could be spent analyzing the contents of existing digital resources.

If it’s difficult to build a large corpus using OCR, what about using a pre-built corpus? There are pre-built corpora, both commercial and non-commercial, available to researchers. But these also have limitations that make them unappealing for many TDM projects. The scope of these pre-built corpora is often opaque. It is difficult to know which books are in a corpus when, as is often the case, direct access to the works underlying the corpus is not allowed. Further, pre-built corpora tend to lack contemporary works, which is a major challenge for me given my interest in studying contemporary issues surrounding equality and representation. Various pre-constructed corpora can only be analyzed using the algorithms provided by the platform that hosts them. These algorithms are also often opaque and prevent me from exercising tight control over the computational methods applied. This lack of control makes it more difficult



to conduct research. For example, if an initial result is surprising, I may need to troubleshoot to ensure that the algorithm is working as intended. But this is not possible when using the proprietary algorithms packaged with licensable corpora.

One example of a platform that does afford some flexibility in algorithmic approaches is HathiTrust's data capsule tool. While HathiTrust is a useful resource, the technical bar for using a data capsule is very high; it requires the user to have well-developed coding skills. For this reason, it's not well-suited for many researchers. Second, the contents and representativeness of the data contained by the HathiTrust is still widely unknown. A great deal of work still needs to be done to better understand this collection's holdings meaning it is not entirely suitable for broad application yet.

Additionally, researchers are often blocked from using their own algorithms to examine a licensed corpus, which is particularly frustrating for those TDM researchers interested in developing novel ways to parse data. Licensed corpora are also subject to differing access permissions, imposing inconsistent levels and types of access on TDM researchers. This variation makes it difficult for a researcher to analyze corpora from different publishers in a consistent manner. The end result is that researchers are not able to construct data sets and their analysis in a systematic way to study real-world problems. Instead research questions are constrained by the license restrictions of library holdings. This is backwards for knowledge discovery.

With these issues in mind, e-books present one of the greatest opportunities for direct access to high-quality text files without requiring a massive time investment. An e-book's text is error-free. Researchers can easily pick and choose a selection of e-books to build a corpus, to which they can directly apply their own, tailored algorithms. Researchers can either apply grant money to purchase copies of desired books or utilize their libraries' existing collections, which reduces cost and access. In order to address concerns about reproduction, it would be straightforward to create registries of datasets whose circulation could easily be tracked. The barrier to this solution is §1201, which effectively outlaws access to e-book text for TDM purposes.

Let me close with an analogy and an opportunity. Imagine if when researchers had the opportunity to map the human genome they were told they had to read all 6.4 billion sequences by hand. Or that only portions of the genome would be available for analysis and researchers did not control which portions. Knowledge about the genetic foundations of life would have been impossible. Today, we are in a similar position with respect to the human textual record. The portions we can access using data-driven methods are insufficient and patchwork in nature. The portions we cannot access are growing by the day. Knowledge of human culture is essential to a healthy society. The current legal framework severely hampers our ability to conduct research within this domain.



If this §1201 exemption were granted, there are numerous projects that would become immediately open to researchers today. In my lab, our number one priority is to build a global collection of literature across a broad array of national and linguistic cultures. So much of TDM and computational work focuses on anglophone documents. And yet we live in a highly connected, richly diverse world of different cultures and sensibilities. Our dream project is to begin to understand how these different cultures tell stories -- where are the fault lines that make one narrative world different from another and where are the lines of commonality, that illustrate for us a common human approach to storytelling? We see this as something comparable to the human genome in its scope. Imagine understanding and comparing all of the stories and sequences from across the world today to arrive at a truly global understanding of the human relationship to creativity and storytelling. Right now this is only possible with a §1201 exemption.

For these reasons, I ask that you grant an exemption for TDM purposes.

Sincerely,

Andrew Piper
Professor and William Dawson Scholar

Appendix J
Letter from Matthew Sag

LOYOLA UNIVERSITY CHICAGO SCHOOL OF LAW

Philip H. Corboy Law Center
25 E. Pearson Street | Chicago, Illinois 60611

Matthew Sag
Georgia Reithal Professor of Law
Associate Dean For Research and Faculty Development
Phone: 312-915-7223
Fax: 312-915-7201
Email: msag@luc.edu



November 16, 2020

To the Register of Copyrights:

Via electronic submission

Dear Register Perlmutter,

I am the Associate Dean for Faculty Research and Development and Georgia Reithal Professor of Law at Loyola University of Chicago, where I am also the Associate Director for Intellectual Property of the Institute for Consumer Antitrust Studies. I am also a member of the American Law Institute.

I write to you in my individual capacity in support of an exemption to 17 U.S.C. § 1201 to enable text and data mining (“TDM”).

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*, the *Journal of the Copyright Society Of the USA*, the *Northwestern Law Review*, and the *Berkeley Journal of Law and Technology*.¹ 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.² I have been a member of the HathiTrust Research Center Advisory

¹ 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).

² Brief of Digital Humanities and Law Scholars as Amici Curiae in Support of Defendants-Appellees and Affirmance, *Authors Guild v. HathiTrust*, 755 F.3d 87 (2d Cir. 2014) (No. 12-04547), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2274832; 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.

Board Since 2016³ 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.

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.⁴ I am the co-founder of ScotusOA.com, a website which applies computational analysis to oral arguments before the U.S. Supreme Court. 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*.

Section 1201 is a barrier for TDM researchers. Through my roles at HathiTrust and Building LLTDM, I have been able to speak with many different TDM researchers. It is evident that § 1201 has a serious negative impact on their research practices. Currently, § 1201 prohibits researchers from incorporating various kinds of digital works, like e-books and DVDs, into their datasets because doing so would require circumventing a technological protection measure (“TPM”). Researchers, who are rightfully concerned about § 1201 liability, frequently curtail their studies because these otherwise useful digital works are protected by TPMs. They instead focus their efforts on low-risk data, such as works in the public domain. Because of the chilling effect of § 1201, the world is deprived of significant research into contemporary cultural works including literature, movies, and tv shows.

TDM is a fair use. Section 1201 presents an unnecessary barrier to TDM: it prevents researchers from making fair use of copyrighted work. *HathiTrust* and *Google Books* confirm that the reproduction of copyrighted works as part of a process of knowledge discovery, as in TDM, is transformative and therefore a fair use of the works.⁵ TDM as a non-expressive use, is a fundamentally fair use that does not infringe the copyright of the underlying works.⁶ To elaborate briefly: The reason why classic transformative uses such as parody, commentary, criticism, are routinely found to be fair use is that, although they reproduce some copyrighted original expression, they do so in circumstances where the risk of expressive substitution is very small. Non-expressive uses, such as TDM research are “quintessentially transformative”⁷ and obviously fair use because they do not communicate the copyright owner’s original

³ 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.

⁴ 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).

⁵ Matthew Sag, The New Legal Landscape for Text Mining and Machine Learning, 66 J. of the Copyright Soc’y of the U.S.A. 291, 293–94 (2019).

⁶ *Id.* at 301.

⁷ Authors Guild, Inc. v. HathiTrust, 755 F.3d 87, 97 (2d Cir. 2014).

expression to the public at all. I discuss why TDM is a fair use in depth in *The New Legal Landscape for Text Mining and Machine Learning*, which is attached.

I ask that you grant the proposed exemption for text and data mining.

Yours sincerely,

A handwritten signature in black ink that reads "Matthew Sag". The signature is written in a cursive, flowing style.

Appendix K
Letter from Rachael Samberg And Timothy Vollmer

November 20, 2020

To the Register of Copyrights:

We are scholarly publishing experts at University of California, Berkeley (UC Berkeley) writing in support of Authors Alliance’s proposed text data mining (TDM) exemption to 17 U.S.C. § 1201’s prohibition against circumvention of technological measures (“Proposed TDM Exemption”). We submit this letter in our individual capacities.

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 navigate the shifting publishing, intellectual property, and information policy landscapes in ways that promote research dissemination, accessibility, and impact.¹

Over the past five years, we have routinely fielded questions from confused or frustrated researchers seeking ways to conduct TDM within legal bounds. Researchers performing TDM face a thicket of legal issues, and a marked absence of community guidance for navigating them. Indeed, a study of humanities scholars’ text analysis needs found that access to and use of copyright-protected texts was a “frequent obstacle” in participants’ ability to select appropriate texts for TDM.²

At the same time, TDM researchers have an appetite for education and training around these issues. We therefore developed a nationally-recognized analysis and workflow to help United States-based scholars and research professionals navigate the law and policy landscape of TDM—including as to copyright, contracts and licensing, privacy law, and ethical considerations (“TDM Legal Literacies”).³ This approach enables researchers to fully and fairly utilize rights-protected works, and disseminate their resulting TDM scholarship broadly. We have also developed and delivered a national training institute to educate and empower digital humanities⁴ researchers and research-adjacent support staff (such as librarians and other professionals).⁵

¹ University of California, Berkeley Library. (n.d.). *Office of Scholarly Communication Services*. Available at <https://www.lib.berkeley.edu/scholarly-communication>

² Green, H., et al., (2016). Scholarly Needs for Text Analysis Resources: A User Assessment Study for the HathiTrust Research Center. *Proceedings of the Charleston Library Conference*. Available at <http://dx.doi.org/10.5703/1288284316464>.

³ Samberg, R. G., & Hennesy, C. (2019). Law and literacy in non-consumptive text mining: Guiding researchers through the landscape of computational text analysis. *Copyright Conversations: Rights Literacy in a Digital World* (pp. 289–315). ACRL. Available at <https://escholarship.org/uc/item/55j0h74g>.

⁴ Digital humanities is a growing academic field concerned with the application of computational tools and methods to traditional humanities disciplines such as literature, history, and philosophy. TDM within digital humanities has been used to conduct research such as understanding how depictions of gender have changed in fiction, evaluating language from body camera footage for evidence of racial disparity, and many other research directions. While the digital humanities would gain greatly from the Proposed TDM Exemption, the benefits of an exemption generalize to all fields of research.

⁵ Samberg, R. (2019, August 14). Team Awarded Grant to Help Digital Humanities Scholars Navigate Legal Issues of Text Data Mining. *Berkeley Library Update*. Available at

On June 23-26, 2020 we welcomed 32 digital humanities researchers and professionals to our institute, *Building Legal Literacies for Text Data Mining*.⁶ The institute—and a follow-on comprehensive open educational collection of training materials, lecture videos, exercises, etc.—has and will continue to build communities of practice to support navigation of the TDM Legal Literacies.

What we have learned from our consultations with researchers and the extensive work we have done to provide training on the TDM Legal Literacies is that: The absence of a § 1201 exemption specifically for TDM research constrains scholarly inquiries into important sociopolitical and humanistic trends and, as a result, inhibits the advancement of knowledge. We can explain this problem by distinguishing two groups of TDM researchers using in-copyright texts:

(1) those who conduct TDM by relying on in-copyright texts that are *not* protected by technological protection measures (TPM), and

(2) those who need to rely on in-copyright texts that *are* protected by TPM.

The researchers in group (1) can proceed with their TDM research because the procedures they need to undertake are considered fair use. TDM research typically involves digitizing or downloading (i.e. reproducing) potentially copyrighted works in order to perform algorithmic extractions upon them. Courts have already found that making such reproductions for the purpose of this type of research constitutes a fair use under U.S. copyright law.⁷ As such, anyone performing TDM on in-copyright works that are not protected by TPM can conduct their research in keeping with copyright law.

However, researchers in group (2)—scholars who are seeking to perform the very same automated processes to answer the very same types of questions as those in group (1)—are currently prohibited from engaging in their research. That is because, while non-consumptive text mining has been found to be fair use, there is no fair use exemption specified in § 1201. Thus, while copying electronic books for the purpose of TDM research is protected by the fair use doctrine, breaking TPM to make that very same copying possible would be unlawful. Fearing that they will violate the law, researchers in group (2) have reported that they abandon or feel compelled to modify their TDM work by resorting to non-TPM texts that do not well serve their research inquiries.

For instance, we have been approached by researchers seeking to analyze recent best-selling novels to compare how gender identity is communicated in contemporary literature. While much

<https://update.lib.berkeley.edu/2019/08/14/team-awarded-grant-to-help-digital-humanities-scholars-navigate-legal-issues-of-text-data-mining/>.

⁶ Vollmer, T. (2020, July 17). What happened at the Building LLTDM Institute. *Berkeley Library Update*. Available at <https://update.lib.berkeley.edu/2020/07/17/what-happened-at-the-building-lltdm-institute/>.

⁷ See *Authors Guild v. HathiTrust*, 755 F.3d 87 (2d Cir. 2014), *Authors Guild v. Google, Inc.*, 804 F.3d 202 (2d Cir. 2015).

of this literature is available in digital form from platforms such as Amazon, most of it is encumbered with technological protection measures that restrict how the content can be accessed and read. So, even if the researcher purchases licensed access to the Amazon digital books, they are unable to conduct text data mining without overriding the TPM embedded in the proprietary Kindle file format. The researcher may decide not to pursue this type of project because they presume there is no legally authorized path to conduct TDM on the Amazon eBooks.

Complicating the matter further is that TDM research teams often cross international boundaries, or use content generated or stored in foreign countries. The laws and directives of various foreign countries permit TPM to be circumvented in the context of TDM research.⁸ For instance, Article 3 of the European Union's Directive on Copyright in the Digital Single Market is interpreted to permit research organizations and cultural heritage institutions to conduct TDM without regard to TPM restrictions,⁹ and cannot be overridden through contractual restrictions. Yet, U.S. researchers collaborating with their European counterparts would not be afforded the same rights under U.S. law, likely quelling U.S. research innovation and discouraging international partnerships.

These legal hurdles do not just deter U.S. TDM research; they also:

- (1) *bias research toward particular topics and sources of data.* In response to legal roadblocks like TPM, some digital humanities researchers have gravitated to low-friction research questions and texts (such as relying only on public domain works not bound by TPM). Restricting research to such sources can skew inquiries, leave important questions unanswered, and render resulting findings less broadly applicable. A growing body of research also demonstrates how race, gender, and other biases found in openly available texts have contributed to and exacerbated bias in developing artificial intelligence tools;¹⁰ and
- (2) *create a perverse incentive for researchers to seek out and use unlawfully "liberated" texts.* It is not a § 1201 violation to work with or conduct TDM on texts for which third parties have illegally circumvented rights management. For example, Carroll (2019) argues that a researcher could legally conduct TDM on content downloaded from Sci-Hub, the massive shadow research library, if such copies were made "only for computational research and that the durable outputs of any text and data mining analysis would be factual data and

⁸ See, e.g. Japan: Copyright Research and Information Center. (n.d.). *Copyright Law of Japan*. Ch. 2, Art. 30-4. Available at <https://www.cric.or.jp/english/cj/cj2.html#art30>; European Union: Directive (EU) 2019/790 of the European Parliament and of the Council on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC. (2019, May 17). *Official Journal of the European Union*. L130/92. Available at <https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32019L0790&rid=1>.

⁹ LIBER & Communia Association. (n.d.). *Articles 3-4: Text and data mining*. Available at <https://www.notion.so/Articles-3-4-Text-and-data-mining-9be17090ebc545b88ed9ac7d39e4e25a>.

¹⁰ Levendowski, A. (2018). *How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem*. 93 Wash. L. Rev. 579. Available at <https://digitalcommons.law.uw.edu/wlr/vol93/iss2/2>.

would not contain enough of the original expression in the analyzed articles to be copies that count.”¹¹ This may incentivize bad actors to break protection measures in the first place, as well as encourage TDM researchers to utilize such unlawful texts. Moreover, TDM researchers concerned about having used unlawfully liberated texts may be less inclined to reveal their research methodology, impeding research transparency and reproducibility.

Certainly, once equipped with the TDM Legal Literacies, researchers and institutional staff are better positioned to understand what the law permits and proceed with TDM research. But merely understanding what is permitted does not go far enough to support vital research, when Section 1201 currently disadvantages TDM researchers who need to rely on TPM-protected content. While other TDM researchers asking the very same queries and performing the very same methodologies can conduct their research lawfully by relying on fair use, anyone who needs to work with TPM-protected materials cannot.

By securing an exemption to 17 U.S.C. § 1201 for literary works and motion pictures, TDM researchers will feel more confident in engaging in research knowing they are permitted to circumvent access controls for their innovative research. Extending § 1201 exemptions to enable TDM on lawfully accessed digital literary works and motion pictures would champion 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.

Thank you for the opportunity to respond to this matter.

Regards,
Rachael Samberg
Timothy Vollmer

¹¹ Carroll, M. (2019). *Copyright and the Progress of Science: Why Text and Data Mining Is Lawful*. 53 UC Davis L. Rev. 893. Available at https://lawreview.law.ucdavis.edu/issues/53/2/articles/files/53-2_Carroll.pdf.

Appendix L
Letter from Dan Sinykin



November 29, 2020

To the Register of Copyrights,

I am an assistant professor of English at Emory University, with a courtesy appointment in Quantitative Theory and Methods. I am writing a book on the conglomeration of the United States publishing industry, under contract with Columbia University Press, for which I require computational analyses of thousands of novels published since 1945. I also teach computational analysis; presently, I am teaching Practical Approaches to Data Science with Text. I am writing in my individual capacity in support of an exemption to Section 1201 of the DMCA for the purposes of text and data mining research (TDM).

My book-in-progress, *The Conglomerate Era*, asks how the conglomeration of US publishing changed fiction. In the 1950s, almost every publisher in the US was independent. By 2000, only six multinational media conglomerates controlled a large majority of the sector. How can I make arguments about change at such scale? I cannot read enough fiction to make judgments myself. Instead, to detect patterns of change across thousands of novels across decades, I use TDM. TDM methods are exciting; they promise to expand considerably our understanding of literary history. But, at present, scholars in my field (post-1945 literature) are severely limited by Section 1201 of the DMCA.

The only option I and fellow scholars have is to use HathiTrust to build sufficient datasets. 1201 makes it otherwise impossible. I am grateful that Hathi exists, because otherwise I would be unable to pursue my research at all, but Hathi has considerable limitations. I can access Hathi's collection because Emory is a partner institution. Colleagues whose employers are not partner institutions, or who are independent scholars, lack access. To access works under copyright through Hathi, I need my own data capsule, a secure virtual computing environment. Hathi's data capsules are cumbersome. Navigating them takes much more time than does navigating a standard computer today. Opening and closing windows, accessing files, and other basic tasks require patience. One must also navigate between data capsules' "secure" and "maintenance" modes. In secure mode, internet is disabled. So if I need the internet for any reason while working in secure mode—for example, if I'm working with code, as I often do, and must do in secure mode to work with text under copyright, but need to debug a bit that's not working, as is common, by searching the web—I need to switch to maintenance mode. Switching between modes can take a few minutes. While, individually, these delays might sound minor, in aggregate they make my work two, three, or four times slower than it would be otherwise. Further, the challenges of working in data capsules are enough to inhibit most scholars from even attempting TDM in my field of study.



Hathi data capsules, further, have limited computing power. When I initially launched my capsule, I ran into limits with fairly basic analyses; for example, I had to cut some very long novels from my corpora because I did not have enough computing power to process them in my models, adding an artificial bias in my corpus. I won one of Hathi's Advanced Collaborative Support awards for the 2019-2020 year, granting me enhanced computing power. Even still, I do not have enough to run the most advanced and demanding models, like certain neural networks and transformers.

Beyond the limits of data capsules, Hathi's holdings themselves limit my research. Hathi is not comprehensive. Its holdings are the holdings of select university libraries, which do not acquire all fiction equally. There are, thus, vast gaps in Hathi. Worse, scholars do not yet know the contours of the gaps. This means that my findings based on Hathi's holdings are necessarily provisional and partial.

In my capacity as a professor, 1201 inhibits my ability to teach TDM. For my students to use TDM in our field of post-1945 literature, they need proficiency with Hathi's data capsules. In most cases, this is too large a barrier to overcome. It takes too much time to teach students of literary studies, whether undergrads or grads, to use Hathi over the course of a semester. In practice, this means students turn to older periods where literature is not under copyright or to text they can acquire from the internet. So long as students do not pursue TDM in the field, the field will be stunted.

If I were exempt from 1201, I would be able to write a better, truer book about conglomeration. Maybe more profoundly, I would be able to teach TDM to the next generation of scholars who would transform our field of study. I am working to build the foundation for this future work. Laura B. McGrath and I have co-founded the Post45 Data Collective, which will launch in coming months. We are building a system of peer review and a single home for metadata such as author gender and race, MFA site, thesis advisor, and titles' publisher, prizes, literary agent. As of now, we have to cross-reference this metadata with Hathi IDs for scholars to study the metadata with the text. Exemption from DMCA 1201 would allow researchers far greater ease to do research with the data in the collective, which we believe will be transformative for our collective knowledge of literature and literary history.

Sincerely,

A handwritten signature in cursive script, appearing to read 'DS'.

Dan Sinykin

Assistant Professor of English; Courtesy Appointment in Quantitative Theory and Methods
Emory University

Appendix M
Letter from Lauren Tilton And Taylor Arnold



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

To the Register of Copyrights,

As the directors of the Distant Viewing Lab¹, we write to support an exemption to DMCA § 1201 to enable non-consumptive text and data mining (TDM) of in-copyright materials. The current policy is detrimental to scholarly research on media, particularly moving images. The negative impact is widespread across fields from data science and digital humanities to communications and media studies to computer science and statistics. The current policy not only curtails domain-specific research but limits innovative interdisciplinary scholarship that is a hallmark of American higher education. Below we demonstrate several areas where the current law adversely affects research in media, specifically audiovisual data that is distributed primarily in the form of DRM protected DVDs.

The Distant Viewing Lab uses and develops computational techniques to analyze visual culture on a large scale. Bringing together our expertise in data science and digital humanities, we develop tools, methods, and datasets that can be re-used by other researchers while making disciplinary interventions in areas such as film and media studies. Our scholarship has received grant support from the Mellon Foundation and National Endowment for the Humanities. Yet, the scope of our research is incredibly restricted due to DMCA § 1201. Essentially, we cannot study the 20th and 21st century visual culture, which is only available through media formats such as DVDs and protected digital files. Below we demonstrate several areas where the current law adversely affects research in media, specifically audiovisual data.

Large scale TDM has opened up new theoretical frameworks. Along with concepts such as distant reading and distant listening, a quickly growing area of research is *distant viewing* (DV). A theory and method for large scale computational analysis of visual materials developed at the intersection of data science and digital humanities, DV harnesses the power of computer vision to answer domain and interdisciplinary questions such as:

- Which new methods and techniques do we need for analyzing large sets of unstructured audiovisual data at scale? (Data Science/ Digital Humanities)
- Which ways of viewing, and therefore which algorithms, do we need to build to conduct various kinds of DV? (Computer Science)
- How do we understand the algorithms and the results from DV? (Data Science/ Statistics)
- What are the politics of representation during the Network era of American Television? (TV Studies)
- What is the visual style of film auteurs? What elements make for a Hollywood blockbuster? (Film Studies)
- To what degree do TV and film send messages impact film style and vice versa? (Media Studies)
- How does film and television communicate meaning visually and aurally? (Communications)

All of these questions require access to large amounts of data. DMCA § 1201 is a major barrier because it eliminates entire areas of study, including much of the available materials in 20th and 21st century film and television. It also means that two of the most culturally, socially, and politically powerful forms of media in the world – US film and television – cannot be studied using computational methods. It also limits methodological innovation at the intersection of AI, machine learning, and audiovisual data. This is particularly pressing for moving images, which is an innovation of the 19th century and became prominent in the 20th century, with most of the work produced currently being held under copyright. Due to market consolidation and the formation of

¹ For more about the Distant Viewing Lab, please visit www.distantviewing.org.



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

the media industries in the 20th and 21st century, the majority of moving images are held and distributed by for-profit multinational corporations. A result is the almost complete foreclosure of computational research on audiovisual data in the United States, except for rare research groups that get special access to materials from a company or have incredible financial resources to pay for access. The comparison with the progress on scholarship with other visual forms such as drawings, illustrations, and paintings from fields such as Art History further demonstrates the impact of DMCA § 1201. Work in the field of Art History has received tens of millions of dollars of funding from organizations such as the Terra Foundation and resulted in new journals such as the *International Journal for Digital Art History*. The lack of access to data has financial, institutional, and international implications, to which we now turn.

We also want to add that this law puts 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 are legally barred from this kind of research. 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.

If there was an exemption for DMCA § 1201, we could eagerly pursue several projects. Understanding how media sends messages as well as which messages they send is an important area of communications and media studies. Yet, until an exemption is granted, the negative impacts include curtailing new work at the intersection of the data science and the digital humanities, thereby placing us at a disadvantage globally.

Sincerely,

Lauren Tilton
Assistant Professor of Digital Humanities
Department of Rhetoric & Communication Studies
Director, Distant Viewing Lab
University of Richmond

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

Appendix N
Letter from Ted Underwood

Champaign, Oct 28, 2020

To the Register of Copyrights:

I am a professor in the School of Information Sciences at the University of Illinois at Urbana-Champaign and also hold an appointment in the Department of English in the College of Liberal Arts and Sciences. I have authored three books about literary history, including *Distant Horizons*, *Why Literary Periods Mattered: Historical Contrast and the Prestige of English Studies*, and *The Work of the Sun: Literature, Science and Political Economy 1760-1860*. I am writing in my individual capacity to support an exemption to section 1201 of the DMCA to enable text and data mining (TDM) research.

My research uses TDM to explore literary patterns that become visible when many books are considered across long timelines, often spanning centuries. By analyzing large numbers of works and using TDM to ask precise questions, we can observe trends such as the marked decline in fiction written from a first-person point of view that took place from the mid-late 1700s to the early-mid 1800s, the weakening of gender stereotypes, and the staying power of literary standards over time.

Without access to large collections of works and modern text and data mining methods, this kind of research would be impossible. There are existing collections provided by organizations such as HathiTrust that have been immeasurably beneficial to my work. However, these collections are incomplete. HathiTrust, for example, tends to reflect the holdings of the major academic libraries with which it partners, and therefore tends to be less inclusive of more popular, modern books. Missing works can severely impede or outright foreclose some avenues of research. For example, I would like to pursue questions about romance fiction, but the lack of such works in HathiTrust's collection makes this impossible.

Even when works are available in collections that are currently available for text and data mining, conducting TDM research can be extraordinarily difficult. For example, HathiTrust’s “data capsules” allow researchers to conduct customized research but present significant workflow challenges. All work must be done within the capsule, with only derived data being allowed out when the researcher is finished. This presents both technical and workflow challenges because the code I use to analyze works can only be refined outside the capsule, but can only be tested inside the capsule (with the “door” to the outside world firmly shut). This means I often have to go in and out of the capsule as many as a hundred times before I have effective analytical tools. Each transition costs me about five minutes of work time (since there’s a lag, for instance, in shutting the “door.”)

In any case, HathiTrust is based on the collections of academic libraries, which tend to collect a specific subset of fiction (emphasizing works with literary aspirations). For some types of research, such as questions about romance fiction, the only viable path forward is to build my own collection because digital libraries do not adequately cover this genre. However, § 1201 prevents me from using electronic copies of books I have purchased to conduct my research because I would have to bypass DRM to access and analyze the text. Without access to electronic versions of the texts, the only option is to manually scan and conduct optical character recognition (OCR) on printed books. This, however, is simply not a feasible path forward. Scanning and OCRing is an intensely laborious process that would be impossible when thousands of works are required to investigate a question.

OCR also introduces significant errors, which is especially problematic for research that spans time periods, genres, or languages. Different fonts, character sets, and printing technologies systematically produce different kinds of errors. This raises serious concerns about the validity of research done using text from scanned sources, where observations could be more reflective of errors in the process of creating electronic versions of the texts than of actual patterns in the works. In fact, this problem is so substantial that I have recently received a grant of \$73,122 from the National Endowment for the Humanities to study errors in optically transcribed text and their consequences (Broadening Access to Text Analysis by Describing Uncertainty, Grant No. PR-268817-20).

But in order to measure the risk of error caused by optical transcription, we need to pair optically transcribed texts with clean, manually transcribed versions of the same books. (That way we can run parallel experiments in the two corpora, and measure the distortion produced by OCR.) Since I only have access to a large collection of clean text through sites like Project Gutenberg—which tend to end where copyright begins—I currently have no



good way even to measure the amount of distortion caused by OCR error in most of the twentieth century.

If this exemption were granted, I would develop corpora of several thousand volumes of fiction and biography stretching across the twentieth century. A small portion of those volumes would intentionally overlap with optically transcribed volumes in HathiTrust. By pairing optically-transcribed and clean versions of the same texts, I would be able to estimate the level of error involved in using the optically-transcribed texts in digital libraries. Those libraries will remain important, because researchers will never be able to purchase all the texts they need to understand contemporary literary production: I will still have to use the sealed data capsules run by HathiTrust to mine really large collections. But if I could also mine the texts of volumes I had purchased, I could develop supplementary collections (for instance of science fiction or romance fiction, or missing works by Toni Morrison) to fill gaps in the academic libraries that created HathiTrust. By pairing my personal library with the larger digital collection, I could get a much better understanding of the range and diversity of twentieth-century literary culture.

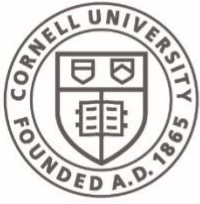
I ask that you grant an exemption to § 1201 to allow circumvention for text and data mining. This exemption will empower researchers like me to tackle important questions that will otherwise remain unanswered.

Sincerely,

A handwritten signature in black ink that reads "Ted Underwood". The signature is written in a cursive, slightly slanted style.

Ted Underwood
Professor of English and Information Sciences
University of Illinois, Urbana-Champaign
Champaign, Illinois, USA

Appendix O
Letter from Melanie Walsh



Cornell CIS
Information Science

Melanie Walsh
Gates Hall
Cornell University
Ithaca, NY 14853
infosci.cornell.edu

Dear Librarian of Congress:

I am writing in support of an exemption to Section 1201 of the DMCA for the purposes of text and data mining research. I am currently a Postdoctoral Associate in Information Science at Cornell University¹, where I use computational methods to study literature and culture—a growing research area known as digital humanities and cultural analytics. I also design and teach classes in this area. For example, in the spring of 2020, I taught “Introduction to Cultural Analytics: Data, Computation, & Culture,” a course that introduces humanities students to a programming language (Python) for the purposes of studying books, songs, social media posts, and other cultural materials. I am advocating for an exemption to Section 1201’s anti-circumvention provisions because they detrimentally impact my research and my teaching, as well as the wider field.

In the last twenty years, text mining methods have brought revolutionary insights to literary scholarship, because they have allowed researchers to study trends across thousands and even millions of books—many more books than a single critic could ever read in a lifetime. These methods hold particular promise for understanding literary trends in the twentieth- and twenty-first centuries because rates of publishing have exploded in this period. For example, *Early English Books Online*, which includes nearly every extant work published in the British Isles and North America between 1470-1700, contains 146,000 books. By contrast, more than 300,000 new print books were published in the U.S. in 2013 alone.² As the number of books published per year increasingly outstrips what an individual critic can glean from human reading, making sense of large-scale cultural patterns has become all but impossible without the help of computers.

Despite the special suitability of computational methods for the twentieth- and twenty-first centuries, this is the literary period that has arguably received the *least* amount of computationally-assisted critical attention, largely due to Section 1201. Because pre-copyright texts are so much easier to access, digital humanities research and teaching have both become strongly biased toward pre-1925 texts. There are many post-1925 research questions that I have personally not pursued because Section 1201 has chilled me from pursuing them. For example, I have studied how #BlackLivesMatter tweets quote the novelist and civil rights activist James Baldwin, using computational methods to archive and analyze these tweets. But I have not attempted a computational analysis of Baldwin’s literary corpus itself (1949-1985) because these texts are so difficult to access due to Section 1201. Similarly, I am eager to compare collections of novels published by authors who graduated from different MFA writing programs—a project idea inspired by literary critic Mark McGurl, who has famously traced the influence of MFA programs on contemporary literature—but again Section 1201 makes it nearly impossible for me to access these collections in the correct format for text mining purposes.

It is true that there are a few alternative ways of accessing in-copyright texts other than circumventing DRM on e-books, but these alternatives are not feasible for me as an early career researcher and as a teacher of introductory classes and beginner programmers. For example, some digital humanities scholars, in lieu of bypassing DRM, hire undergraduate and graduate students to scan physical books, use Optical Character Recognition (OCR) technologies to convert the scans to text files, and manually clean the resulting text files

¹ I am making this statement in a personal capacity, not on behalf of my employer.

² See, for example, 2013 publishing reports from Bowker:

<http://www.bowker.com/news/2014/Traditional-Print-Book-Production-Dipped-Slightly-in-2013.html>

since they often have errors—a process that can take 2-10 hours for each book depending on length and quality of the scans. However, as an early career researcher, I do not have my own research funds, and I do not have my own undergraduate or graduate student advisees. The fact that OCR is more expensive and time-consuming than breaking DRM is not merely an inconvenience. It actively prevents me from researching and teaching about literary culture after 1925.

There are two other alternatives for accessing in-copyright texts through the HathiTrust Digital Library, but these are not viable alternatives to circumventing DRM, either. First, HathiTrust makes available “extracted features,” or word counts per page, for all the volumes in their collection. Second, HathiTrust can provide member-affiliated researchers who fill out a lengthy application with a “Data Capsule,” a remote computing environment that has access to their in-copyright texts. While I applaud HathiTrust for these efforts and for their recognition that copyright poses an obstacle to research, these are not viable alternatives because 1) “extracted features” are not sufficient for the most cutting-edge natural language processing techniques, which often require syntax and not simply word counts 2) HathiTrust “Data Capsules” are currently so complicated and difficult to work with that they are not feasible for my own research nor for classroom use 3) the HathiTrust Digital Library does not contain every published book, and it is especially deficient in twenty-first-century holdings (even twenty-first-century blockbusters such as Stephenie Meyer’s *Twilight* and Suzanne Collins’s *The Hunger Games* are not included in the library’s collection).

The bias toward out-of-copyright texts not only prevents us from better understanding our own recent past and present, but it also contributes to already existing racial and gender biases within the field. Because the majority of available, digitized, pre-1925 texts are authored by white men, the focus on pre-copyright texts further reinscribes white men as the center of the field and further marginalizes women and people of color. When designing curriculum in the past, I have personally run up against Section 1201 as an obstacle that prevents me from making my syllabus more diverse and inclusive. For example, when I teach text mining methods, I want to draw on collections of twenty-first-century writings by Asian American authors or Latinx authors, collections that have resonated well with students in my non-computational literature courses. But because of copyright law and Section 1201 (as well as the deficiencies of OCR and HathiTrust explained above), I cannot teach these collections or design the class the way I want to. This is especially damaging for my course “Introduction to Cultural Analytics,” because many students enroll in this course specifically because they have had prior negative experiences with programming that often stem from gender or racial biases in computer science fields, making it all the more urgent to foster an inclusive classroom environment.

I would like to close with one final observation about how Section 1201 harms the production of knowledge in my field. I have become aware that some researchers do bypass DRM for text mining purposes and choose to disregard Section 1201. But even in these instances, Section 1201 continues to prevent the spread of knowledge because it makes researchers reluctant to share the details of their methodology and impossible for other researchers to replicate their results. When I was just starting out as a graduate student in this field, I experienced some of the harms of this implicit self-censorship. As I began to read articles that used computational methods on in-copyright texts, I tried to figure out how the authors accessed the in-copyright texts, but it was never explicated in the articles themselves. Only months and years later, through private conversations with senior scholars, did I discover that these researchers likely circumvented DRM on e-books. Such whisper networks are antithetical to the production and dissemination of knowledge.

In conclusion, I believe that the exemption to Section 1201 proposed here will offer a critical path forward for text and data mining research. Based on my personal and professional experience, I am confident that it will help accelerate and diversify knowledge about computational methods and twentieth- and twenty-first-century culture.

Sincerely,

A handwritten signature in cursive script that reads "Melanie Walsh". The signature is written in black ink on a light-colored, slightly textured background that resembles a piece of paper or a card.

Melanie Walsh | +1 (512) 762-7259 | melanie.walsh@cornell.edu

Appendix P
Letter from Henry Alexander Wermer-Colan

Henry Alexander Wermer-Colan
Scholars Studio, Temple University
1900 N. 13th Street, Philadelphia, PA 19122
781-264-1992; alex.wermer-colan@temple.edu



November 10th, 2020

To the Register of Copyrights,

I am writing in support of an exemption to the anti-circumvention provisions of the Digital Millennium Copyright Act to enable researchers like myself and the students and faculty I support to pursue their academic work in text mining and data analysis. I am a Digital Humanities Postdoctoral Fellow with a Ph.D. in English literature and a specialization in text mining and literary study. In my role as a coordinator of digital scholarship across the curriculum at Temple University Libraries' Loretta C. Duckworth Scholars Studio, I support students, librarians, and faculty in the development of research and teaching projects involving data curation and analysis. I write today in my individual capacity to request that you grant an exemption to § 1201 to enable this important field of scholarship.

Throughout my doctoral studies at The Graduate Center of the City University of New York, which focused on Euro-American twentieth-century literature, and my postdoctoral research and work developing projects digitizing and analyzing cultural history through data, I have faced obstacles imposed by the DMCA for accessing the vast set of texts and cultural material produced over the last hundred years. I was unable to include sophisticated data analysis methods in my dissertation because there was no way, through my university, available library databases, or other means to build a representative dataset of post-WWII literature. During my postdoctoral fellowship, I have worked in Temple Libraries to digitize twentieth-century literary texts. This process involved collaborating with the Digitization and Metadata Services department to purchase twentieth-century canonical novels, de-bind the books, scan them through a sheet-feed scanner, convert the scanned images through Optical Character Recognition software into machine-readable text, and then, using automated and manual methods, fix errors produced during the conversion into machine-readable text as well as clean those texts of paratextual information such as copyright statements and other front matter. This process is so laborious that even at a R1 university (doctoral degree-granting institutions with very high research activity) like Temple, with multiple library departments and half-dozen staff and student workers contributing to various stages of the project, we have only been able to digitize approximately 500 books over a three-year period.

Furthermore, we have faced continuous obstacles to making these texts available to researchers at Temple University. For researchers in text mining, it is very difficult to do sophisticated work without access to the full texts, including the ability to manipulate these texts through complex algorithmic processes tailored to the particular purposes of any given research project. Disaggregated texts that do not enable any sort of consumptive access greatly reduce the complexity and range of options for the researchers' analysis. While certain forms of text mining, such as topic modeling, can be somewhat successfully achieved on disaggregated texts, recent innovations in text mining and machine learning for generating predictive models like word embeddings require full-text access to the corpora.

To address the problem of access to full texts, we went through the equally laborious process of ingesting the corpora into the HathiTrust Digital Library, whose Research Center has the computing infrastructure to provide researchers with virtual access to a Data Capsule where, under controlled and thus onerous conditions, researchers can conduct computational analysis on full-text files. However, HathiTrust's repository is only available to students and faculty at member institutions, limiting the corpora to only a couple hundred institutions in the world. Beyond this form of access, on a case-by-case basis, I have worked to create extracted features datasets with limited applicability, while collaborating with specific researchers who have sent me programming scripts so that I can run the processing on the corpora before providing the results to the researcher. This iterative process is so time-consuming that most researchers have decided to find other means of acquiring the texts or simply abandoned their projects.

The obstacles to digitization have also adversely impacted the representativeness of available databases. We are ingesting materials into the HathiTrust Digital Library, as we've identified that the vast majority of materials held in Temple Libraries' Special Collections Research Center's Paskow Science Fiction Collection are not contained in HathiTrust. Databases are also siloed repositories, and research access for text mining to one database does not enable researchers to conduct research in a holistic process across multiple databases. As these databases develop proprietary corpora and text mining tools for their corpora, they continue to further silo researchers into limited available datasets. As a result, these balkanized databases both lack datasets representing the diversity of cultural production in the twentieth century by underrepresented groups and restrict researchers to relatively random sets of that data in each database. This problem could be overcome if it were feasible for researchers to build suitable corpora for their own research, rather than relying on proprietary datasets.

The obstacles imposed by the DMCA's anti-circumvention provisions on researchers seeking to access cultural data across the disciplines cannot be overestimated. For senior academics, the problem is still so significant that in the field of literary study, the most significant work of scholarly study on literature at scale, Ted Underwood's *Distant Horizons: Digital Evidence and Literary Change* (2019), still depends on a very small set of literary corpora in genre fiction. Underwood's chapter analyzing science fiction, for instance, only considers approximately 300 books, amounting to far less than one percent of the total books published in the science fiction genre in the twentieth century. Beyond the obstacles imposed on major researchers in the field, who might have access to corpora developed at digital humanities centers for the limited purpose of in-house research, the obstacles imposed on early career researchers like myself and the students I support are so significant that most scholars in the field simply choose to not do this sort of research on twentieth-century materials. The amount of work required to build the corpora they would need, and the limits on their ability to access representative corpora, ensure their research will produce relatively meaningless results within the available timeframe. I regularly consult with undergraduate and graduate students who want to study twentieth-century literature and culture, and I feel it necessary to advise them on how they can develop research projects that think through these theoretical problems, while prototyping research projects we hope can be achieved at scale if the law and the infrastructure supporting digital literary study were to change.

If this exemption were granted, I would be able to scale my digital humanities research on twentieth-century literature by building a corpus of canonical works for text mining. My current research is limited to an unrepresentative dataset, and I hesitate to publish my findings when I doubt that they will be conclusive. The research I could then publish would, I believe, contribute greatly to broader conversations in academia and popular culture around the history of genre fiction in the twentieth-century. Beyond the immediate changes this exception would enable for my research, I believe I would begin to radically rethink my research plans around this exception, considering new research questions I'd previously avoided because I knew I couldn't answer them adequately. Likewise, in my consultations with students and faculty, I would be able to support their research in a new and direct way, opening up more possibilities for their research and new avenues for them to pursue their research.

For these reasons, granting an exemption to enable researchers to efficiently produce large-scale corpora for non-consumptive research would revolutionize our understanding of cultural history in the twentieth century. Previous studies would be revealed as radically narrow and short-sighted in their analysis of small datasets, and early and advanced career scholars alike would be able to finally explore through computational methods the complex patterns and textures our culture has produced in the written word. I sincerely believe these changes can be made in a wise and careful fashion. I strongly believe academic libraries and researchers can develop protocols and workflows for developing research corpora that, rather than imperiling sales of individual books, would actually improve sales of those books. Researchers want to buy books and read them, the better to understand what they see when they look at these texts at scale through the lens of computational algorithms. But the current impositions on text mining at scale discourage many students and scholars from studying these books at all. Furthermore, by limiting their access to these texts, algorithms being developed by researchers will lack these important texts in their generative models. We will continue to ignore underrepresented works by diverse writers who flourished in the twentieth and twenty-first century as they were finally allowed more of a voice in the publishing industry. As a result, the models upon which our contemporary culture depends to design our predictive texts and to classify the sentiments and meanings of writing and speech will remain woefully biased.

I ask that you grant an exemption to § 1201 to allow circumvention for text and data mining to enable myself, my students, and my colleagues to finally ask these important questions about contemporary culture that will otherwise remain unanswered.

Sincerely,

Henry Alexander Wermer-Colan
Digital Humanities Postdoctoral Fellow
Temple University Libraries, Loretta C. Duckworth Scholars Studio