ach 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?

- 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,

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

UNIVERSITY OF CALIFORNIA, BERKELEY

BERKELEY • DAVIS • IRVINE • LOS ANGELES • MERCED • RIVERSIDE • SAN DIEGO • SAN FRANCISCO SANTA BARBARA • SANTA CRUZ

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 objects[1] 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 Saywer*), 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

[1] See, 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 McConnaughey, Jennifer Dai and David Bamman (2017), "The Labeled Segmentation of Printed Books," EMNLP 2017.

[2] See: 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,

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

GRAMBLING
STATE UNIVERSITY
GRAMBLING, LOUISIANA

Department of English
Jeanes Hall, Rm. 126
clawsonj@gram.edu

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 M. Clawson

James Clawson, Ph.D.
Grambling State University

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

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 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,

Eric Hoyt

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/

WASHINGTON STATE UNIVERSITY

College of
Arts and Sciences

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,

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

THE UNIVERSITY OF CHICAGO
HUMANITIES

DEPARTMENT OF EAST ASIAN LANGUAGES AND CIVILIZATIONS

1050 EAST 59TH STREET
CHICAGO, ILLINOIS 60637

TELEPHONE: 773-702-1255
FAX: 773-834-1323

WEB: humanities.uchicago.edu/easian
E-MAIL: ealc@humanities.uchicago.edu

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 de-creases 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,

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

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,

Jes Lopez

McGill | Department of Languages, Literatures, and Cultures | Département de langues, littératures et cultures

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

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 2016[3] 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 HRTC 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,

Matthew Sag

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 humanities[4] 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/clj/cl2.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.

EMORY UNIVERSITY

Department of English

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 larger 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,

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

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 Lab[1], 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.

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 assess 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 C Tilton

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

Taylor B. Arnold

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

ILLINOIS

SCHOOL OF INFORMATION SCIENCES
501 E. Daniel St., MC-493
Champaign, IL 61820-6211

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,

Ted Underwood

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

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 University[1], 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,

Melanie Walsh

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

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