



Open Library of Humanities

## Access and Advocacy: Text & Data Mining and DMCA §1201

Quinn Dombrowski, Stanford University, US, [qad@stanford.edu](mailto:qad@stanford.edu)

Lauren Tilton, University of Richmond, US, [ltilton@richmond.edu](mailto:ltilton@richmond.edu)

---

Significant work in DH has focused on access, labour, and infrastructure. They fundamentally shape the field, from who is included to what we study (for example, see Graban et al 2019, McGrail et al 2022, and Losh and Wernimont 2018). Often less discussed are the legal systems that impact the field. Laws defining data use including ownership, access, circulation, and privacy shape the analytical possibilities and differ according to geography and nation state. In this article, we address the impact of a specific United States legal framework governing access to data and outline advocacy to change the legal code to facilitate DH research. We specifically discuss the Association for Computers and the Humanities (ACH) work to support a change in text and data mining policy by supporting an effort led by Samuelson Law, Technology & Public Policy Clinic and Authors Alliance. This policy, section 1201 of the US copyright code, imposes a set of restrictions on how people can use digital materials whose access is mediated using technical protection measures, including digital rights management (DRM) code—even when the use would otherwise be permitted by fair use. (Fair use in US copyright law is roughly equivalent to fair dealing in Canada, although inherently more open through not enumerating a specific list of exemptions.) By collaborating with one another and with legal practitioners, DH scholars can play a significant role in changing the legal regime that protects multinational corporations and limits humanities research. We end with a discussion of next steps and how US-based scholars can get involved in changing this legal landscape.

---

D'importants travaux en DH se sont concentrés sur l'accès, l'emploi et l'infrastructure. Ils façonnent fondamentalement le domaine, de qui est inclus à ce que nous étudions (par exemple, voir Graban et al 2019, McGrail et al 2022, et Losh et Wernimont 2018). Les systèmes juridiques qui ont un impact sur le domaine sont souvent moins discutés. Les lois définissant l'utilisation des données, y compris la propriété, l'accès, la circulation et la confidentialité, façonnent les possibilités d'analyse et diffèrent en fonction de la géographie et de l'État-nation. Dans cet article, nous abordons l'impact d'un cadre juridique spécifique des États-Unis régissant l'accès aux données et décrivons les actions de plaidoyer visant à modifier le code juridique pour faciliter la recherche en DH. Nous discutons spécifiquement du travail de l'Association for Computers and the Humanities (ACH) pour soutenir un



changement dans la politique d'exploration de textes et de données en soutenant un effort mené par Samuelson Law, Technology & Public Policy Clinic et Authors Alliance. Cette politique, section 1201 du code des droits d'auteur américain, impose un ensemble de restrictions sur la façon dont les gens peuvent utiliser les matériaux numériques dont l'accès est médiatisé par des mesures de protection technique, y compris le code de gestion des droits numériques (Digital Right Management ou DRM)—même lorsque l'utilisation serait autrement autorisée par l'utilisation équitable. (L'utilisation équitable dans la loi des États-Unis sur le droit d'auteur est à peu près équivalente à l'utilisation équitable au Canada, bien qu'elle soit intrinsèquement plus ouverte du fait qu'elle n'énumère pas une liste spécifique d'exemptions). En collaborant entre eux et avec les praticiens du droit, les chercheurs en sciences humaines peuvent jouer un rôle important dans la modification du régime juridique qui protège les multinationales et limite la recherche en sciences humaines. Nous terminerons par une discussion sur les prochaines étapes et sur la manière dont les chercheurs basés aux États-Unis peuvent s'impliquer dans la modification de ce paysage juridique.

---

## Introduction

Significant work in DH has focused on access, labour, and infrastructure. They fundamentally shape the field, from who is included to what we study (for example, see Graban et al 2019; McGrail, Nieves, and Senier 2021; Losh and Wernimont 2018). Less discussed are the legal systems that impact the field. Laws constraining data use including ownership, access, circulation, and privacy shape the analytical possibilities and differ according to geography and nation state. In this article, we describe efforts by the Association for Computers and the Humanities (ACH), the US-based professional association for Digital Humanities, to advocate on behalf of its members for more reasonable laws and positive outcomes to court cases that directly impact the work that scholars are able to do. We begin with an overview of text and data mining within Digital Humanities, as well as the relevant legal landscape in the US, before describing how ACH has engaged with the legal process at key moments to advocate for its members' right to legally do this kind of research. We then focus on a recent example, where ACH and individual Digital Humanities scholars worked with the Samuelson Law, Technology & Public Policy Clinic and Authors Alliance to create an exemption to section 1201 of the Digital Millennium Copyright Act (DMCA) in order to facilitate text and data mining. We then discuss the implications of the exception, which was granted in 2021, and conclude by reflecting on the role scholars can and should play in the increasingly complex regulatory landscape to ensure that commercial interests do not lead to undue restrictions on research.

## The legal and technical landscape of text and data mining

Text and data mining has been a part of the broad interdisciplinary field of Digital Humanities since its inception. Once known as humanities computing, "Digital Humanities" emerged as researchers sought ways for computers to analyze and read texts (Hockey 2004). (Computational analysis is not, and never has been, the entire extent of Digital Humanities, but it is an area of the field that is particularly likely to push legal boundaries.) Punch cards and magnetic tape led to more powerful microprocessors and then CPUs that increased the scale and speed by which researchers could process data. As researchers built machine-readable corpora, scholars from areas such as classics, linguistics, literary studies, and related fields focused on canonical texts such as the Bible and Shakespeare. Initiatives such as the TEI emerged in the late 1980s to support the representation of texts in digital form and add analytical nuance (for more about their history, see their website at <https://tei-c.org/about/history/>). Initiatives such as #transformDH and scholars coming from cultural studies have made an effort to open text analysis and digital text representation beyond the canon. The development of

diverse corpora, paired with rapidly developing methods out of fields such as natural language processing (NLP), has led to the development of Digital Humanities subfields such as computational literary studies (CLS).

Particularly in the last two decades, there has been an expansion of data mining to include other types of data, such as images and sound. Early work in DH anticipated these possibilities, but the technology for analyzing sound and images took longer to develop. Creating data for analysis remains a major area of work in the field, enabled by image and sound digitization technologies in the 1990s as well as the development of annotation tools, which became particularly prominent in the last decade (Arnold et al. 2021). While one should always be careful about not overstating the power of technological developments, recent advancements in storage, GPUs, and machine learning have supported the development of computer vision algorithms that have brought video data closer to parity with text in terms of ease, speed, and scale of analysis. Digitizing and using computer vision to analyze a film used to take days; now running several state-of-the-art models can take merely hours. With this shift have come new questions, scales of evidence, and areas of exploration such as distant reading and distant viewing. Getting access to data at the necessary scale has become a key issue.

TDM is made possible by digitization and the creation of born-digital materials. Commitments to digitization and open access, particularly by cultural institutions such as the Library of Congress, Rijkmuseum, and Smithsonian have opened collections of art, film, and music, in addition to modelling a commitment to open data for other members of the cultural heritage community. Initiatives such as Archives Unleashed are building on these efforts to make born-digital historical content accessible. Yet, their commitment to openness is not shared by other industries, even for research use. Capitalism and expansions to US copyright law have an outsized effect on access, particularly a 1998 law that has imposed significant restrictions.

The Sonny Bono Copyright Extension Act of 1998 ensured that no new works entered the public domain in the United States between 1998 and 2019, when a 95-year rolling window after the creation of a work began to move again. As of 2023, works created after January 1, 1928, are still protected by copyright under US law; works created after January 1, 1978, are protected for the life of the author plus 70 years. Consequently, works of the late 20th century may well not come into the public domain until the 22nd century. For fields such as film and media studies, the decision effectively makes the study of forms such as TV through computational methods off limits. Relying on the public domain as the primary way to source materials for computational analysis is simply not viable for humanists who wish to engage with the contemporary or the recent past. The effect of these laws on the trajectory of scholarship cannot be understated:

legal difficulties with doing computational analysis on contemporary and recent materials are steering cohort after cohort towards materials that can be more easily accessed—though they may largely be seen as less relevant by the public. This, in turn, has a negative impact on humanities programs, which are already put in the position of having to justify their existence in an era of budget cuts. Large corporations with an ownership stake in cultural heritage have the financial resources to pay lobbyists who can advocate for more and more restrictions on how that recent cultural heritage can be accessed. It is incumbent, then, on professional organizations to try to advocate for the interests of scholars and their research. The work of the Association for Computers and the Humanities in this space has demonstrated the impact of bringing humanities scholars' voices to the courtroom and the Copyright Office, where they are given more thoughtful consideration than in many corners of the university.

### **History of ACH advocacy**

Founded at the Modern Language Association (MLA) conference in 1978, the Association for Computers and the Humanities (ACH) has been committed to supporting and amplifying digital humanities work since before that name was used for the field. The organization's scope and activities have evolved over the years, ranging from holding regular conferences, to running a one-on-one mentoring program, to arranging newcomer dinners at the international DH conference, to offering microgrants for DH projects. One common thread over the decades has been advocacy on behalf of the DH community in legal cases that directly impact scholars' ability to do this kind of work.

Digital Humanities has been impacted more than many humanities subfields by changes in the legal landscape over the past 50 years, as both lawmakers and courts have made decisions about whether and how the “digital” differs from the analog. Sometimes these changes result from negotiations held by international bodies, such as the Digital Millennium Copyright Act of 1998, which codified into US law two 1996 treaties of the World Intellectual Property Organization (WIPO). This law—particularly when combined with the 1998 Sonny Bono Copyright Term Extension Act described above, and generally increased the copyright term to life of the author plus 70 years—has had a chilling effect on DH scholarship on 20th- and 21st-century materials, particularly work involving data analysis at scale, such as text analysis, which requires large corpora.

Copyright is not the only legal issue vexing Digital Humanities scholarship. The debates around net neutrality in the early 2010s—whether internet service providers could charge different rates for different kinds of traffic—were among the first to galvanize a response from ACH along with a broad network of digital (and less-digital)

humanities organizations. In 2014, 27 digitally-oriented professional organizations and journals wrote a letter to the chairperson of the FCC, advocating for net neutrality. This letter served as the basis for a similar response from the MLA, and in April 2015, the FCC voted in favour of a rule that defined ISPs as a Title II common carrier telecommunication service, which would maintain net neutrality.

The mid-2010s also saw ACH take action in relation to an important copyright case, the *Authors Guild v. Google* lawsuit. In this lawsuit, the largest professional association for writers sued Google for scanning authors' books without permission, as part of its Google Books project. Google settled with the Authors Guild in 2008, which allowed the book scanning project to continue in exchange for sharing ad revenue with Authors Guild members. This settlement was opposed by entities including the US Copyright Office and the US Department of Justice, out of concerns that it would lead to Google having a monopoly in making orphan works available. The settlement was ultimately rejected by the courts in 2011 after an attempted revision in 2009, leading to a new wave of lawsuits including *Authors Guild v. HathiTrust*. The lawsuit against HathiTrust focused on providing screen reader accessible copies of books to users with disabilities, offering full-text search (though the full text itself was only visible if in the public domain), and providing replacement copies of books that libraries had verifiably owned, if their copy was lost or stolen, and a replacement was unavailable for a fair price.

Matt Jockers, who served on the ACH executive council from 2008–2012, worked with Matthew Sag and Jason Schultz, two law professors and copyright experts, to submit the Brief of Digital Humanities and Law Scholars as Amici Curiae in *Authors Guild v. Google*, representing ACH and a group of 64 scholars from fields including law, computer science, linguistics, history, and literature (Jockers, Sag, and Shultz 2012b). This brief makes two major set of points: that “the freedom to make non-expressive use of copyrighted works is vital to the ‘progress of science’ in the Digital Humanities” and that “text mining creates value by facilitating the advancement of our collective knowledge; to protect that value, mass digitization and similar intermediate copying for data mining and other non-expressive purposes should be considered ‘fair use.’” Simultaneously, the same group laid out the same case for a broader audience through an article in *Nature*, entitled “Don’t Let Copyright Block Data Mining” (Jockers, Sag, and Schultz 2012a), which speaks clearly and directly to stakes of the case:

Among the issues at the heart of this dispute is what researchers in the emerging field of digital humanities will be allowed to analyse: only public-domain books (mostly those published before 1923 in the United States), or all known literary works. The answer may define the future of the field. (Jockers, Sag, and Schultz 2012a)

Ultimately, the rulings found Google Books and HathiTrust to both be within the realm of fair use, which was a significant win for the Digital Humanities community (Nowviskie 2012). At the same time, the major focus of the lawsuit was on the fair use status of search, rather than text and data mining, even as the brief gestured towards computational research as being an important area for the future of the field that would be foreclosed by a negative ruling in the Google Books and HathiTrust cases. ACH's next major advocacy effort would confront this issue directly, by supporting a petition for an exemption to the DMCA specifically for text and data mining.

### **Challenges of corpus-building for the computational humanist**

Before turning to the exemption petition itself, we will first describe the process scholars must undertake to acquire a corpus of digitized or born-digital material from the recent past for computational analysis. The complexity and expense of this process is core to the argument of the petition, that it is unreasonable to require scholars to work this way when faster, easier, less expensive, and more accurate options exist.

Depending on a scholar's interest and financial resources, the simplest path towards corpus acquisition is purchasing (or licensing) a commercially-available package of this data. For-profit companies have digitized in-copyright collections perceived to be of commercial value, offering page images, full text, and/or metadata, and making these available to libraries. These offerings are expensive, and sometimes come with license restrictions on use that can be negotiated away by a skilled licensing specialist, but not all libraries have staff with this expertise. Furthermore, relying on commercial solutions has the pernicious effect of skewing what can meaningfully be researched using computational methods towards materials that companies think have a commercial payoff. This inevitably disadvantages non-English materials, non-“literary” books (such as genre fiction or books for young readers), or anything that aligns with neither the canon nor trendy topics. While these services have recently turned their attention to postcolonial literature, writings by minoritized groups in the United States, and other materials that have historically been overlooked, there is little reason to think that this reflects a fundamental dedication to making underrepresented materials accessible, rather than a desire to capitalize on a rising wave of interest in topics like ethnic studies.

What options are left for the computationally-oriented digital humanist who is interested in studying materials from the recent past? Let's look specifically at text analysis. Scanning books and performing optical character recognition (OCR) is the safest approach—in essence, completely sidestepping all of the additional laws and entanglements that come with working with already-digital materials as a result of the DMCA. But book scanning is time-consuming, and therefore expensive. It becomes



unfeasibly so when scaling up to the very large corpora needed to answer some kinds of computational research questions. A much more viable path would be to take ebooks—which are either born digital or have been digitized (usually more accurately than when done by individual Digital Humanities projects)—and convert them to plain text files to build the necessary corpora. Even this is not without its distorting effects: not all books from the 1930s to the 2000s are equally likely to have been converted into ebooks for ongoing sales. For many projects, purchasing (sometimes rare and expensive) used books and scanning them is the only viable route. But the impact of commercial assessments of value is at least dampened compared to a situation where in-copyright materials are available only in purchased packages from vendors.

It is not difficult to find ebooks, freely available for download, on the internet. These books have usually had their technological protection measures (TPM), also known as digital rights management, stripped before being uploaded, leaving an ebook file that can be converted to the plain text used for computational DH research without breaking any laws. In contrast to European law, which has explicit language around how materials must be acquired, there is nothing directly prohibiting scholars from downloading and using these ebooks, assuming the scholar did not request that someone else acquire the books, break the TPM, and post it online. In such a situation, legal liability would fall to the person who violated the DMCA (by breaking the TPM) and copyright law (by distributing the book). The scholar may face liability under copyright law if they further distribute their corpus containing these in-copyright works, but there are not DMCA-related issues on the scholar's part since the files as they acquired them had no TPM, and the scholar's analysis of the materials, publication of limited excerpts of the materials, etc. should be covered under fair use.

Nonetheless, many scholars are risk-averse and reluctant to adopt practices that fall into a gray area of “not illegal” rather than clearly and positively “legal.” Prior to 2021, scanning and OCRing books, or running moving pictures through an analog capture process were the only options that fell into this category of clearly legal ways to acquire corpora for computational research, as a result of the restrictions imposed by the DMCA. However, one window of opportunity built into the DMCA is an exemption-seeking process, where every three years the US Copyright Office will hear petitions for why people need to be able to legally circumvent technological protection measures. Exemptions, if granted, must continue to go through a renewal process every three years—simpler than the initial petition, but nonetheless posing some degree of paperwork burden. In 2021, ACH joined groups of Digital Humanities researchers who use computational methods in providing a letter of support for an exemption petition initiated by the Samuelson Law, Technology & Public Policy Clinic on behalf of their client, Authors Alliance.



### Petitioning for a DMCA exemption

The Authors Alliance emerged in the wake of the Authors Guild lawsuits, as an alternative organization for authors who were interested in disseminating their works more broadly. Founded in Berkeley, California, in 2014, the Authors Guild has advocated for authors' rights from a different perspective than the Authors Guild, framing their mission as "promoting authorship for the public good by supporting authors who write to be read." "Being read," here, gestures towards barriers to reader access that often emerge on the publishing side, including books that go out of print while publishers continue to hold exclusive rights, preventing the author from taking steps to make their works more available. Part of their work has involved informing authors about relevant details of copyright law, including rights reversion, that are not widely understood. Another facet of their work has involved representing authors who *want* their works to be made available for research, including computational research. This group of authors includes many academics: a prolific group of often-overlooked authors, who may find themselves on both sides of the writer/researcher divide. Their engagement with academic authors has also led to their involvement in discussions around open access policy. Pamela Samuelson, a law professor at UC Berkeley and director of the Samuelson Law Clinic, was a founding member of the Authors Alliance, leading to the longstanding relationship between the two entities. Lawyers working at the Clinic, with support from UC Berkeley law students, have represented the Authors Alliance in multiple contexts, including developing a DMCA exemption request around text and data mining in 2021, which would allow people to circumvent technological protection measures for purposes of text and data mining research.

A 2020 NEH Advanced Institute workshop on the legal landscape for text and data mining laid the coalition-building groundwork for the exemption proposal. This workshop covered major legal and ethical considerations—including DMCA related complications—when doing text and data mining research. The audience was a mix of academics and librarians, and at the end of the workshop, the organizers put out a call for participants and their colleagues who might be interested in participating in an effort to put together a DMCA exemption for text and data mining in 2021. Among the volunteers was Quinn Dombrowski, then co-VP of ACH. Lauren Tilton became involved in her role as a member of the Executive Council of ACH. Together, we prepared a letter on behalf of ACH, which was signed and submitted by then-president Kathleen Fitzpatrick; in addition, we prepared letters on behalf of our own research groups, the Data-Sitters Club and Distant Viewing Lab.

Looking back on the exemption petition process, Associate Director of the Samuelson Law Clinic Erik Stallman noted that the letters and presentations from scholars as part

of the hearing process had a substantial impact on the outcome. Altogether there were 14 letters submitted, some from individual scholars, some from research groups and organizations such as ACH (ACH 2020). As part of our letter from ACH, we conducted an informal survey on Twitter to elicit perspectives from scholars that we wanted to make sure were represented, without placing the burden on them to write a full letter. We chose to include many direct quotes from the survey responses in this letter, to support our members in speaking, unmediated, to the Copyright Office. We described the impact of copyright as warping the trajectory of scholarship, but our members' responses made it clear that these issues were not just hypothetical. 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 the DMCA extended beyond individuals' research, and into the classroom. "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. We suggested that an exemption to the DMCA for text and data mining could more effectively put computational methods in the hands of students at a moment when machine learning and computational analysis are becoming a key research priority globally. 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." Respondents

also noted that the time-based limitations of copyright law create a filter that has more dimensions than time alone: “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.” The restrictions extend to media forms such as film and TV where sources such as DVDs are inaccessible due to DRM and therefore place off limits much of 20th-century time-based media. Overall, the restrictions have drastically limited the academic inquiry that animates humanities research and teaching.

The exemption proposed by the Samuelson Clinic on behalf of the Authors Alliance was expansive and was intended to cover the research and classroom scenarios described by the respondents to the ACH survey. Item C, section 1d (p. 10) specifically called out the adverse effects of the current law on teaching, and item E, section 3c notes that “The prohibition on the circumvention of technological measures applied to copyrighted works adversely impacts criticism, comment, news reporting, teaching, scholarship, and research”—including all of those activities within the scope of the desired exemption. There are numerous phases to the exemption process, including multiple response phases, a hearing, and rounds of private meetings between each side and the Copyright Office. Those opposed to the exemption included the Motion Picture Association, the Alliance for Recorded Music, the Entertainment Software Association, the DVD Copy Control Association, and the Advanced Access Content System Licensing Administrator, LLC, which we will call the “content industry.”

The opponents argued against the entire premise of the exemption, that text and data mining was not covered by fair use at all, and that there was no basis for an exemption as broad as what had been proposed. In consultation with the parties deeply involved with the exemption, including ACH, Erik Stallman from the Samuelson Clinic used the reply comment period to narrow the exemption to cover non-commercial research by academic institutions, libraries, and museums, focusing on the kinds of uses represented by the letters included in the petition—which primarily dealt with the impact on research by scholars at academic institutions. As the ACH representatives in this process, it was a hard concession to make, especially in light of how many scholars work beyond the academy. Nonetheless, all advocating for a change in policy recognized that any exemption would be difficult to win, and we could continue to work towards expanding even a narrow exemption.

### **Ruling and implications**

After months of deliberation, the ruling was released in September 2021—granting the exemption, in a limited way, that imposed a considerable security burden on

the researcher (Copyright Office 2021; Authors Alliance 2021). Per the final decision, researchers residing in institutions of higher education could bypass DRM for research under a series of conditions. This included staff and students if they are a part of a research team or as a part of teaching. The university must own the source of the data (i.e., not accessing it via a subscription service), and the researchers must take “effective security measures” to protect the data. As a result, researchers at institutions of higher education can conduct TDM with sources such as DVDs and ebooks, which is a significant development for the Digital Humanities.

Successful arguments by colleagues helped secure this exemption. For example, the content industry argued that even if one is allowed to use TDM, they should not be allowed to look at the actual material and should only be allowed to explore the computational results. David Bamman explained to the Copyright Office that people who executed algorithms on large corpora without being able to double-check their results would not be treated with credibility in the scholarly community, and the Copyright Office took this concern seriously. The letters from individuals, labs, and associations informed and bolstered the persuasive arguments of the Authors Alliance and Samuelson Clinic Team, and as a result, one can look at the data to verify findings.

The ruling did come with drawbacks. While parties agreed to exclude independent researchers from the exemption during the negotiation process, the final ruling also cut out libraries and archives that are not affiliated with a university. Scholars are permitted to circumvent TDM for purposes of conducting research on ebooks and digital films—but “only on copies of the copyrighted works that were lawfully acquired, and that the institution owns or for which it has a non-time-limited license.” This restricts both the use of scholars’ personal collections of materials (unless they were purchased with grant or university funds and can therefore be argued to actually belong to the university), and the use of materials that are part of an ongoing subscription plan but not purchased outright (e.g., anything on a streaming service). Furthermore, the ruling included a requirement that an institution “storing or hosting a corpus of copyrighted works ... implement either security measures that have been agreed upon by copyright owners and institutions of higher education, or, in the absence of such measures, those measures that the institution uses to keep its own highly confidential information secure.” We expect that, in practice, the security requirement will be a barrier. Scholars at research institutions—especially those with medical schools—are more likely to have access to data storage and compute infrastructure compatible with “highly confidential information.” Other types of institutions also handle highly confidential information for their internal operations but are less likely to have developed user-facing services for doing the same. We anticipate this to be an ongoing challenge for access to TDM in DH.

Another limitation on the impact of the ruling became evident working through the implications of the wording: while the ruling blunted the impact of legal liability on account of the DMCA, a DMCA exemption has no impact on a separate set of issues in the realm of contract law. Most people's path to acquiring ebooks runs through a small number of online stores run by large conglomerates such as Amazon and Google. People typically click through the thousands of words that constitute the terms of service for these stores without reading them, but almost all online ebook vendors include a clause prohibiting circumvention of TPM. To purchase things from the store, users must agree to these terms of service, and the DMCA exemption does not change the fact that circumventing TPM on ebooks purchased from the store would involve violating a contract that the user has agreed to. The copyright office has no jurisdiction over restricting this practice, and recent attempts at legislative remedy (e.g., in Maryland, where a law was passed requiring that ebooks be made available to libraries on reasonable terms) have been overturned as an unconstitutional interference with business practice (Albanese 2022). In Europe, there has been more success regarding legislative changes; at the EU level, there are now laws that prohibit contractual overrides to user rights otherwise encoded into copyright law, and these have made their way into the laws of individual member and ex-member states.

A significant amount of effort went into preparing this exemption request, both on the part of the lawyers from the Samuelson Clinic, and on the part of the scholars and ACH representatives who prepared letters and advised the lawyers as the exemption process progressed. The biggest winners are US scholars at well-resourced institutions. For scholars conducting research on moving pictures, this is particularly a significant win: this group can now rip and analyze DVDs, as long as they're able to meet the security burden. While the licensing situation muddies the waters for literary scholars, if their university library has carefully negotiated with vendors around their ebook purchases, they can likewise circumvent TPM and analyze the texts, if they have access to an environment for working with highly secure data.

While the exemption granted in 2021 is a significant development, we are also particularly sensitive to the fact that the people most in a position to take advantage of it are those already advantaged through affiliation with relatively well-resourced institutions. The exemption falls far short of what we had hoped for in terms of the number of people covered by the exemption, the range of activities, and the barriers imposed by the security requirements. It is not the full solution that our community needs. These limits are exactly why we need the community involved. The more projects that pursue this kind of work, and show the challenges that remain, the more likely the Copyright Office is to extend the exemption. It will be important to demonstrate the

burden of the security requirements from perspectives such as labour and cost, which are directly tied to access issues. It is also essential to involve scholars from more, and more diverse, institutional contexts, amplifying their stories so we can work to expand the exemption to the entire community, from research libraries to community colleges.

As part of her recommendation to approve the exemption, Register of Copyrights Shira Perlmutter included a discussion about whether computational analysis was fair use—a point the content industry had disputed. Her evaluation was that—at least within the parameters described in our exemption request—text and data mining would likely be considered fair use largely because it was both non-commercial and transformative. This alone is a positive development that lays the groundwork for future requests to improve the exemption by lowering security barriers to something more commensurate with the actual risk to the data, and by expanding access to additional users and contexts. In other words, we need DH scholars to show what is possible by taking advantage of this exemption, despite the current barriers.

Computational analysis encompasses one set of methods in Digital Humanities and should not be treated as a proxy for the field as a whole. Nonetheless, it is an area of work that has been well represented in DH research and at recent ACH conferences, and one entangled in a set of legal issues that other types of DH scholarship are able to navigate more easily. As DH also expands into critical code and algorithm studies, having access to text and data mining data to test, experiment, critique, and rebuild algorithms will also be key as we intervene in important debates about our computational algorithmic world. Without advocacy to address these legal issues, they will continue to play an outsized role in shaping the direction of individual research agendas and careers, and the field as a whole. We cannot afford to ignore this legal landscape and hope that it gets better on its own, even though the alternative is time-consuming, slow, and as likely to lead to frustration as relief. While the US is only one country, it has exerted pressure on other countries to adopt comparable laws around copyright. This has led, for example, to the 2012 C-11 bill in Canada with “digital locks” provisions comparable to DMCA §1201—but without the exemption petition process offered by the DMCA, leading to significant and difficult-to-resolve limitations on the potential scope of archivists’ work (Hinze and Sutton 2012; Macklai 2021).

For Digital Humanities scholars in the US whose computational research is impacted by copyright law, the DMCA, and license terms that include contractual overrides on fair use rights, we hope that this case offers some ray of hope that the Copyright Office is open to listening our requests and granting exemptions to address the needs of our community. Much as researchers are loath to take on additional administrative work, ongoing copyright advocacy is an unavoidable part of the future of Digital Humanities if



we want to remove the barriers facing our computational work. The more the Copyright Office hears from researchers and teachers from a wide range of institutional contexts, the more likely it is that they will continue to re-evaluate the restrictions that get in the way of our work.

Pushing back against national copyright laws and other restrictions from the DMCA and DMCA-like add-ons has to also happen on a country-by-country basis. Especially since many Digital Humanities projects involve collaboration across borders, we must not limit our vision to only our national context. There are opportunities to work together, comparing notes on effective advocacy strategies, and using one another's successes as positive examples to point to. Realistically, these efforts may not substantively improve the scholarship and teaching situation for those in a position to advocate for change today, at any point in their careers. But for a field to thrive, it depends on more than self-interest. We are deeply appreciative of the work of Authors Alliance and the Samuelson Clinic for continuing to work to expand this exemption, and we are excited to keep supporting their efforts. If you are in a position to support advocacy efforts, please get involved.

---

## Competing interests

The authors have no competing interests to declare.

## Contributions

### Authorial

Authorship in the byline is Quinn Dombrowski and Lauren Tilton. Author contributions, described using the NISO (National Information Standards Organization) CrediT taxonomy, are as follows:

### Author name and initials

Quinn Dombrowski (QD)

Lauren Tilton (LT)

Authors are listed in descending order by significance of contribution. The corresponding author is QD.

Conceptualization: QD, LT

Methodology: QD, LT

Resources: QD, LT

Writing – Original Draft: QD, LT

Writing – Review & Editing: QD, LT

### Editorial

#### Special Issue Editor

Emmanuel Château-Dutier, Université de Montréal, Canada

Barbara Bordalejo, University of Lethbridge, Canada

Roopika Risam, Dartmouth College, United States

#### Section, Copy and Layout Editor

A K M Iftekhar Khalid, The Journal Incubator, University of Lethbridge, Canada

#### Translation Editor

Davide Pafumi, The Journal Incubator, University of Lethbridge, Canada

#### Production Editor

Christa Avram, The Journal Incubator, University of Lethbridge, Canada

---

## References

ACH (Association for Computers and the Humanities). 2020. Letters of Support, December 3. Authors Alliance. Accessed October 30, 2023. <https://www.authorsalliance.org/wp-content/uploads/2020/12/letters-of-support.pdf>.

Albanese, Andrew. 2022. "Maryland Gives Up on Its Library E-book Law." *Publishers Weekly*, April 11. Accessed October 30, 2023. <https://www.publishersweekly.com/pw/by-topic/industry-news/libraries/article/89017-maryland-gives-up-on-its-library-e-book-law.html>.

- Arnold, Taylor, Stefania Scagliola, Lauren Tilton, and Jasmijn van Gorp. 2021. "Introduction: Special Issue of AudioVisual Data in DH." In *Digital Humanities Quarterly* 15(1). Accessed October 30, 2023. <http://www.digitalhumanities.org/dhq/vol/15/1/000541/000541.html>.
- Authors Alliance. 2021. "Update: Librarian of Congress Grants 1201 Exemption to Enable Text Data Mining Research." *Authors Alliance: Latest News*, October 27. Accessed October 30, 2023. <https://www.authorsalliance.org/2021/10/27/update-librarian-of-congress-grants-1201-exemption-to-enable-text-data-mining-research/>.
- Copyright Office. 2021. "Exemption to Prohibition on Circumvention of Copyright Protection Systems for Access Control Technologies." *Federal Register* 86 FR 59627. Docket No. 2020-11. 2021-23311: 59627-59641. Library of Congress. Accessed October 30, 2023. <https://www.federalregister.gov/documents/2021/10/28/2021-23311/exemption-to-prohibition-on-circumvention-of-copyright-protection-systems-for-access-control>.
- Graban, Tarez Samra, Paul Marty, Allen Romano, and Micah Vandegrift, eds. 2019. "Special Issue: Invisible Work in Digital Humanities." *Digital Humanities Quarterly* 13(2). Accessed October 30, 2023. <http://www.digitalhumanities.org/dhq/vol/13/2/index.html>.
- Hinze, Gwen, and Maira Sutton. 2012. "Canada's C-11 Bill and the Hazards of Digital Locks Provisions." *Electronic Frontier Foundation (EFF)*, February 10. Accessed October 30, 2023. <https://www.eff.org/deeplinks/2012/02/canadas-c-11-bill-and-hazards-digital-locks-provisions>.
- Hockey, Susan. 2004. "The History of Humanities Computing." In *A Companion to Digital Humanities* edited by Susan Schreibman, Ray Siemens, and John Unsworth. Accessed December 4, 2023. <https://doi.org/10.1002/9780470999875.ch1>.
- Jockers, Matthew L., Matthew Sag, and Jason Schultz. 2012a. "Don't Let Copyright Block Data Mining." *Nature* 490, October 3. 29-30. Accessed October 30, 2023. <http://www.nature.com/nature/journal/v490/n7418/full/490029a.html>.
- . 2012b. "Brief of Digital Humanities and Law Scholars as Amici Curiae in Authors Guild v. Google." SSRN, October 30, 2023. Accessed December 4, 2023. <http://doi.org/10.2139/ssrn.2102542>.
- Losh, Elizabeth, and Jacqueline Wernimont, eds. 2018. *Bodies of Information: Intersectional Feminism and the Digital Humanities*. University of Minnesota Press. Accessed December 4, 2023. <https://doi.org/10.5749/j.ctv9hj9r9>.
- Macklai, Sabrina. 2021. "Fair for Who? In Favour of Digital Lock Exceptions for Canadian Archives." IP OSGOODE Intellectual Property Law & Technology Program (blog), May 12. Accessed October 30, 2023. <https://www.iposgoode.ca/2021/05/fair-for-who-in-favour-of-digital-lock-exceptions-for-canadian-archives/>.
- McGrail, Anne, Angel David Nieves, and Siobhan Senier, eds. 2021. *People, Practice, Power: Digital Humanities Outside the Center*. University of Minnesota Press. Accessed December 4, 2023. <http://www.jstor.org/stable/10.5749/j.ctv2782dmw>.
- Nowvickie, Bethany. 2012. "ACH Advocacy News." ACH (Association for Computers and the Humanities) blog, October 11. Accessed October 30, 2023. <https://ach.org/blog/2012/10/11/ach-advocacy-news/>.

