The Chilling Effects of Obstacles to Accessing, Using, and Sharing In-copyright Data for Quantitative Research

Patricia Aufderheide, Brandon Butler, Kimberly Anastacio

Copyright's Obstacles to Quantitative Research

Bio: Patricia Aufderheide is University Professor in the School of Communication, American University. Brandon Butler is Director of Intellectual Property and Licensing, University of Virginia. Kimberly Anastacio is a PhD Candidate at American University

Abstract

An international survey of researchers doing text- and datamining research, followed by in-depth interviews, shows a range of obstacles. Researchers experience challenges in access, use, sharing of data, and storage. The sources of the problems are high prices for proprietary data, terms of use that inhibit research, and legal policies including copyright, privacy and anti-hacking. Consequences of facing this range of obstacles include changing research design, delaying research, abandoning research, and failure to collaborate across jurisdictional borders. The right to conduct research should be asserted in designing relevant legal policies.

Keywords: copyright, fair use, quantitative research, datamining

Introduction

Text and data mining (TDM) is a basic feature of daily digital life. It enables search (of the internet and other digital resources); it powers targeted advertising; it feeds predictive policing; and increasingly for scholars it is a crucial tool to track networked behaviors and identify patterns relevant to their subject disciplines.[1] Those disciplines are as wide-ranging as medicine, political science, engineering, law, computer science, linguistics, history, and communication.

Data mining, as studies from the Organization for Economic Co-operation and Development (OECD) have identified, is growing both as a driver of economic growth and as a research activity.[2] A *sine qua non* for data mining is access to data. In some cases, researchers create their own databases from publicly accessible raw material, for example by scanning physical texts, assembling personalized databases out of larger ones (e.g., the Internet Archive), or crawling publicly accessible websites. In other cases, the data has already been gathered and formed into a database by someone else.

Once the data is gathered and the analysis is conducted, the right to use and share the results of data analysis is crucial to generating new knowledge. Sharing data with other researchers is also important, as it supports reproducibility and transparency in research and spares others' duplicative effort in creating databases for related projects.

The advent of "big data" research has generated a wave of scholarly literature, including sunny optimism about the immense opportunities, cautionary notes about the many ethical questions involved, and even theoretical exploration of what "Big Data" is, exactly.[3]

Meanwhile, such research has yielded astonishing results. In science and medicine, unsuspected correlations have led to life-changing and life-saving innovations. Better treatment for Raynaud's disease and better diagnosis of Huntington's disease are only two of the achievements.[4] Digital humanities projects have burgeoned, too, thanks to new ways to analyze cultural phenomena. For instance, the Kinomatics Project (https://kinomatics.com) uses both big data and mapping software to chart global flows of cultural products, allowing a robust analysis of their cultural, economic and political implications. Early work with a corpus of over 5 million digitized books (created as part of the Google Books digitization project) demonstrated the rapid growth of the English lexicon (the number of unique words appearing in the literature), trends in English grammar (the regularization of verbs like “dwelt” and “knelt”), a growing cultural obsession with the present year, an accelerating cycle of fame and obscurity, and the censorship of German books under the Nazi regime.[5] Data mining techniques can also provide foundational evidence in the conversation about diversity in literature, for example that 95% of works published by major firms between 1950 and 2018 were written by white authors, and that there was little improvement over that period, as white authors represented 89% of the output of the major publishers in 2018.[6]

Unlike traditional research methods, text and data mining research faces potential obstacles at every stage based on policies and practices designed to limit access to and sharing of in-copyright and other proprietary data. Our research indicates that real and perceived legal barriers play a particularly significant role in limiting text and data mining research.

Research on TDM practices

The advent of affordances for text and data mining has resulted in a flourishing of new research, and analysis of new questions raised by this research, such as preservation and interoperability, along with meta-analyses of the field itself.[7]

Some research has identified obstacles, including training and experience in tools and skills with which to do research, maintenance of sites, lack of appropriate recognition for work in the field, and funding. Copyright problems of researchers, primarily of access, have been identified among these obstacles. For instance, a study based on interviews of 75 people involved in production of the digital humanities found:

Those working with post-1924 United States materials reported more problems with copyright. The new affordances of text-mining complicate questions of access, as many academic distributors only permit browsing access to copyrighted materials. Others charge fees for large-scale, computational use cases (e.g., text-mining). In the United States, analysis of copyrighted materials for scholarly purposes qualifies as fair use, but the re-distribution of copyrighted materials as part of a dataset does not. Some vendor licensing agreements bar any redistribution of data, even if copyright is not a factor. Such complexities have led to creative solutions like the HathiTrust Research Center's (HTRC) original "walled garden," which enabled pre-defined text analysis algorithms to run on copyrighted materials. HTRC has also provided "non-consumptive" versions of texts in the form of term frequency tables.[8]

A recent study of licensing challenges for TDM research found that, "TDM licensing still seems like a Wild West, where too many vendors are taking too many approaches, causing too many librarians to have to figure out far too many different licensing options."[9] Rachael Samberg and Cody Hennessy have developed a "legal literacies" approach to guiding scholars through the challenges involved in TDM research in the US, including copyright but also contract law issues, as well as ethical considerations.[10] With funding from the National Endowment for the Humanities, a larger group of scholars led a weeklong institute to train a cohort of scholars and librarians to use this literacies approach.[11]

Copyright and TDM

Researchers have charted an uncertain path as they navigate the relationship between copyright and TDM. While a body of international agreements creates broad commonalities regarding basic elements such as duration and subject matter, copyright law can vary substantially from country to country in many respects, including its treatment of databases and limitations favoring research. In most countries, databases are subject to at least some level of copyright protection, while in some countries (including members of the European Union), databases receive additional protection against copying.[12] Some kinds of works within databases may be copyrighted, and others may not be.

In some countries, including the US, researchers can (if they understand the law) employ fair use to create databases of all kinds for TDM research without seeking permission, assuming they can lawfully access the data. It is widely accepted in the US that fair use permits the copying requisite to enable TDM, something that effectively provides the US a competitive advantage over nations without this exception.[13] Decisions in two lawsuits brought by the Authors Guild targeting book digitization by Google and partner academic libraries firmly established the application of fair use in this context.[14]

In other jurisdictions, including the EU, researchers face a more difficult path, as permission-seeking is the norm and recently adopted exceptions are unduly limiting.[15] In many places, including the US, data owners can use contracts to limit access and use even when copyright would permit it. In others, e.g., Brazil, user rights under law cannot always be revoked by private agreements.[16] International copyright agreements like the Berne Convention and the World Intellectual Property Organization (WIPO) Copyright Treaty do not addreses TDM directly, leaving countries to design their own approaches (or not) subject only to the broad requirements of the treaty system. Among these is the so-called "three-step test," which is sometimes advanced as a barrier to TDM-enabling policies.[17] At least one scholar has argued that in fact the international copyright regime has no application to TDM practices, leaving signatory countries free to design approaches to TDM unconstrained by the generally protectionist approach of the copyright treaties.[18]

As the amount of data potentially available grows exponentially, and as more and more tools become available to analyze bodies of text and data, uncertainty about how to apply exceptions, and indeed whether exceptions are even needed, has a cost to productivity and innovation.[19] Christian Handke, Lucie Guibault, and Joan-Josep Vallbé have documented (2021) that production of scholarly research is correlated with copyright terms and copyright enforcement.[20] They found that in countries with a requirement to get permission from the owner of the database to do TDM and a strong "rule of law" in copyright (meaning vigorous enforcement), less work employing TDM is published.

Some common problems are indirectly related to copyright. Libraries may have contracts with vendors, enforceable by the draconian penalties of copyright law as well as by contract, that limit their ability to provide access to entire databases to researchers.[21] Commercially hosted data such as social media content may be publicly accessible but constraints on use (e.g., bars or limits on downloading data or sharing results) may make TDM research seem risky or impossible. While contracts, anti-hacking laws, and other measures play an important role in imposing these limits, copyright is a powerful tool for legal control of publicly available proprietary data.[22] Owners of in-copyright data may zealously enforce their rights, or, perhaps worse, they may be unreachable, creating a dilemma about how to deal with apparent orphan works. The price of access to commercial databases may be too high. Researchers working across national borders may encounter a clash of copyright regimes.

Copyright obstacles to research exist not only in the law but in people's understanding of the law.[23] Just as the process of judicial interpretation of equitable doctrines like fair use depends intuitively on the context of use and expectations in the field of practice under scrutiny, so does the individual's ability to invoke legal rights depend on their own understanding of what is normal and what is possible under the law.

Methods

We developed a survey using the Qualtrics platform, which permits full anonymization and secure storage. We received a waiver from the American University Institutional Review Board and fielded the survey via our own social media networks and professional organizations. Professional groups included the Association of Internet Researchers, the International Communication Association, and the Right to Research network at the Program on Information Justice and Intellectual Property at American University's Washington College of Law. We consulted with colleagues in the Right to Research Network to develop Arabic, Portuguese, Spanish and French versions. We received 262 responses between March 2021 and November 2022. Respondents were free to choose which survey questions they wanted to answer. For each question, they could also choose any number of answers from a set of multiple choices. Thus, the results and percentages reported in this paper correspond to the total of answers in each question of the survey, and not the total respondents that answered a particular question.

We then followed up with survey respondents who volunteered to talk with us and did open-ended interviews with ten people, all either librarians, technologists assisting researchers, or researchers. We asked them simply to tell us about their frustrations with doing or supporting TDM research, and the workarounds they used.

Results

Our respondents came from diverse regions of the globe, have a range of disciplines and a range of experience. When asked in which countries the respondents primarily do research, we obtained 185 answers that mentioned a total of 40 countries. The United States (US) represents a third of the countries cited (33%, cited 71 times). Countries in the European continent totaled 43% (cited 80 times, and the United Kingdom alone was cited 13 times, as well as Germany, each corresponding to 6% of the total). Another 6% (cited 13 times) were from Canada, 7% from Brazil (cited 16 times), and 4% (cited 9 times) from India. Thus, the great majority were from the global North. The remaining 14% were sprinkled throughout the globe.

When asked in what kinds of institutional contexts they work, 220 respondents cited a total of 251 contexts. Coming in first was the academic environment (78%), cited 195 times). 9% (22 times) cited nonprofits, 6% (15 times) government and around 6% (16 times) cited private corporations. Between those who answered in what particular areas they research (220 respondents), digital humanities was the most cited field (14%, cited 89 times). Next, it was Communication (14%, 87 times) and Internet Studies (11%, 72 times). Information and library science corresponded to 9% (cited 59 times), followed by computer science (8%, 51 times) and linguistics (7%, 45 times). Other answers were dispersed among different fields, primarily in the humanities or social sciences, such as Literature, History, Sociology, and Education. Among 184 respondents, almost half (43%) had been working in this area 5-14 years, and 29% had been working less than five years. A little more than a quarter (28%) were veterans working 15 or more years in the area.

Users get their data in a range of ways, sometimes more than one way. 176 respondents cited 553 sources of data. A fifth of the total responses (20%) reported using off-the-shelf software to collect digital information. Another fifth (18%) collected information digitally using purpose-built software. 14% of the answers correspond to access to data for free from a third-party provider. 8% of the answers said they use a library-licensed database. 7% reported using a library/archive-owned database, and another 7% of the answers correspond to digitizing analog information using off-the-shelf software. 6% paid to access a vendor's database (not obtained through their library), and 6% relied on internal corporate databases. 5% use purpose-built software, and 4% reported relying on a database the library specially paid for to support their TDM research. 3% reported special arrangements with the library to access their data.

Access to data – Vendors

We asked respondents first about their relationship with vendors. What problems did TDM researchers encounter with vendors? As in other answers of the survey, respondents could answer as many options as they found relevant. 85 respondents reported 276 problems at every level: initial contact, access, use, sharing, and archiving. Out-of-reach pricing was most common; almost a fifth of the answers (18%, 50 times) reported this as an obstacle to access. Almost as common (13%, 36 times) was that the terms of the license were confusing or that there was deliberate vendor interference, using technical means, to impede access to data (13%, 36 times). Respondents also reported delays in getting a response from a vendor (13%, 35 times). In 10% of the answers (28 times), respondents said the vendor impeded their appropriate use of the data; almost the same amount reported vendor impeding sharing of data (9%, 26 times); and 9% of the answers reported not being permitted to archive the data (25 times). In 10% of the answers (27 times), the relevant entity did not even answer their query. 4% of the answers reported no challenge when dealing vendors (mentioned 11 times).

Interviewees gave interesting examples and more detailed accounts of the kinds of problems they encounter in working with vendors. In one interview, we heard about a snarl of issues associated with JSTOR, a major scholarly journal platform, including a rolling embargo that blocked TDM access for material less than 6 years old (a term the interviewee assumed was required by the platform's publisher partners). The upshot was that alternative sources had to be found to plug the gap in the researcher's corpus to ensure their claims were up-to-date, and inconsistent data formats had to be reconciled to create a coherent body of material for study.

Indeed, several interviewees cited inconsistent data formats across different publishers and vendors, as well as changes in data format over time, as stumbling blocks for TDM. Relatedly, researchers told us they would benefit from re-processing collections as technology improves (feeding original scans through successively improved optical character recognition programs, for example), but vendors aren't necessarily motivated to do this work or to make originals available openly to scholars so they can do the work themselves. Zoe LeBlanc, Assistant Professor at University of Illinois Urbana-Champaign (UIUC), told us that, "Out-of-copyright materials are so much richer because they don't rely on providers, scholars can act independently to improve quality." LeBlanc explained how an Arabic language newspaper headline is 100% wrong in the Optical Character Recognition (OCR) used by the vendor, but Google Vision OCR gets all but 2 words right. All this inconsistency and unreliability among data providers creates additional work for researchers who must make the data commensurable to study it in a unified TDM corpus. In most cases, copyright gives the vendor leverage to exercise dominion over these technological uses.

Interviewees described wide disparities in vendors' familiarity and comfort with TDM, with some vendors simply unable to handle requests for TDM-ready data, while others charge a hefty fee for a hard drive full of raw data that requires substantial processing to make it usable.

Vendor efforts to keep data within proprietary silos were a consistent theme in our discussions. Interviewees expressed concerns that for some vendors, the profit motive is at cross purposes with research mission: where researchers need data to be accessible and compatible to facilitate their work, some vendors prefer to control access and impose proprietary formatting and other limits to maximize their opportunities for profit. Good models were cited, as well, including the Text Creation Partnership, a collaboration between libraries and vendors to facilitate digitization of early print books, where a limited exclusive window lets the vendor profit reasonably but afterward the public gets full open access.[24]

Interviews revealed another dimension of the data access problem: identifying a corpus of data to be studied sometimes requires access to specialized data sets such as bestseller lists, which provide a kind of external criteria for inclusion in a particular TDM corpus. Within certain bounds (texts in English, the New York *Times* lists up to a certain year, and so on) these data may be easily accessible, but beyond them the price and difficulty of access can escalate quickly. One interviewee reported a quote for access to a non-English best-seller list that was higher than the price for access to the entire corpus of the *Times*.

Access matters, and corporate researchers are aware of that. Among corporate researchers, everyone unsurprisingly agreed that they have better access to their data than outsiders. Out of 48 answers from 31 respondents, 38% (18 times) agreed they have better access to internal corporate data than outsiders, and 15% (7 times) that they have better tools to work with such data. In 23% (11 times) of the answers, the respondents said the access to corporate data gives them better insight into their research questions than publicly available data. In 13% of the answers (6 times), respondents said that their company makes its data available to outsiders on the same terms, and to the same extent, as insiders.

Access to data - Libraries

Next, we asked respondents about problems they encountered with libraries. 64 respondents reported 219 problems. The most common problems reported had to do with resources licensed to the library by vendors, and echoed the problems reported when dealing directly with vendors. In other words, the libraries' barriers to use were most likely created on behalf of the vendors. For instance, the most reported problem with library access (11%, cited 25 times) was vendor pricing, greater than the library could afford. The terms of vendor licenses to libraries caused several problems. Those terms were confusing (11%, 25 times), banned data sharing (11%, 25 times), technically impeded access (10%, 23 times), limited or banned TDM (9%, 20 times), impeded appropriate use (9%, 20 times), or banned archiving (8%, 16 times). In 5% of the answers (mentioned 10 times), the vendor simply didn't reply to the library's inquiry regarding access for TDM research.

Researchers also experienced some obstacles as being caused by the library directly. Some of the answers reported that librarians did not return to vendors who banned TDM to renegotiate access

(6%, 13 times) and a similar percentage of answers reported library delays (7%, 15 times). In 5% of the answers (11 times), researchers believed the library had imposed restrictions that went beyond the law or vendors' terms. 7% of the answers reported no challenge when dealing with libraries and archives (mentioned 15 times).

In interviews, we heard more detailed stories about library- or university-created barriers to TDM. One scholar lamented that repositories of open access research in India do not have the technological capacity to support TDM, despite being openly licensed and thus theoretically free to re-use. Another interviewee bemoaned limits on access and use of materials in the HathiTrust Digital Library (HTDL), characterizing it as the "Yucca Mountain of digital libraries, where toxic materials are buried and not allowed to see the light of day." Several characterized the "bag of words"-style data that is currently the only form of data that can be released from the HTDL. The interviewee found this method insufficient to support the most powerful tools for independent computational analysis (i.e., analysis outside of the research "capsule" controlled by HathiTrust). Users ascribed these limitations in the HathiTrust's functionality to the library's fear of copyright liability.

At the same time, interviewees expressed hope that library projects, including HathiTrust, are helping to solve some of the problems endemic to TDM research. For example, Glenn Layne-Worthey explained that the HTDL "is just like libraries generally [in that] it solves the problem of prohibitively expensive collecting for res earch." The problem of "start-up cost" associated with finding, acquiring, and preparing data for analysis was cited by virtually all of our interviewees as a barrier to TDM work, and libraries are uniquely positioned to help solve it, especially through collective projects like HathiTrust.

Policy and obstacles

From a policy perspective, where do researchers lay blame for the problems they encounter? They widely believe that the trouble starts with copyright law and is then filtered through institutions. 143 respondents chose 304 answers in assessing blame. One fourth of the answers (24%, cited 73 times) attributed the problem to copyright law. A similar amount (21%, 65 times) blamed vendor contracts, e.g., Terms of Service or End User License Agreements with users. Next, answers said legal issues that affect TDM research are also connected to the library's contract with the vendor (19%, 58 times). One respondent wrote, "I teach web scraping to students as well, and the legal standing of terms of service of various sorts tend to be as strong an impediment to work as copyright law per se, especially in cases where redistribution isn't necessary." 19% of the answers (57 times) lay the problem in privacy laws. 7% of answers (cited 21 times) found anti-hacking (anti-circumvention) laws a problem, and almost the same amount reported no legal issues (8%, 23 times). One person noted that ethical considerations, no matter what the law says, were an obstacle for them.

Zeroing in on copyright law specifically, out of 134 respondents, the majority indicated problems (67%). Those who did typically had problems with copyright law throughout the research cycle. The biggest single problem was with sharing of data (17%, cited 33 times), but there was plenty of friction in other areas. Some also said copyright law was unclear (17%, 28 times), and a similar number (14%, 27 times) said it impedes collection of or access to data (26 times). 11% percent said copyright law impedes data use (22 times), and 7% said it impedes archiving of data (13 times).

Among specific problems mentioned, researchers cited (in their exact words):

- European General Data Protection Regulation (GDPR) makes collection of personal data complicated;
- Missing fair-use principle in the United Kingdom;
- European GDPR makes it impossible to publish non-anonymized data;
- The law for the library we work with (yes, there is a law for this library), demands that the data is used on their physical premises, so we have to travel there to do our research (no virtual private network etc. allowed);
- Digital Rights Management (DRM) on ebooks makes it difficult to gather data, and buying print books and digitizing them is expensive and not time-effective;
- Many of my historical sources are under copyright and therefore few of these materials have been digitized or if they have been they are not available for TDM;

- Storage security requirements for new Digital Millennium Copyright Act (DMCA) exemption makes it very hard to use.

Researchers also cited:

- Inability to share raw data (which limits reproducibility);
- Out-dated and pre-digital laws;
- DRM on various media;
- Problems with sharing data with researchers in other institutions;
- Problems with determining copyright of texts;
- Orphan works;
- Fear of fair use ("determined post hoc");
- Risk-averse cultural institutions;
- Requirements to store data on a university server with unclear capacities to move the data should the researcher switch institutions.

Cross-border research

What about researchers working across legal jurisdictions? Out of 132 respondents and 153 answers, most answers (52%, 79 times) state that researchers had no cross-border experience. Among those who did, there were a variety of issues. In 18% of the answers (cited 28 times), researchers said that they simply ignored the differences in the law. For a similar group (15%, 23 times), the differences in law affected their project. One noted in a comment, "I have been unable to get around the GDPR and had to drop a European partner." For 10% of the total answers (15 times), lack of backstopping legal expertise affected the project. What happens then, reported one respondent, "is that, most of the time, we decide to act cautiously and drop any risky data, even though we may have been allowed to use it according to official laws." Fear evidently was a factor. As one respondent wrote, "Pretty much hoping not to get sued!" Fear could inhibit action. Another respondent wrote, "I try to avoid international collaboration so as not to worry about this."

But so was the hope to fly under the radar. One wrote, "We ignored them [conflicting regime issues]. It's common." Another wrote, "Ignorance is bliss!" One respondent believed this was quite common: “Honestly, people are breaking copyright left and right and just not saying anything about it. We have to be realistic about that, on top of all the other complications. It is easy to OCR [optical character recognition] books, just time consuming. The bigger concern to me is what happens to the files afterwards.”

Chilling effects

What are the effects of facing this range of obstacles? Out of 140 respondents and 155 answers, most of the answers said either that these problems had in some way impaired their research (43% total), or that they were unsure whether it had (17%). The most commonly reported problem was having to change the design of the research (23%). 14% percent of the answers reported avoiding taking on a project. 6% percent of the answers mentioned having to abandon a project. As one interviewee said, “We have these fun conversations and then one of us asks, ‘OK, what do you have on disk?’ And 85% of the time the conversation just ends there, because it’s so hard to get things that are in copyright.”

According to one respondent, they have “No more interest in doing research on social media data (e.g., messages from twitter, etc.).” On the other hand, one respondent mentioned that “The tool I have to collect real-time data (Twitter/TAGS) is much easier to use than the tools that collect historical data (paid Application Programming Interface (API) access, command line, etc.). This has encouraged me not to ask questions that require historical data.”

Individually, people reported various choices that result in suboptimal research design. They tailor their research to what they can find: “Honestly, I just look for platforms (Constellate, Twitter, HTRC that allow for TDM before even approaching a project like this. Don't want to deal with clearance/permission/negotiation so try to scope research based on what's available,” one said. The choices they must make for access shrink the scope of their project: “Instead of the TDM I had planned to do with a larger run of issues digitized by the vendor, I had to use a smaller range of issues that were ones the library originally contributed to the project.” Also, “instead of using the full content that was collected, I had to choose a very small part.”

Even when they have the data, other constraints arise. For example, one respondent mentioned that they “had the data, but couldn't use it in an official setting due to unclear laws”. Delays limit what can be done as well. For instance, one respondent mentioned that the problem they faced “slows down the research on news data analytics, restricts data sources” One also told that, “Once, I needed access to a dataset and had to sign a fairly simple contract. That was a year ago, and my university's legal department still hasn't gotten back to me yet.” Even involving experts on copyright law, for some, did not help: “Getting access to data is the thing that I spend way more time on than I wish for. It's a time sink that's often not built properly into research projects which ultimately causes delays. Involving lawyers usually makes this process even longer.”

People are skewing research questions entirely away from copyrighted materials. Comments included: “I no longer work on materials where copyright could be an issue” and “I pretty much shifted my whole area of research to avoid worrying about post-1922 issues.”[25]One respondent was clear and direct, “I have stopped research on projects where copyright is confusing or otherwise impedes sharing of data.”

In doing so, they are aware of what is lost: “Students would be so much more engaged if we could use more contemporary literature,” one survey respondent wrote. An interviewee put it more bluntly: “It’s the most interesting stuff that’s most radioactive, copyright-wise. It’s really hard to get undergrads interested in eighteenth century novels.” A third was even more direct: “There are all these complaints about literary scholarship not being relevant, et cetera, and no shit, because we can’t work directly with the cultural production of the twentieth and twenty-first century.” Such issues, as noted by one of the respondents, directly affect the condition of their research: “(I) had to use older data of lower quality because newer material is copyrighted. This decreases the quality of research.” Thus, researchers expressed fear of legal troubles: “It's not worth my efforts to do the research when doing it might make me subject to legal action.”

Interviewees shared similar stories and sentiments. “Anything in-copyright is a problem,” and “Older is always better!” were representative of the general attitude toward TDM with in-copyright materials. Graduate researcher Cristiano Therrien referred us to his dissertation, in which he describes how he curtailed his use of text and data mining methods due to legal uncertainty: “Even considering copyright exceptions for scientific research, the text mining on protected PDF files leaves too many grey areas…, so the more intensive and extensive forms of data analytics with machine learning were soon set aside.”[26] Another interviewee described the impact of copyright on the availability of text for TDM research as “The Calvin Coolidge apocalypse - it’s as if an asteroid struck the earth in 1926 and wiped out all civilization for 75 years.”

Several interviewees described projects that were artificially curtailed in deference to perceived copyright barriers. In one example, a scholar interested in using computer analysis to see trends in the use of certain kinds of metaphors in scientific writing was told by our interviewee (a digital scholarship specialist) that they should confine their research to pre-1923 texts to avoid copyright headaches. Another described realizing that the promise of TDM for answering long-running questions about a particular genre of fiction—how it has evolved, whether it has devolved, what is it “really” about, and so on—would be consistently frustrated by the limitations that copyright placed on creating corpora that fully represented twentieth century texts.

Scholars of recent history and culture worried they would be consistently frustrated by copyright in their efforts to use TDM and similar techniques. One interviewee told us, “I’m worried that for twentieth century historians, so much remains under copyright it’s just hard to see how we will grow.”

Policy knowledge

Researchers work in a world where, most of the time, they are not getting much professional help in making their decision, even though they theoretically have access to it. 128 respondents gave 236 answers for where they get helpful information in making their TDM decisions. 29% of the answers (68 times) were “colleagues/peers” and 18% (43 times) “myself.” 13% (31 times) selected “librarian,” 11% (26 times) “lawyer,” and also 11% (26 times) “superior/boss.” Friends was the least cited source (7%, 17 times). Among the 11% answers (25 times) that selected “other,” respondents mentioned their department staff, university law experts, and the Internet through online communities.

One noted, “I have not been able to find anyone ever willing to give me advice on legal issues, for fear of being legally responsible if I do something wrong.” Another respondent offered a reminder that independent researchers are particularly vulnerable: “Not every researcher is embedded in an institutional context or affiliation, there is greater lack of clarity on how to proceed.” Similarly, one interviewee noted the likely disparate impact of legal uncertainty depending on a scholar’s institutional home, their seniority, their job security, and other factors: “People like me at places like [my institution] can do this, will do this, have the resources and lack of bureaucracy looking over their shoulders, etc. to do this. Others will not.”

Yet another noted the constellation of expertise needed to offer constructive support: “It is difficult to find someone who is an expert in the area of computer science / natural language processing AND copyright / data privacy laws. People are usually only proficient in one of those, and it is hard to bridge the gap. It would be great to have an official committee to review research proposals / ideas and to evaluate if it is possible to do research with the data (for the country/countries in question).”

This pattern of lack of professional support and dependence on one’s own judgment is similar when respondents were asked who raises legal concerns that could impede the research. Here, among 124 respondents and 213 answers, the most common answer was “myself” (29% of choices, mentioned 61 times) and “colleagues/peers” (28%, 59 times). Librarians were chosen 11% of the time (24 answers). A superior was appointed in 13% of the answers (27 times), and lawyers came in at 8% (17 times). Considering the 7% of answers that chose “other,” respondents mentioned ethics boards and university staff among other things.

Thus, researchers depend on the knowledge of peers, and their own experience, more than any legal or library expert. At the same time, many of them are remarkably confident about their legal understanding. Out of 131 respondents, almost half (46%) are “somewhat” or “very” confident of their understanding of copyright law. 37% are “somewhat” or “very” unsure, leaving 17% in the middle. We see the same split when we ask about the legal knowledge of their team or organization. Out of 132 respondents, 45% are somewhat or very confident, 36% are somewhat or very unsure, and 19% are in the middle.

Does confidence in one's knowledge lower the level of abandoning and avoiding a project? Not necessarily. Of those who were somewhat or very confident, 49%, reported experiencing a problem, as opposed to 43% of the total group. Fewer of them than the general population were unsure whether they had experienced a problem or not. The total proportion of people experiencing problems was still quite large, and the patterns were the same, overall; the most common problem was making design changes, then avoidance, and finally abandoning. Interestingly, one interviewee reported a similar effect in their experience teaching digital research methods: an essay they assigned with the goal of reassuring students that fair use would generally permit them to conduct TDM seemed to have the opposite effect, replacing their blissful ignorance of copyright with questions about where the boundaries lie.

Cross-jurisdictional differences

We compared US-based researchers with Europe-based researchers, to see if there are differences correlating with copyright policy in different jurisdictions. In the US, fair use broadly enables this research, and fair use knowledge is widely (if often inaccurately) distributed. In Europe, different nations have remarkable diversity in copyright exemptions, and they can be challenging to interpret.

In some ways, the two populations were similar. Of a total of 52 answers, most of those working in European settings stated they faced problems due to copyright law (63%), while 37% reported no problems. Considering a total of 68 answers for those working in the United States context, the numbers are similar: 69% reported problems and 31% did not. Those working in a European setting, presumably because of the GDPR, mentioned they faced privacy-law problems twice as often as those working with US-based research. Besides, both were very likely (80% for US, 82% for European) to prefer open-access or public domain materials, simply for access.

Researchers working in a European context had similar levels of experience with cross-border research to those working in an American context (47% of European respondents had worked across borders, 45% for US). By contrast, researchers working in Europe were overall less sure about their copyright knowledge than those in America. 35% were somewhat or very confident, while in the United States (US) that percentage is 52%. This is plausible given the thicket of jurisdictions and limited exceptions they encounter.

In terms of chilling effects, those working in an American context were more than twice as likely as Europeans to avoid potential trouble altogether by deciding against doing research (23% for US, 9% for Europe). Researchers reported a similar experience of having to change their research design to accommodate the obstacles they encountered (24% for Europe, 21% for US).

Proposed solutions

When asked for good practices to replicate, suggestions ranged from the individual to the global, and from rule-of-law to subversion. Several suggested de-identifying and sharing datasets open-access whenever possible. Another suggested sharing source code, for peer review, and increased collaboration between data owners. Among the organizations and tools praised for good practices/good guidance are HathiTrust/HTRC, JSTOR, JSTOR Data for Research, Internet Archive, Constellate, Twitter Academic Access API, The Text Creation Partnership, the Association of Internet Researchers (which has applicable ethics standards), the British Library, and the National Library of Scotland. One person suggested standardizing good practices and developing a common vocabulary for TDM, another mentioned there should be good documentation of the data. Several suggested libraries could develop pro-TDM practices, including checking contracts for limits, and putting a protection clause into contracts, e.g. “Nothing herein should be understood to abrogate or deny fair use rights." One noted that others’ copyright violations could be leveraged; “using pre-liberated ebooks from some large-scale Russian online collections has made for a most excellent experience!”. Suggestions also included collaborative, user-generated databases, especially concerning social media data.

Discussion

Of the obstacles singled out, we note that copyright remains not only the largest single obstacle reported but is also implicated in other problems. Challenges researchers face with vendor terms, whether with the vendor directly or through their institution, are ultimately problems of terms of service typically enforced by copyright. Anti-circumvention/anti-hacking laws have protection of copyright as their key justification. Hesitations or refusal to engage cross-jurisdictionally are typically because researchers cannot find support for a copyright analysis that can assuage their concern for risk. Indeed, most of the legal risk that researchers flag is associated with copyright.

These results confirm a pattern we have seen in other areas,[27] of copyright creating chilling effects on creative research design. This is occurring at a time of rapid development, particularly in the corporate and military spheres, of text- and data mining. The public is left without the skill development, the research results, and the evolution of knowledge that the corporate sector exploits.

The losses to public knowledge surface only occasionally and murkily, since the costs are to work that could not be done, and in some cases, even imagined. The majority of the survey respondents (42%) said they encountered obstacles to their TDM research reported changing their project, avoiding, or abandoning it as a result. Although none of the survey respondents volunteered description of a specific project, one noted: "I would like to do a really wide scale analysis of news sources over time, but data availability is a huge problem." Interviewees were more forthcoming in describing abandoned and curtailed projects, with maddening results: artificial copyright cut-offs just as data are getting interesting, frustrating busywork to route around vendor copyright silos, and rules of thumb that place most of the twentieth century off-limits. Some worried that entire fields of study—history, literature, media studies—were being negatively impacted as students and administrators took their failure to engage with recent history as a sign of their growing irrelevance. Beyond the projects imagined but set aside are the projects that have not even been imagined; they are beyond what researchers consider possible in light of their frustrating experiences.

The researchers often know well what they have lost when they attempt and fail. Their choices are rarely documented, of course, and remain private knowledge. But even the researchers themselves may not know what they have lost in the case of learned avoidance, where researchers “know” never to attempt something that may not ultimately be able to be executed.

The cross-jurisdictional comparison suggests, but not definitively, that the US-based fair use doctrine enables researchers working in US settings to be more confident in their copyright knowledge and may be tied to fewer changes in their research design. However, US researchers, who work in a more litigious jurisdiction than Europeans, also avoid projects more often.

Conclusion

A combination of strong copyright law, other laws including privacy and anti-hacking laws, a low appetite for risk, researchers’ lack of legal knowledge, and private corporations’ and governments’ proprietary grip on data inhibit TDM research. The cost to the public is failure to create or make accessible new knowledge. This is in sharp contrast with the exploitation of vast databases by corporate researchers in big tech and cybersecurity.

This research demonstrates a need for governments to address policies that inhibit non-corporate TDM research. To consider copyright alone, an international instrument requiring recognition of the right to research in copyright regimes could be a step toward lowering some of the obstacles to TDM research within nations and across borders. Another would be a broader exception to anti-circumvention laws (such as the DMCA in the US) to permit TDM research, as is typically given for a range of other purposes in the US by the Copyright Office. (The US currently has a narrowly drawn exemption favoring only certain kinds of TDM uses, and other jurisdictions, such as Canada, have no exemptions at all.) Others would include expanding open access to government databases and requiring government-funded research to be made available in open-access, machine-readable versions. The US Office of Science and Technology Policy’s 2022 memo on public access to federally funded research is a big step in the right direction.[28] Finally, legal provisions could protect researcher rights (such as fair use) against contractual waiver.

Notes

Acknowledgements

Author Declaration: The authors do not have any conflict of interest.

[1] Eleonora Rosati, "Copyright as an obstacle or an enabler? A European perspective on text and data mining and its role in the development of AI creativity," *Asia Pacific law review* 27, no. 2 (2019): 198-217, https://doi.org/10.1080/10192557.2019.1705525.

Michael W. Carroll, "Copyright and the Progress of Science: Why Text and Data Mining Is Lawful," *U.C. Davis Law Review* 53, no. 2 (2019): 893, Available at SSRN: https://ssrn.com/abstract=3531231

Matthew Sag, "The new legal landscape for text mining and machine learning," *Journal of the Copyright Society of the USA* 66, no. 2 (2019): 291-[vi], http://dx.doi.org/10.2139/ssrn.3331606.

[2] Organization for Economic Co-operation and Development. *Measuring the digital economy: A new perspective* (Paris: OECD Publishing, 2014).

Organization for Economic Co-operation and Development. *Data-driven innovation: Big data for growth and well-being* (Paris, OECD Publishing, 2015).

[3] American Council of Learned Societies*, Our Cultural Commonwealth: The Report of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Sciences* (New York: American Council of Learned Societies, 2006).

Alex H. Poole, "Now is the Future Now? The Urgency of Digital Curation in the Digital Humanities," *Digital Humanities Quarterly 7*, no. 2 (2013).

Cathy O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (New York: Crown Publishers, 2016), https://dl.acm.org/doi/10.5555/3002861.

danah boyd and Kate Crawford, "CRITICAL QUESTIONS FOR BIG DATA: Provocations for a cultural, technological, and scholarly phenomenon," *Information, communication & society* 15, no. 5 (2012): 662-679, https://doi.org/10.1080/1369118X.2012.678878.

[4] Matthew Sag, "The new legal landscape for text mining and machine learning," *Journal of the Copyright Society of the USA* 66, no. 2 (2019): 291-[vi], http://dx.doi.org/10.2139/ssrn.3331606.

[5] Jean-Baptiste Michel et al., "Quantitative Analysis of Culture Using Millions of Digitized Books," *Science* 331, no. 6014 (2011): 176–82, https://doi.org/10.1126/science.1199644.

[6] Richard Jean So, *Redlining Culture: A Data History of Racial Inequality and Postwar Fiction* (New York: Columbia University Press, 2020), http://cup.columbia.edu/book/redlining-culture/9780231197731.

Interestingly, one of our interviewees used Richard So's research as an example of high-profile TDM research where the author's exact methods (the sources of his digitized novels, in particular) are not disclosed in the final published work. As discussed below, this may be due to fear and uncertainty about legal repercussions from TDM methodologies, which is widespread in the field.

[7] Hamid Ekbia et al., "Big Data, Bigger Dilemmas: A Critical Review," *Journal of the Association for Information Science and Technology* 66, no. 8 (2015): 1523-1545, https://doi.org/10.1002/asi.23294.

Alison Langmead et al., "Towards Interoperable Network Ontologies for the Digital Humanities," *International Journal of Humanities and Arts Computing* 10, no. 1 (2016): 22-35, https://doi.org/10.3366/ijhac.2016.0157.

See Eduardo Navas, Owen Gallagher, and xtine burrough, *The Routledge Handbook of Remix Studies and Digital Humanities* (Milton: Taylor and Francis, 2021), https://doi.org/10.4324/9780429355875 .

See Melissa M. Terras, Julianne Nyhan, and Edward Vanhoutte, ed., *Defining Digital Humanities: A Reader*, Enhanced Credo Edition (Farnham, Surrey: Ashgate, 2016), https://doi.org/10.4324/9781315576251 .

See Kristen Schuster and Stuart Dunn, ed., *Routledge International Handbook of Research Methods in Digital Humanities* (Abingdon, Oxon: Routledge, 2020), https://doi.org/10.4324/9780429777028 .

[8] Matt Burton et al., "New Scholarship in the Digital Age: Making, Publishing, Maintaining, and Preserving Non-Traditional Scholarly Objects," *Digits* 22, no. 1 (2019): 7. [no d.o.i]

[9] Peter McCracken and Emma Raub, "Licensing Challenges Associated With Text and Data Mining: How Do We Get Our Patrons What They Need?" *Journal of Librarianship and Scholarly Communication* 11, no. 1 (2023), https://doi.org/10.31274/jlsc.15530.

[10] Rachael Gayza Samberg and Cody Hennesy, "Law and Literacy in Non-Consumptive Text Mining: Guiding Researchers Through the Landscape of Computational Text Analysis," in *Copyright Conversations: Rights Literacy in a Digital World* (UC Berkeley, 2019), https://escholarship.org/uc/item/55j0h74g.

[11] Scott Althaus et al., Building Legal Literacies for Text Data Mining (California: University of California, Berkeley, 2021), https://doi.org/10.48451/S1159P.

[12] J. Band and Brandon Butler, "Overlapping Forms of Protection for Databases," in *Overlapping Intellectual Property Rights*, 2nd ed., ed. Neil Wilkof and Shamnad Basheer (New York: Oxford University Press, 2023), https://doi.org/10.1093/oso/9780192844477.001.0001

[13] Michael W. Carroll, "Copyright and the Progress of Science: Why Text and Data Mining Is Lawful," *U.C. Davis Law Review* 53, no. 2 (2019): 893, Available at SSRN: https://ssrn.com/abstract=3531231

[14] Matthew Sag, "The new legal landscape for text mining and machine learning," *Journal of the Copyright Society of the USA* 66, no. 2 (2019): 291-[vi], http://dx.doi.org/10.2139/ssrn.3331606

[15] Thomas Margoni and Martin Kretschmer, "A Deeper Look into the EU Text and Data Mining Exceptions: Harmonisation, Data Ownership, and the Future of Technology," *GRUR International* 71, no. 8 (2022): 685–701, https://doi.org/10.1093/grurint/ikac054.

Rossana Ducato and Alain M. Strowel, "Ensuring Text and Data Mining: Remaining Issues With the EU Copyright Exceptions and Possible Ways Out," *SSRN Scholarly Paper* (Rochester, NY, 2021), https://papers.ssrn.com/abstract=3829858.

Gabriella Svensson, "Text and Data Mining in EU Copyright Law," (Master thesis, Uppsala University, 2020).

Rossana Ducato and Alain Strowel, "Limitations to Text and Data Mining and Consumer Empowerment: Making the Case for a Right to 'Machine Legibility'," *IIC - International Review* *of Intellectual Property and Competition Law* 50, no. 6 (2019): 649–84, https://doi.org/10.1007/s40319-019-00833-w.

[16] Allan Rocha de Souza, "Copyright, Human Rights, and the Social Function of Property in Brazil," in *Global Intellectual Property Protection and New Constitutionalism: Hedging Exclusive Rights*, ed. J. Griffiths and T. Mylly (Oxford: Oxford University Press, 2021), 295-317, https://doi.org/10.1093/oso/9780198863168.001.0001.

Sean Flynn et al, "Research Exemptions in Comparative Copyright," Joint Pijip/TLS Research Paper Series no. 75, (American University: 2022), Available at SSRN: https://ssrn.com/abstract=3961017.

[17] Martin Senftleben, "Copyright, Limitations and the Three-Step Test – An Analysis of the Three-Step Test" in *International and EC Copyright Law* (The Hague: Kluwer Law International, 2004).

[18] Martin Senftleben, "Compliance of National TDM Rules with International Copyright Law: An Overrated Nonissue?" *IIC - International Review of Intellectual Property and Competition* Law 53, no. 10 (2022): 1477-1505, https://doi.org/10.1007/s40319-022-01266-8.

[19] Rita Matulionyte, "Australian Copyright Law Impedes the Development of Artificial Intelligence: What Are the Options?" *IIC - International Review of Intellectual Property and Competition Law* 52, no. 4 (2021): 417-443, https://doi.org/10.1007/s40319-021-01039-9.

[20] Christian Handke, Lucie Guibault, and Joan-Josep Vallbé, "Copyright's Impact on Data Mining in Academic Research," *Managerial and Decision Economics* 42, no. 8 (2021): 1999-2016, https://doi.org/10.1002/mde.3354.

[21] The intersection of contract and copyright can be complex and the courts are not entirely consistent in their treatment of situations where copyright allows what a contract would not. The threat of a copyright infringement claim, however dubious, is still enough to chill behavior at the risk-averse institutions where most research takes place.

[22] Kieran McCarthy, "Hello, You've Been Referred Here Because You're Wrong About Web Scraping Laws (Guest Blog Post, Part 2 of 2)," Technology & Marketing Law Blog, December 9, 2022, https://blog.ericgoldman.org/archives/2022/12/hello-youve-been-referred-here-because-youre-wrong-about-web-scraping-laws-guest-blog-post-part-2-of-2.htm.

[23] Patricia Aufderheide and Peter Jaszi, *Reclaiming Fair Use*, 2d ed. (University of Chicago Press, 2018), https://doi.org/10.7208/9780226374222.

[24] See Text Creation Partnership, https://textcreationpartnership.org (last visited March 23, 2023).

[25] For 20 years after the passage of the Sonny Bono Copyright Term Extension Act (CTEA) in 1998, no published works entered the public domain in the United States. The year 1923 (75 years prior to 1998) became a rule of thumb for the border between copyright and the public domain—pre-‘23 works had shed their copyright before the extension, but 1923-and-later works had been awarded another 20 years of protection. In 2019 the 20-year extension granted by the CTEA expired and the public domain began to grow again. As of 2022, works published through 1926 have all entered the public domain in the US. Perception of copyright limits, however, is as important as real limitations when it comes to chilling researcher behavior.

[26] Cristiano de Souza Therrien, “Law in the Present Future: Approaching the Legal Imaginary of Smart Cities with Science (and) Fiction” (PhD Diss., Université de Montréal, 2020).

[27] Patricia Aufderheide and Peter Jaszi, *Reclaiming Fair Use*, 2d ed. (University of Chicago Press, 2018), https://doi.org/10.7208/9780226374222.

[28] Alondra Nelson, “Ensuring Free, Immediate, and Equitable Access to Federally Funded Research,”, accessed April 26, 2023, https://www.whitehouse.gov/wp-content/uploads/2022/08/08-2022-OSTP-Public-Access-Memo.pdf.

ANNEX I

Right to Research Using TDM Protocol (excluding open-ended answers)

Q1 Thank you for taking this survey about research use of text and data mining (TDM). We understand that by taking it, you consent to giving us this information. This research is led by AUTHORS. We are looking at the challenges faced by researchers doing TDM, especially in regard to copyright issues. This study was given a waiver by the Institutional Review Board of ANONYMIZED. You will be anonymous to us, unless you choose to give us your name, in which case your identity and all information associated with it will be kept confidential. The data are stored securely, via the Qualtrics platform and American University. We may reuse this information in later studies, but always keeping either anonymity or confidentiality, according to the option you have chosen.

Q2 In which countries do you primarily do your research work? Check all that apply.

▼ Afghanistan…Zimbabwe

Q3 In what kind of institutional context do you work? (Check all that apply)

Academic (1)
Governmental (2)
Private Sector Corporate (3)
Private sector nonprofit (5)
Other (4) ____

Q4 What are the areas of research that you conduct or assist others in conducting? Choose all that apply

Communication (1)
Internet (2)
Computer Science (3)
Education (4)
Linguistics (5)
History (6)
Sociology (7)
Criminology (8)
Languages (9)
Literature (13)
Libraries/Information Science (15)
Law (16)
Digital humanities (14)
Other (12) ____

Q5 Do you conduct text mining or data mining (TDM), using material you didn't create yourself, in your work? This means that you research using a body of text or data, produced for other purposes by others, using computer analysis rather than human reading of the text. For instance, you might search a body of Twitter data to find out at what times people are more likely to repost news.

Yes (1)

No (2)

End of Survey "If Do you conduct text mining or data mining (TDM), using material you didn't create yourself, in your work... = No"

Q6 How many years of experience do you have in research?

Less than 5 years (1)

5-14 years (2)

15-24 years (3)

25 years or more (4)

Q7 How do you get access to the text or data you mine? (Check all that have applied in your projects)

I pay a vendor or other business (such as a publisher, data aggregator, government, academic center or other service provider) that hosts this material. (1)

The entity that hosts this material (such as a publisher, data aggregator, government, academic center or other service provider) gives me free access (2)

The library or archive I use has purchased access to the database, which permits text TDM as a standard feature for all users (4)

The library or archive I use made a special arrangement with a vendor to permit datamining in my case (8)

The library or archive I use owns the collection that I datamine and makes it available for that purpose to any interested researcher. (12)

The library or archive I use owns the collection that I datamine and made it available to me under a special arrangement. (13)

I collect the information digitally (e.g., by gathering digital materials from the open web), using off-the-shelf software (5)

I collect the information digitally, using purpose-built software (6)

I digitize analog information, using off-the shelf software (9)

I digitize analog information, using purpose-built software (10)

Other (7) ____

Display This Question:

If “How do you get access to the text or data you mine? (Check all that have applied in your projects) = I pay a vendor or other business”

Or “How do you get access to the text or data you mine? (Check all that have applied in your projects) = The entity that hosts this material (such as a publisher, data aggregator, government, academic center or other service provider) gives me free access”

Q8 I have experienced the following challenges when working directly with an entity, e.g. a publisher, data aggregator, government, academic center or other service provider (check all that apply):

None (1)

The entity didn't answer my initial inquiry. (2)

The entity was slow to reply to my questions. (3)

The vendor's pricing was out of my reach. (21)

The vendor impeded my access to data technically. (20)

The vendor impeded appropriate use of the data. (22)

The vendor did not permit my archiving the data. (15)

The vendor didn't let me share the data appropriately. (23)

The vendor's terms of use/license were confusing. (6)

Other (please explain). (12) ____

Display This Question:

If “How do you get access to the text or data you mine? (Check all that have applied in your projects) = The library or archive I use has purchased access to the database”

Or “How do you get access to the text or data you mine? (Check all that have applied in your projects) = The library or archive I use made a special arrangement with a vendor to permit datamining in my case.”

Or “How do you get access to the text or data you mine? (Check all that have applied in your projects) = The library or archive I use owns the collection that I datamine and makes it available for that purpose to any interested researcher.”

Or "How do you get access to the text or data you mine? (Check all that have applied in your projects) = The library or archive I use owns the collection that I datamine and made it available to me under a special arrangement."

Q9 I have experienced the following challenges via a library or archives (check all that apply):

None (1)

Not applicable (22)

Vendor did not reply to the library/archive's inquiry. (2)

Library was slow to address my questions. (14)

Vendor's pricing was greater than the library's budget permitted. (5)

The terms of use/license were confusing. (6)

The terms of use impeded my access to data technically. (23)

The terms of use impeded appropriate use of the data. (8)

I could not share the data appropriately for my research. (9)

I couldn't keep an archive of the data (15)

The library has contracts with vendors that limit or ban TDM research (19)

The librarians/archivists did not negotiate special terms with vendors to permit TDM, when the standard agreement did not allow it (21)

The library imposed restrictions on use of data that go beyond the law or vendor requirements (20)

Other (please explain): (13) ____

Q10 Various legal issues can affect TDM research. Please check all that you have experienced.

None (9)

Copyright law (1)

Vendor terms of service or EULAs that I had to agree to. (2)

License agreements or terms of service the library or archive signed with a vendor (3)

Anti-hacking laws (4)

Privacy laws (7)

Other (8) ____

Q11 Have you experienced impediments doing TDM, specifically because of copyright law? Please check all that apply:

No problems (1)

The law impedes collection or access to data. (Please explain or give examples.) (2)

The law impedes use of the data. (Please explain or give examples.) (3)

The law impedes retention/storage of the data. (Please explain or give examples.) (9)

The law impedes sharing data. (Please explain or give examples.) (10)

The law is unclear. (Please explain or give examples.) (4)

Other (5) ____

Q12 Different countries have different copyright laws. Please check all that apply in your experience:

I have no experience with cross-border copyright research (1)

The fact that other countries had different laws affected our project (please explain, if so) (2)

We lacked expertise to manage the different copyright regimes (please explain what happened as a result, if so) (3)

We proceeded without worrying about different countries' different copyright laws (4)

Other (5)

Q13 Have the problems referenced in this survey discouraged you from doing TDM research? (check all that apply)

Yes, I have avoided taking on projects because of one or more of these problems. (If so, please explain) (1)

Yes, I have had to change a project because of one or more of these problems. (If so, please explain) (2)

Yes, I have abandoned a project I started because of one or more of these problems. (If so, please explain) (5)

Unsure (3)

No (4)

Q14 Who gives you helpful support on legal issues that come up in your TDM research? Check all that apply.

a superior/boss (1)

a lawyer (2)

my colleagues/peers (3)

my friends (4)

librarian (8)

myself (5)

other (please explain) (7) ____

Q15 Who flags legal problems or raises concerns about legal issues that can impede progress in your TDM research? (check all that apply)

a superior/boss (1)

a lawyer (2)

my colleagues/peers (3)

my friends (6)

a librarian (8)

myself (4)

other (5) ____

Q16 How confident are you that you personally know enough about copyright law concerning TDM to make appropriate decisions?

Very confident (1)

Somewhat confident (2)

Neither confident nor unsure (3)

Somewhat unsure (4)

Very unsure (5)

Q17 How confident are you that your team or organization has enough knowledge about copyright law concerning TDM to help you make appropriate decisions?

Very confident (1)

Somewhat confident (2)

Neither confident nor unsure (3)

Somewhat unsure (4)

Very unsure (5)

Q18 Have you selected materials that are open-access (i.e., licensed to the public for free use) or in the public-domain (i.e. free of copyright) for TDM, specifically because they did not present as many legal concerns as copyrighted materials?

Yes (1)

No (2)

Unsure (3)

Q19 Would you like to share any good experiences in doing TDM research that you hope would become best practices?

____

Q20 Would you like to share any other experiences doing TDM research that we haven't asked about?

____

Q21 Would you like to have a conversation/interview with us? If so, please put your name and email here. You will lose anonymity by doing this, but you will not lose confidentiality.