Building Legal Literacies for Text Data Mining

Building Legal Literacies for Text Data Mining

SCOTT ALTHAUS; DAVID BAMMAN; SARA BENSON; BRANDON BUTLER; BETH CATE; KYLE K. COURTNEY; SEAN FLYNN; MARIA GOULD; CODY HENNESY; ELEANOR DICKSON KOEHL; THOMAS PADILLA; STACY REARDON; MATTHEW SAG; RACHAEL SAMBERG; BRIANNA L. SCHOFIELD; MEGAN SENSENEY; TIMOTHY VOLLMER; AND GLEN WORTHEY

UNIVERSITY OF CALIFORNIA, BERKELEY
BERKELEY, CALIFORNIA

Contents

Part II. Teaching the Literacies

Acknowledgements

NATIONAL
ENDOWMENT
FOR THE
HUMANITIES

The Building Legal Literacies for Text Data Mining Institute (Building LLTDM) was made possible by a grant from the National Endowment for the Humanities. Any views, findings, conclusions, or recommendations expressed in this book do not necessarily represent those of the National Endowment for the Humanities.

Thank you to the project team, institute participants, and staff at the UC Berkeley Library for making Building LLTDM a success.

Introduction and structure of this book

Until now, humanities researchers conducting text data mining in the U.S. have had to maneuver through a thicket of legal issues without much guidance or assistance.

UC Berkeley Library led more than a dozen institutions in submitting (and receiving) a grant to create a National Endowment for the Humanities Institute entitled Building Legal Literacies for Text Data Mining (Building LLTDM). We wanted to empower digital humanities researchers and professionals (librarians, consultants, and other institutional staff) to confidently navigate United States law, policy, ethics, and risk within digital humanities text data mining projects—so that they could more easily engage in this type of research and contribute to the advancement of knowledge.

On June 23-26, 2020, we welcomed 32 digital humanities researchers and professionals to the institute. After months of preparation, we had been looking forward to working and learning together at UC Berkeley, but the world had other plans. Due to the global health crisis, we had to transform our planned in-person, intensive workshop into an interactive and relevant remote experience.

The pandemic meant we had to transition everything online. The substantive content was pre-recorded and delivered in a flipped classroom model. Faculty created a series of short videos, and shared readings relevant to the legal literacies. We also provided the video transcripts and slides to participants to promote accessibility and accommodate multiple learning styles.

This book explores the legal literacies covered during the virtual institute, including copyright (both U.S. and international law), technological protection measures, privacy, and ethical considerations. It describes in detail how we developed and

delivered the 4-day institute, and also provides ideas for hosting shorter literacy teaching sessions. Finally, we offer reflections and take-aways on the institute.

PART I
SUBSTANTIVE LITERACIES

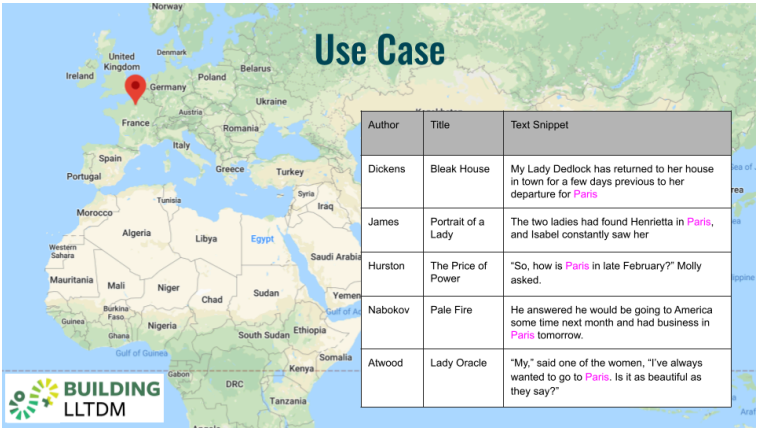
1. Copyright

DAVID BAMMAN; BRANDON BUTLER; KYLE K. COURTNEY; AND BRIANNA L. SCHOFIELD

Copyright use case

To illustrate some of the copyright issues that arise for text data mining ("TDM") research, we can consider a use case that raises a number of common issues in different applications in TDM. Let's assume we have some collection of texts in varying copyright status, and we want to carry out some algorithmic transformation of those texts and publish the results. So envision this scenario: you're a researcher who has a large collection of texts already digitized, and what you want to do is perform some natural language processing on those texts and visualize their results for the broader public.

In particular, you have a large collection of fictional texts and what you want to do is extract all of the mentions of place names from each of these texts and plot those place names on a map. This is an aspect of text mining that's known by a number of different terms, including toponym resolution and geolocation, but it starts from the fundamental problem of named entity recognition ("NER")—of simply recognizing all of the names in the text that refer to places. So you extract those place names, georeference them to latitude/longitude coordinates on a map, and the visualization you want to present is effectively an organizing system for your fiction corpus: whenever a user clicks on a place in a map, you want to present to them a list of all the times when that place was mentioned in a book in your collection, including a snippet from the text where that place name was mentioned.



Disambiguating place name mentions in TDM analysis.

In the example above, a user has clicked on "Paris" and we can see that "Paris" shows up in works by Charles Dickens, Henry James, Zora Neale Hurston, Vladmir Nabokov, and Margaret Atwood. This involves a fundamental transformation of the data in several ways—not least of which is the fact that you are disambiguating place name mentions—and asserting, for example, that when Charles Dickens mentions "Paris" in *Bleak House*, he's not talking about Paris, TX, he's talking about Paris, France.

In this use case, the books you hold in your collection of fiction are relatively heterogeneous, and span over two hundred years—being published anywhere between 1800 and 2020. All of these books originate in print form (so, for example, they are not born digital as markdown files or Kindle editions); they're print works that you've scanned and OCR'd (that is, an optical character recognition tool has been used so that all of text on a page image has been recognized). Your corpus also includes some unpublished manuscripts that are housed within your own library collections. The transformations that you are performing on this dataset is NER and toponym—where you extract all mentions of place names from

text, and then ground those place names in specific coordinates on a map.

But your use case doesn't just stop at running an NER system on your dataset and plotting those names on a map. You know that just about all of the existing NER systems out there are trained on data that's not fiction, and you know you can do better if you train your own system on data that actually includes it. So what you want to do in your project is create training data in the domain you care about—fiction written between 1800 and 2020. This data is going to help you train better NER systems for recognizing places as they show up in literature. To achieve this, you take 1,000 novels from your dataset and annotate all of the place names that show up in a 500-word sample of each one, effectively creating a total labeled dataset that's 500,000 words long. Your primary goal in creating this dataset is to make NER better for your visualization, but at the same time you recognize that this dataset really would be of tremendous value to the research community. It would allow computational researchers to train and evaluate models for NER on a domain that simply does not have much annotated data, and you would be helping the community be less focused on news while at the same time helping improve these tools for other researchers in the humanities who work with these texts. So in addition to publishing your interactive visualization of place names mentioned in fiction, you also want to publish your annotated dataset of 500,000 words for others to use. You value reproducibility as a scientific goal and want to have that dataset out there in the world. You can see below what one of these annotations would look like—you want to publish a 500-word snippet of, for example, Vladimir Nabokov's *Pale Fire*—along with your annotations for which words are places within it, for all of the 1,000 novels in your annotated dataset.

Examples of TDM annotation.

So those are the two main aspects of this use case we're working with: (1) creating a visualization plotting place names extracted from fiction on a map using algorithmic transformations of NER and toponym resolution, and (2) publishing a new annotated dataset of place names mentioned in these works. Let's keep this use case in mind through this chapter, and we'll return to it at the end.

Copyright basics

Copyright law is part of a legal system that covers both creation and use. Here we will cover the copyright basics: what copyright is, what copyright protects, and how long copyright protection lasts. Additionally, copyright law is filled with exceptions and exemptions that strike a balance between the exclusive rights granted to creators and the rights of many users, including TDM researchers. It is critical that TDM researchers understand both the rights and the exceptions, with an emphasis on fair use, which in the TDM context is one of the most important rights that provides a legal justification for using the material that drives a TDM project. However, before the exceptions, which are covered in a later section, let us start with the copyright basics.

In 1710 the English parliament passed the Statute of Anne. This new law gave authors, for the first time in history, an economic incentive to create new works: Authors had control of their own works, and the copies made, via a limited economic monopoly—not unlike our modern understanding of copyright. This captured the first balance between authors' rights and the public benefit of copyright, when works drop into the public domain. This temporary economic right was enough incentive for authors to continue to create new works. And, of course, when the rights expired (after 14 years) the work would drop into the public domain, and anyone could use the work thereafter without permission. This encapsulated the cycle of copyright: creation, control, and

expiration, with the hope that further works could be created using what dropped into the public domain. And in fact, the Act starts with the language, "An Act for the Encouragement of Learning."

This concept moved into the U.S. system in our Constitution. Certainly, the members of the United States Constitutional Convention were aware of the ideas of control and censorship as the U.S. emerged from English rule. In 1790, pursuant to their Constitutional authority under Constitutional Clause: Article 1, section 8, clause 8: "To promote the progress of science and useful arts, by securing for limited times to authors and inventors the exclusive right to their respective writings and discoveries," the Congress passed, and George Washington signed, the first copyright law in the United States. It was also titled "An Act for the Encouragement of Learning" and featured the same balance that the English had revolutionized with the Statute of Anne: an incentive of a limited economic monopoly granted to authors over their works, followed by the expiration of those rights when the work then would drop into the public domain.

The current copyright law on the books is based on that initial 1790 law, but now it is in the U.S. code as the Copyright Act of 1976. It protects original works of authorship that are fixed in any tangible medium of expression.

But what is an "original work of authorship"? An original work must embody some "minimum amount of creativity." Courts have held that almost any spark beyond the trivial will constitute sufficient originality. On the other hand, the Supreme Court ruled in 1991 that a garden variety alphabetical, white pages telephone book lacks the minimum creativity necessary for copyright protection. This is called the *Feist* case. The U.S. Supreme Court held in *Feist Publications v. Rural Telephone Service* that copying of a white pages book was not infringement because there was no existing copyright. However, although facts themselves are not copyrightable, the way the items are categorized and arranged may be original enough to satisfy the originality requirement.

Ultimately, this creativity threshold is also touched upon in

another part of the Copyright Act, section 102(b), which states that copyright's threshold for originality does extend to "any idea, procedure, process, system, method of operation, concept, principle, or discovery." From this we gather an important point for authors: facts are not copyrightable.

But, beyond creativity, what is copyright, really? Is it a "bundle of rights"? A limited economic monopoly for authors? Or, in the Constitutional narrative, is it a system "to promote the progress of science and the useful arts"?

Well for copyright to work, it has to be all three. The cycle of creation, dissemination, and expiration of rights into the public domain is a critical component of copyright law. Without this balance, the system loses its value, or prevents the public from receiving the benefit of the bargain. The bargain is made by granting limited economic monopolies to incentivise creation, and then, after expiration of the monopoly, the benefit is effectively giving that material to the public for unimpeded use, thus inspiring more works to be harnessed and used.

When a work is creative and fixed, creators automatically get this exclusive bundle of rights. These are the rights: to reproduce the work copies; to prepare derivative works; to distribute copies; to perform the copyrighted work publicly; and to display the copyrighted work publicly.

In 1790, when George Washington signed our country's first copyright law into existence, copyright protection was for books, maps, and charts. However, under the Copyright Act of 1976, the subject matter of copyright has been extended into these eight extensive categories: (1) literary works; (2) musical works, including any accompanying words; (3) dramatic works, including any accompanying music; (4) pantomimes and choreographic works; (5) pictorial, graphic, and sculptural works; (6) motion pictures and other audiovisual works; (7) sound recordings; and (8) architectural works. As Congress indicated in the creation of these categories, there is a great deal of material that has the potential to be protected by copyright.

Occasionally we learn about copyright by understanding what's not copyrightable. For example, there are other parts of intellectual property law that are not under the umbrella of copyright. Slogans and logos, for example, are part of trademark law. Trademark law is generally all about what the mind of the consumers think as the source of the material when they see a logo. Patent law covers new and useful ideas such as processes, methods, and systems that are separate from copyright. Secret formulas and recipes that are not disclosed to the public are generally considered trade secrets. They derive economic value by not being disclosed to the public. And then, of course, there's raw data. As we know from *Feist*, our white pages telephone book case, you can't copyright a fact. Applying that holding here, raw data then, viewed as a set of facts, is uncopyrightable.

In order to know understand copyright, you need to know these six things: that creators get copyright if the work is original, creative, and fixed in a tangible medium of expression; that no registration is required to get copyright—the work is automatically granted protection under copyright if it's creative and fixed; that the grant of rights to the author is represented by the exclusive bundle of rights in section 106; that there is a wide range of protected works; and they have a long term of protection. However, as we will cover, despite all of these rights there are numerous exceptions and limitations. The focus of our inquiry for TDM will be section 107 fair use.

However, before we move to the exceptions, we will cover a critical part of the copyright cycle: the public domain. When copyright was first passed by Congress in 1790, Congress set a term of protection for 14 years, with a potential of an additional 14 years if the creator renewed the copyright. In 1909, Congress doubled that timeline and copyright moved to a 28-year term of protection with a potential 28-year renewal. In 1976, in accordance with harmonizing international copyright law, as part of the Copyright Act of 1976, the term was set to life of the author plus 50 years. And in 1998, that term was expanded by Congress for an additional 20 years. And so,

copyright today is measured by the life of the author plus 70 years. But what happens after expiration? Our next segment will cover that which is in the public domain.

The public domain

The previous section of this chapter covered what copyright is, what copyright protects, and how long copyright protection lasts. This section addresses the flip side of copyright: the public domain.

In copyright, the public domain is the commons of material that is not protected by copyright. Anyone is free to use, copy, share, and remix material that is in the public domain. The public domain includes works for which the copyright has expired, works for which copyright owners failed to comply with "formalities," and things that are just not copyrightable at all. This section discusses each of these categories in turn.

A word of caution: Some people mistakenly think that the "public domain" means anything that is publicly available. This is wrong. The public domain has nothing to do with what is readily available for public consumption. This means that just because something is on the internet, it doesn't put it in the public domain.

Remember that under today's copyright laws, a work of creative, original expression simply needs to be "fixed in a tangible medium" to be eligible for copyright protection. If Philippa Photographer takes a photograph and puts it online on her blog, it doesn't mean that she

is also granting you permission to reuse it. The default is that Philippa's photo is protected by copyright and not in the public domain.

Copyright expiration

One way content enters the public domain and becomes free of copyright protection is through copyright expiration.

Copyright protects works for a limited time. After that, copyright expires and works fall into the public domain and are free to use. Under United States copyright law, in 2021 (the year this book is being released) all works first published in the US in 1925 or earlier are now in the public domain due to copyright expiration. That said, unpublished works created before 1926 could still be protected by copyright. And under today's copyright laws, works created by an individual author today won't enter the public domain until 70 years after that author's death.

When copyright does expire, the work is in the public domain and there are no copyright restrictions. For example, the book *Alice in Wonderland* is in the public domain, as are *New York Times* articles from the 1910s, because their term has expired. This means anyone may do anything they want with the works, including activities that were formerly the "exclusive right" of the copyright holder, like making copies and selling them.

Failure to comply with formalities

Another way a work may enter the public domain is through a failure to comply with formalities.

Copyright law used to require copyright owners to comply with certain requirements called "formalities" in order to secure copyright protection. These formalities included things like requiring the copyright owner to mark the work with a copyright notice and renew the initial term of copyright. These requirements existed in some form through March 1989. Because many authors failed to comply, many works from between 1926 and March 1989 may be in the public domain. But this analysis needs to be done on a case-by-case basis based on the facts surrounding a particular work. In some cases, a fair use analysis may be easier than making a conclusion about the copyright status of a work. (Fair use is discussed later in this chapter.)

If a work is in the public domain for failure to comply with formalities, as with copyright expiration, there are no copyright restrictions.

> **Additional Resources:** For more information on how to evaluate whether a work is in the public domain due to copyright expiration or a failure to adhere to the previously required formalities, see Peter Hirtle's Copyright Term and the Public Domain in the United States and the Samuelson Law, Technology & Public Policy Clinic at Berkeley Law's Is it in the Public Domain? handbook and flowcharts.

Uncopyrightable subject matter and other

exclusions

In addition to copyright expiration and a failure to comply with formalities, copyright law also sets out things that are simply not protected by copyright, and those things are also in the public domain. This goes back to a point about the purpose of copyright: The public domain is important to the production of creativity; authors need these essential building blocks with which to work.

For example, facts are a category of things that are not copyrightable—even if those facts were difficult to collect. For instance, suppose that a historian spent several years reviewing field reports and compiling an exact, day-by-day chronology of military actions during the Vietnam War. Even though the historian expended significant time and resources to create this chronology, the facts themselves would be free for anyone to use. That said, the way that the facts are expressed—such as in an article or a book—is copyrightable.

Under United States copyright law, other types of works and subject matter do not qualify for copyright protection include: names, titles, and short phrases; typeface, fonts, and lettering; blank forms; and familiar symbols and designs. It is worth noting that other areas of intellectual property, such as patent or trademark law, could provide protection for categories that are not eligible for copyright protection.

The Copyright Act also provides that works created by the United States federal government are never eligible for copyright protection, though this rule does not apply to works created by U.S. state governments or foreign governments. And under the government edicts doctrine, judicial opinions, administrative rulings, legislative enactments, public ordinances, and similar official legal documents are not copyrightable for reasons of public policy.

Additional Resource: For more information on what is not protected by copyright, see the United States Copyright Office's Circular 33: Works Not Protected by Copyright.

Public domain and TDM projects

If a text data mining project involves only public domain materials (like federal government documents or newspaper articles published in the United States in the 1890s), there is no need to investigate whether accessing, using, and sharing of these public domain materials is allowable under an exception to copyright or whether you need permission from the copyright owner to use the work. This is because anyone can use public domain materials without infringing on copyrights, including activities that were formerly the "exclusive right" of the copyright holder like making copies of, sharing, and adapting the work.

A word of caution: Just because a work is in the public domain, this does not preclude consideration of other legal issues. Moreover, it is important to note that working with "low-friction" data like public domain works can exacerbate social biases that can exist in the collection. For example, pre-1926 works in the public domain are likely to be dominated by white, male authors.

Copyright, licensing, and permissions

You can learn about licenses in more detail in the Licensing chapter of this book, but copyright and licensing are so closely connected that we think it's important to say a bit about them here, too.

A license grants permission, and may limit your rights, too

A license is a grant of authorization from a copyright holder to exercise one of their exclusive rights—in a research library context, typically the license is to copy or display protected works on your computer. Databases, journal literature, and other electronic content is often made available under a license either directly to the user or to an institution (typically a library) on behalf of its users. The license tells you which uses have been authorized, and authorization is often conditioned on the licensee doing certain things (most importantly, for commercial entities: paying a fee!).

A license may also include promises by the institution or the user not to engage in certain uses, or only to use licensed content under certain circumstances.

What this means for TDM researchers is that your institution may already have a license that defines what sorts of uses you can make of licensed content. You'll need to read the license, or talk to someone who understands the license terms, to learn more about what uses are possible. You may also need to negotiate a new license to enable your use, especially if you require special kinds of access to a vendor's content in order to conduct your research.

We talk a lot more about this in the chapter on licensing, but the key thing to understand, here, is that if your use is permitted by a license, then you don't have to worry about copyright. If it is not clearly permitted, you will need to think about fair use and other alternatives. Fair use may permit uses that are not mentioned explicitly in a license, because a fair use does not require permission. If your use is **forbidden** by the license, then even if your use doesn't violate copyright law, you or your institution could still face liability for breach of contract. The most likely negative consequence for violating a license is that you or your institution lose access to the resource, at least temporarily.

Creative Commons and other open licenses

Some works are available under open licenses that allow anyone to make specific uses of copyrighted works without the need to pay or seek additional permission from the owner. Creative Commons ("CC") licenses are the most well-known open licenses. Creative Commons is a nonprofit organization that offers a simple, standard way to grant copyright permissions for creative works, and a suite of license options that lets authors impose some commonly-sought limitations on would-be users. Instead of the "all rights reserved" default, copyright owners can apply a CC license that allows others to use and share their works without seeking permission. It is important to pay attention to the specific terms of the license: almost all of the CC licenses require attribution, some can require you to "share alike" (i.e., to attach the same license to any work you create using the licensed work), and some restrict commercial uses or the creation of derivative works (like translations). For example, a work marked CC-BY-NC means that it is licensed for other people to use and share as long as the work is appropriately credited, but commercial uses are not allowed.

Creative Commons also offers a tool, CC0, that allows a copyright

owner to waive all copyrights (and some related rights) in works. Because it is a complete waiver of rights, CC0 doesn't require attribution.

CC licenses are especially common in the academic world, and research funders increasingly require their grantees to use them. But even non-academic works may be made available under CC licenses. For example, some museums distribute photographs of works in their collections under open licenses.

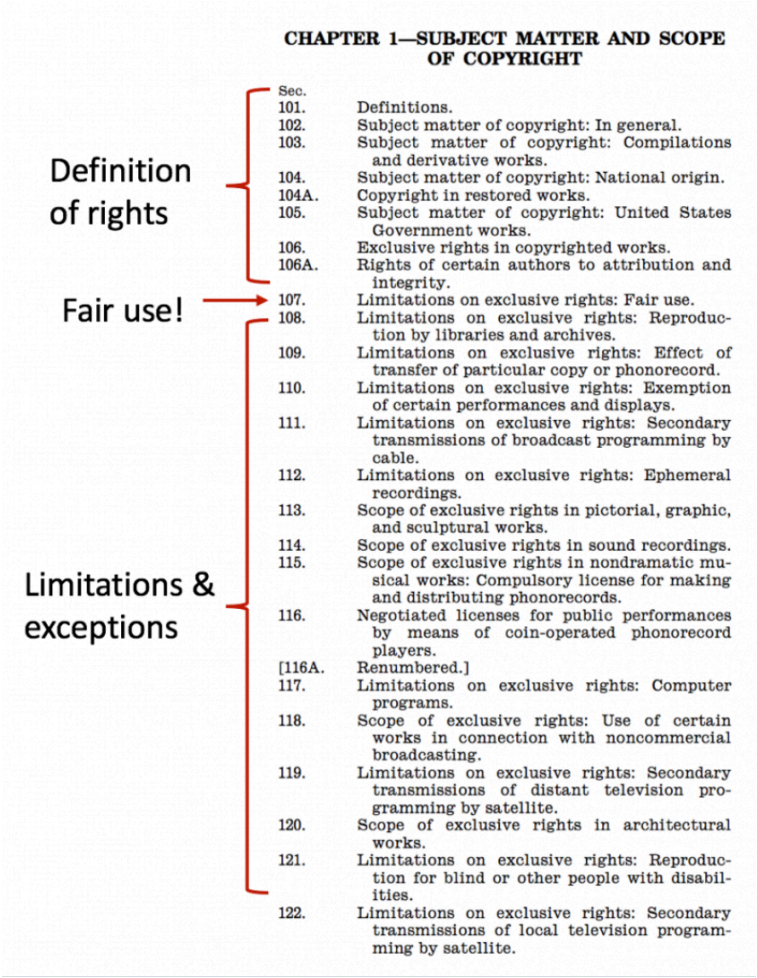
Bottom line: If works are made available under a public license, then (just like any other license) these works can be used in ways that comply with the terms of the license. If a TDM project involves works that are made available under a license, including a public license (like a CC license), these works can certainly be used in ways that comply with the terms of the license. If your use is beyond the terms of the license, or forbidden, things get more complicated. This issue will be discussed further in the chapter on licensing.

A word of caution: Don't forget to consider other legal and ethical issues discussed in this book when using works made available under license. For example, researchers have documented a bias in machine learning resulting from the widespread use of "low-friction" data. Datasets like the Enron email corpus are widely used because they present few legal concerns, but the predominantly white, male, corporate context in which they were created can impart a bias to analyses derived from the corpus.

Fair use: A critical copyright exception

Imagine if all creators had to wait for a copyrighted work to be in the public domain before they used that work? Or if scholars always had to seek permission to use or quote, and that permission could be denied with no recourse? Happily, copyright law gives us the flexibility to allow some uses that are made during the copyright term without permission. One of the most famous of all the copyright limitations in the Copyright Act does just that: the fair use exception.

Under fair use, a person may use certain amounts of copyrighted material without permission from the copyright owner in some circumstances. The doctrine itself was rooted in both English and U.S. case law, but was eventually codified in section 107 of the U.S. Copyright Act. Fair use, as you can see in the image below, sits in the middle of the organized balance in the Copyright Act; it is squeezed right between the exclusive rights and more specific exceptions.



Chapters of U.S. copyright law.

Fair use is for everyone. And since TDM often involves copying large amounts of copyright material in order to mine the content, it is useful to the TDM researcher, because TDM involves access, coping, and processing works that may be in copyright.

Even if TDM researchers have authorized access to the materials,

copying a substantial part of these works may infringe copyright in those works. And so might distribution after the copying and processing is over.

If a use is a fair use, it is not infringement. Again, imagine if you had to get permission to provide analysis, commentary, or criticism of someone's copyrighted work. If there were no fair use, and copyright holders could forbid you from using the work without permission, this would vastly stifle free expression and new modes of scholarship, like TDM.

Fair use is a user's right that allows individuals to exercise one or more of the exclusive bundle of rights of the copyright owner, without obtaining the permission from that copyright owner, and without the payment of any license fee.

To decide whether a use is fair, courts must consider at least four factors that are specifically mentioned in the Copyright Act.

17 U.S.C. §107

Notwithstanding the provisions of sections 106 and 106A, the fair use of a copyrighted work, including such use by reproduction in copies or phonorecords or by any other means specified by that section, for purposes such as criticism, comment, news reporting, teaching (including multiple copies for classroom use), scholarship, or research, is not an infringement of copyright.

In determining whether the use made of a work in any particular case is a fair use the factors to be considered shall include—

1. the purpose and character of the use, including whether such use is of a commercial nature or is

for nonprofit educational purposes;

2. the nature of the copyrighted work;
3. the amount and substantiality of the portion used in relation to the copyrighted work as a whole; and
4. the effect of the use upon the potential market for or value of the copyrighted work.

The first factor is the purpose and character of the use. Here courts ask whether the material has been transformed by adding new meaning or expression, or whether value was added by creating new information, meaning, or understanding. When a work is used for a different purpose than the original, the factor will likely weigh in favor of fair use. If it simply acts as a substitute for the original work, the less likely it is to be fair. Courts may also look at whether the use of the material was for commercial or noncommercial purposes under this factor, but this is rarely a determinative consideration.

The second factor looks at the nature of the copyrighted work. Here courts look at whether the copyrighted work that was used is creative or factual in nature (a song or a novel vs. technical article or news item). The more factual the work, the more likely this factor will weigh in favor of fair use. On the flip side, the more creative the copyrighted work, the more likely this factor is to weigh against fair use. Courts may also consider whether the copyrighted work is published or unpublished. If the work is unpublished, this factor is less likely to weigh in favor or fair use. Note that this factor has been slightly deemphasized by the courts over the last twenty years.

The third factor is the amount and substantiality of the portion taken. Under this factor, courts look at how much of the work was taken, both quantitatively and qualitatively. Quantitatively, courts look at how much of the original work was used (e.g., all the pages,

the entire work of art). Qualitatively, some courts look at whether the "heart" of the work was taken (e.g., the essential bit of the work that is why people want to engage and acquire the work). The more that is taken, quantitatively and qualitatively, the less likely the use is to be fair. That said, copying a full work can absolutely be a fair use depending on the circumstances.

Finally, the fourth factor is the effect of the use on the potential market. The essential question courts ask here is whether this use will undermine the market, or the potential market, for the work that was copied. In assessing this factor, courts consider whether the use would hurt the market for the original work (for example, by displacing sales of the original). There's a lot more nuance to this factor, but let's move ahead to transformative fair use.

Transformative fair use

In 1841, the U.S. decided its first fair use case. And, as case law developed, so did new and different fair use theories. One of the more interesting developments in fair use litigation was the emergence of transformative fair use. Use of any copyrighted materials is substantially more likely to pass fair use muster if the use is transformative. A work is transformative if, in the words of the Supreme Court, it "adds something new, with a further purpose or different character, altering the first with new expression, meaning or message." Transformative fair use is still a use without permission, but it is the legal engine which drives scholarship, research, and teaching.

The last two decades has seen a shift in courts analysis of the fair use test in creative endeavors like these. In transformative fair use, we see the courts collapsing the traditional "four fair use factors" to ask the following questions:

1. Does the new use transform the material, by using it for a

different purpose?
2. Was the amount taken appropriate to the new, transformative purpose?

And, importantly, it helps to identify that this new transformative use has a different purpose than the original item's purpose. For example, the original purpose of the fictional books in the Copyright Use Case was for entertainment. The new use should be for a different purpose—and arguably, the new purpose would be to add commentary or analysis that reveals a new meaning or message, altering the original works with new commentary, expression, meaning, or message.

Fair use law is well equipped to be adaptable to various scenarios. That's the purpose of fair use: flexibility. Fair use is not mechanically applied or even weighed equally. Courts take into account all the facts and circumstances of a specific case to decide if use of copyrighted material is fair. And scholars, TDM researchers, librarians, lawyers, students, staff, and faculty can also use the fair use statute and legal decisions to evaluate their own fair use risk calculus for their own scenarios.

In the next section, we'll look to see how fair use is applied specifically in the TDM mining context.

Fair use & TDM

As you've seen, fair use is a judge-made right that evolves as it is applied, case-by-case. Lawsuits about research and education are few and far between, so TDM researchers are unusually fortunate to have a long and deep line of cases that provides fairly clear support for the kinds of things they do with in-copyright material. Search engine operators like Google were sued early in their history, then related machine learning and computer analysis technologies were challenged, and finally massive digitization of research materials

was challenged in the Google Books and HathiTrust cases, which we'll explore in depth.

What's key for TDM researchers to know is that courts have now blessed core TDM practices many times over. If anything is knowable in fair use law, we now know that these core TDM research methods are well-suited for fair use.

Key case: *Authors Guild v. Google*

Let's take a look at how fair use applies to text data mining using a recent case, *Authors Guild v. Google*, as an example. This case arose when Google made digital copies of millions of books from partner research libraries, and made the resulting corpus searchable through its Google Books service. Google provided digital copies to the libraries who provided print books from their collections, and the libraries banded together to create the HathiTrust to manage the collective collection of those scans, together with other digital content.

Using Google Book Search, users could identify books that contained a desired word or phrase. Google's search results showed limited snippets of the text (about an eighth of a page) so users could see their term in context and get a better sense of the result's relevance to their interest. They also linked users to local libraries and online bookstores where copies of the work could be found. When the Authors Guild sued alleging infringement, Google argued that Book Search was a quintessential fair use. The influential Second Circuit court of appeals agreed. The Authors Guild sued HathiTrust and some of its members in a separate case, with the same result—fair use.

For TDM researchers, it is important to look at the three key uses that the court was evaluating in the Google Books case. Comparing your activities to the ones analyzed here will be extremely helpful as

you figure out how fair use might apply to your research. The uses in the Google Books case were:

1. Copying millions of complete in-copyright books to create a search index;
2. Displaying "snippets" of in-copyright text as search results to users in the public; and
3. Ngram graphs showing the frequency of words and phrases in the corpus over time.

These kinds of practices—compiling works into a machine-readable corpus and revealing relevant portions of the corpus to the public to substantiate or instantiate the results of machine analysis—are likely to recur in many, many TDM projects. Researchers will learn a great deal from a close reading of the court's clear and detailed application of fair use to both practices.

First factor and transformative use

Recall that the first factor asks us to look at the purpose and the character of the use, and central to the analysis is whether a use is "transformative," with transformative uses being much more likely to be fair use.

In *Authors Guild v. Google*, the Second Circuit held that three key activities by Google were all "highly transformative":

1. Copying of the entire text of books to create a searchable index;
2. Display of snippets from books as part of the search process, to help users identify relevant search results; and
3. Creating the ngram tool to show frequency of words and phrases in the corpus over time.

The court said that the purpose of Google Books "is to make

available significant information about those books." The court held that this purpose is exactly the type of transformative purpose that fair use should enable.

For example: Google Books allows users to track the frequency of references to the United States as a single entity ("the United States is") versus references to the United States in the plural ("the United States are") and how that usage has changed over time.

In this way, TDM does not merely supersede the objective of the original work but "instead add[s] something new, with a further purpose or different character."

Second factor

The court gave fairly cursory treatment to the second factor which requires courts to look at the nature of the copyrighted work, saying that nothing influenced it one way or another with respect to this factor in isolation.

Third factor

For the third factor, the amount and substantiality of the portion used, the court evaluated whether the amount of copying was reasonable in relation to the purpose of the uses. In this case, copying entire works was "literally necessary" to achieve the purpose. If Google copied any less than the totality of the original, the search function would not be reliable. It also noted that Google does not display a copy of the entire work to the public. The snippets of in-copyright text that Google does display are not a competing substitute for the original works.

Fourth factor

Under the fourth factor, the court concluded that snippet display does not give searchers access to effectively competing substitutes and therefore does not threaten rights holders with any significant harm to the value of their copyrights.

The creation of the search index did not make any of the works available to consumers, so it had no direct market effect. The court also considered whether the search index was a "derivative work" that required a license, and concluded it was not. Unlike sequels, film adaptations, and translations, a search index does not "represent the expressive aspects of the original work." The transformative purpose of a search index means it is not covered by copyright's derivative works right.

Conclusions

The Second Circuit held that the Google Books service was a fair use, finding that:

1. "The purpose of Google's copying of the original copyrighted books is to make available significant information about those books," a different function from that of the original books;
2. The amount copied was reasonable to enable the transformative use;
3. The amount revealed to users was tailored to the legitimate transformative purpose and did not threaten to substitute for ordinary consumer purchase; and
4. The unlicensed use would not cause any harm to a traditional, reasonable, or likely future market for the original work.

Another TDM case: *A.V. v. iParadigms*

Let's take a look at another TDM case: A.V. *v. iParadigms*. iParadigms created a plagiarism detection database of student-authored papers. Teachers could submit student papers to iParadigms, which would check its database for matches and, in some cases, iParadigms would retain the paper for use in checking future submissions. A student, "A.V.," brought a lawsuit claiming that iParadigms infringed students' copyrights by using their papers without permission. Citing the internet search engine cases, the Fourth Circuit held that iParadigms' database was transformative because it was used for plagiarism detection, an entirely different purpose from the term papers. Including entire works was appropriate to serve that new purpose. The use, therefore, was fair.

Lessons learned about key TDM uses

So, let's review the lessons we learn from the leading cases on TDM when it comes to three core uses that are likely to occur in most TDM research projects: copying to create a database for TDM analysis, using the data derived from TDM analysis, and publishing data sets used in or derived from TDM research.

1. When creating a database or corpus, the cases tell us TDM analysis is highly transformative and is strongly favored by fair use.
2. The appropriate amount of a work for inclusion in a TDM database is typically the entire work (even millions of entire works), and that's OK.
3. Creating such a database has no market effect, and is not a licensable "derivative work."
4. Derived data does not infringe on the rights of the copyright owner when it consists of unprotectable facts and ideas (like

the ngram tool in Google Books). Copyright in a work does not include a monopoly over facts about that work; facts belong to everyone and are free to share.

5. Publishing a data set (or excerpts from a data set, as in the snippets from Google Books), requires a separate fair use analysis. Before publishing data, TDM researchers should look at the effects of data publication on the traditional market for the works in the dataset. It's especially important to consider the amount that will be released publicly and the security measures in place to prevent the kinds of access that could provide a ready market substitute for consumer access to the work.

Fair use mythbusting

The previous two sections in this chapter have addressed what fair use is and how it interacts with activities associated with text data mining. Unfortunately, there are some persistent misconceptions which circulate about what fair use does and does not allow. This section will debunk some of these common misconceptions so you are better informed about what fair use does and does not allow.

Denied permission requests

Myth: An author cannot rely on fair use if she asks for permission and is denied.

Reality: The truth is, you do not have to ask for permission or even alert a copyright holder when a use of materials is protected by fair use. But if you do inquire about permission, you can still claim fair use if your permission request is refused or ignored. In some cases, courts have found that asking permission and then being

rejected has actually enhanced fair use claims. The Supreme Court has even said that asking for permission may be a good faith effort to avoid litigation.

Using entire works

Myth: An author cannot rely on fair use if he is using an entire copyrighted work.

Reality: The amount of the work copied is just one factor courts consider alongside the other factors, and in particular courts look at whether the amount used was reasonable in light of the purpose of the use. In some situations, courts have found use of an entire work to be fair. This was the case in the Google Books case examined in detail earlier in this chapter: Even though Google copied entire books when making its searchable index, the court found that copying of the entire work was reasonably appropriate to the transformative purpose—indeed, the court said it was "literally necessary" to achieve the purpose.

Using unpublished material

Myth: An author cannot rely on fair use if they are using unpublished material.

Reality: Congress amended the Copyright Act in 1992 to explicitly allow for fair use when using unpublished works after several court decisions suggested that the use of unpublished materials would rarely be fair use.

A court may still consider a work's unpublished status to weigh against fair use when evaluating the "nature of the work" under factor two, but this factor is rarely decisive on its own and courts still must weigh all of the fair use factors, including the purpose

of the use. The purpose of the use may weigh against fair use if the unpublished material is being used in a frivolous or exploitative manner. On the other hand, the purpose of the use may weigh in favor of fair use if the unpublished material transforms the original material and contributes to the public's interest in advancing knowledge.

Using highly creative works

Myth: An author cannot rely on fair use if she is using highly creative copyrighted work.

Reality: While courts do consider whether the copyrighted material used is primarily factual or creative under the second factor, "the nature of the work," this factor is rarely decisive on its own. Courts still must weigh all four factors, again including the "purpose of the use." Where the purpose of the use is transformative and the amount used is reasonable, the second factor rarely affects the final outcome of fair use cases.

Making commercial uses

Myth: An author cannot rely on fair use if he is making a commercial use of a copyrighted work.

Reality: The truth here is that while "noncommercial" uses may be a plus in a fair use analysis, there are no categorical rules: Commercial uses can be fair use, and not all noncommercial uses will be fair use. In fact, some of the important court victories for fair use over the past two decades have been won by defendants whose activities were commercial, including musicians, publishers, and artists who sell their works (sometimes at substantial prices).

Copyright risk analysis: remedies and risk reducers

The material in this chapter has hopefully helped you feel more confident that you can evaluate copyright questions that arise in TDM research. In particular, you will have overcome perhaps the greatest myth in copyright: that fair use is unpredictable and unreliable. Even with this newfound confidence, it is important to remember that very little in life is certain, and we often have to think about risk and uncertainty in order to act with imperfect knowledge about the future.

Weighing risk using expected value

One way to think about the risk involved in doing a particular thing, popular among economists (and lawyers who wish they were economists), is to think about the "expected value" of taking that action: multiply the magnitude of each outcome's good-ness or bad-ness (is the result totally awesome or truly terrible, +$1000 or -$100,000?) by the likelihood of that outcome coming to pass (is there a 20% chance this will happen, or an 80% chance?). The sum of the resulting numbers can give you a sense of the overall risk/reward for any course of action.

When you think this way, a few interesting things emerge: if something is really, truly terrible (or really, totally amazing), even a low likelihood of it happening can meaningfully change the overall value of your choice. This can explain the extreme risk aversion that many folks feel as they approach copyright: they have heard about the insanely high penalties imposed on folks for sharing just a few songs online, so even if it seems unlikely that someone will sue you, if they did, you worry that things could go very very badly.

Luckily, the law has several mechanisms that make this bad

outcome exceedingly unlikely for researchers and research institutions.

Risk reducers in copyright law

The first is section 504(c) of the Copyright Act, which includes a carve-out that favors non-profit, educational institutions, libraries, and archives, and their employees. When these folks have a "good faith belief" that their reproduction of copyrighted works is fair, courts "shall remit" statutory damages. In other words, only actual damages are available in these cases. (And as we saw earlier, these are likely to be low-to-zero in TDM research cases). Those hefty penalties you may have heard about in file sharing cases are simply not on the table when section 504(c) applies.

Note, however, that this only applies to the reproduction right, which is just one of the several statutory rights in the law. Distribution (sharing copies) and adaptation (creating derivative works) are not covered. So if you are relying on section 504(c) in your risk calculus, think carefully about whether everything you are doing in your project will be shielded.

State sovereign immunity and qualified immunity protect state institutions and their employees against money damages in most cases. The Supreme Court reaffirmed the application of state sovereign immunity to copyright cases in its 2020 opinion in *Allen v. Cooper*. These immunities will protect state institutions and employees from money damages in most copyright cases, but the court can still order injunctions (judicial commands that the losing party do or refrain from doing particular things). The key remaining risk for state institutions may be that all the time and effort invested in a project could be lost if a court were to find the project infringing and order the resulting data destroyed or made inaccessible pending rightsholder permission. And of course private institutions (even non-profits) are not covered by these immunities. Also

important: Congress can waive these protections in cases of willful infringement, if it drafts its legislation carefully. At the time of this writing, the United States Copyright Office is drafting a report on the feasibility and desirability of new legislation to do so.

Another limitation on remedies, which may be helpful in working with archival materials, is that timely registration is required in order to seek statutory damages. While most commercial works (e.g., novels, academic journals) are likely to be registered, other classes of works may be much less so. Amateur works such as snapshots, ephemera and advertising material, and unpublished and archival works all may be less likely to be registered. If your corpus doesn't include commercially published works, you may face a much lower likelihood of statutory damages.

Risk reducing strategies

Notice and takedown-style policies can give concerned or upset rights holders a channel for expressing their concern, and can give you an opportunity to accommodate them without anyone ending up in court. Hot tip, though: you don't have to promise to take things down, and it can actually help shape expectations if you frame your notice mechanism in terms that are less negative, like "We welcome you to contact us to ask a question or share information about this research collection." Anecdotal evidence suggests that responses to these kinds of prompts are more likely to lead to amicable resolutions.

Reasonable attribution is really important to some authors and rightsholders, and can go a long way to avoiding temper flare-ups. Of course, some won't be placated by this, but surprisingly many folks who raise complaints about content reuse are (or would have been) satisfied by just getting the credit they felt they deserved.

Plaintiffs face risks, too. A recent study found that the average copyright case costs $300k to litigate to a verdict. If a plaintiff loses,

courts have the discretion to force them to pay the defendant's court costs and attorney fees, if the court finds the suit was frivolous or unwarranted. (This is called "fee shifting.") And the Streisand Effect can mean bad press for a copyright holder who sues sympathetic defendants, like libraries and researchers.

Remember (and weigh) the risk of inaction

Any particular research project may pose a variety of risks, but you wouldn't consider embarking on the project if it didn't present an opportunity to do something good. Too often in academia we treat all risk as unacceptable, and ignore the upside value of fulfilling our mission, or, the downside of failing to pursue our mission. The rational course is not to insist on zero risk; it's to consider both the upsides and the downsides of action and inaction, then make choices that are more likely to do good than harm. As you consider a project through the lens of risk (and develop strategies to mitigate it), don't lose sight of the value that drew you to the project in the first place, or the loss associated with abandoning the project due to excess caution.

Copyright use case revisited

Let's return to our case study we outlined at the very beginning—gathering together a dataset of materials of varying copyright status, and allowing users to browse through works in this collection according to the geographical places that are mentioned within them. In this case, a user has searched for "Paris," which brings up a selection of results where "Paris" is mentioned in text, and that "Paris" has been disambiguated to refer to Paris, France, and not Paris, Texas.

The works that comprise this collection have mixed copyright status—we might be relatively confident that works published in 1925 or earlier are in the public domain, while those published afterward are more likely to still be subject to copyright (unless those authors failed to comply with formalities—such as notice—during that time period). This collection also contains works of fiction—so not just purely factual content, but "highly creative works."

We can see this use case as being analogous to that of Google Books—we're performing a transformation of the original (perhaps copyrighted) text in order to present information that's not directly accessible in any single work (here, using geography as an organizing principle to index the entire collection). We use the entire work for the index that we are creating here, but only present small snippets from the original work (single sentences) to users.

The more complex component of this use case comes in the goal of annotating selections from this dataset (having people mark where in the text a place is mentioned), and then publishing those annotations along with the original texts. This requires its own fair use determination separate from that of the indexing-and-visualization use case; while in the former use case only snippets are published, here we want to publish larger samples of text—perhaps a passage of 500 or 1,000 words from a single novel.

The first question to ask is: do we need to publish anything from the original texts at all? Other alternatives may exist. One possibility would be to only publish the annotations (not linked to the original texts), along with a description of the process by which another user could map those annotations back onto the original text—for example, publishing an annotation that says that word 171 on page 37 in the original work is a "place." If another user has access to the same copy of the original work, and can follow your process to align an annotation with that work, then publishing the original work isn't necessary.

In many cases, however, users simply don't have access to exactly the same copy of the original text that would make reproducibility

possible, so let's consider that the annotations we create need to be published alongside their original work. What do we need to consider when making decisions about the scope of this project? As we've seen, there are a number of factors that determine whether this specific case study qualifies as an instance of fair use—so without making a recommendation for this case, we can outline the different factors that would go into a determination. First is the purpose and character of use—in this case, we could reasonably argue that the annotations that we publish alongside the original works are adding new meaning and expression to the original work; we're not simply republishing parts of the original works alone, but only to support the human judgments of place names we've layered on top of them. Second is the nature of the copyrighted work—many of the works in this case study are works of fiction, and so constitute creative works—which (as we've seen) would be more likely to weigh against fair use. Third is the amount and substantiality of the samples we are considering publishing—how much can the samples we publish be seen as a substitute for the original, copyrighted work? While the use of entire works may qualify for fair use, one main consideration is whether the amount of the work used is appropriate for the use—and for the task of enabling reproducibility of NER models, a smaller sample (e.g., publishing only 1% of a 100,000-word novel) may be reasonable. And finally, what is the effect of publishing these samples on the market for the original work? We might imagine that publishing a large amount of a contemporary popular work like *Harry Potter* may impact its sales, while publishing smaller samples that don't get at the heart of work would not.

So these are some of the factors to weigh when deciding on the design of this project—what data sources to use, and how to best use them to help realize the goals of the project. There is risk in all decisions—for this particular project, we need to weigh the risks of using texts in copyright with the risks of not using them—in this particular case, using texts published after 1925 in a reasonable way enlarges the pool of sources beyond the primarily white and male

authors represented in texts published before then. But knowing the landscape of copyright should provide some strategies for weighing and deciding upon these risks yourself.

2. International and cross-border copyright

SEAN FLYNN AND MATTHEW SAG

Introduction

Suppose that you are managing a collection of 1970s environmental catastrophe themed fiction and making it available for text data mining research in the United States. Here are some basic questions to think about:

- Should you allow foreign researchers to query the corpus?
- Should you accept new additions to the collection from an overseas library?
- Are you in a position to send a copy of the corpus to overseas researchers?
- Does it matter if those researchers are housed in a university, a corporate sponsored think tank, or a for-profit corporation?

These questions illustrate some of the issues raised by text data mining research in an international or cross-border environment.

In the materials that follow, we are going to introduce some of the conceptual building blocks that you will need to be able to understand and address these kinds of issues. Our aim isn't to make you experts in comparative and international copyright law, but we hope to give you enough information so that you can identify potential areas of concern and understand how to structure cross-border collaboration in TDM research without taking on unnecessary risks.

The relation between domestic and international copyright law

The first step in appreciating the kinds of international and cross-border copyright law issues that might be relevant to text data mining research is understanding the relationship between domestic and international copyright law.

Copyright law is harmonized across the globe by virtue of various international agreements. The most relevant international copyright treaties are the Berne Convention and the World Trade Organization Agreement on Trade Related Aspects of Intellectual Property Rights (or the TRIPs Agreement, for short). These agreements establish minimum standards for copyright protection, that more or less every country in the world has agreed to adopt as part of their domestic copyright law.

There is a lot of agreement about many aspects of copyright law around the world, but that agreement is often at a high level of generality. Digging a bit deeper, we find meaningful diversity in how countries choose to implement their international copyright obligations.

As a result, particularly in relation to the issues surrounding text data mining research, copyright law can vary significantly from one country to the next.

So, although international agreements provide important background principles, the law we generally need to focus on is the domestic copyright law of individual countries.

That sounds simple enough, but we have to complicate this story slightly with respect to the European Union. Copyright law in the EU is harmonized by a series of EU directives. These directives must be implemented in the national law of the various member states, but in many cases the EU directives also have direct effect. This feature of European law explains why in some cases you will hear us talk about European copyright law as though it was a single consistent body of law—sometimes this is just a helpful

generalization—and yet in other cases we focus in more detail on the laws of individual countries.

Copyright protection and limitations and exceptions for TDM research

Here we want to go over the basic steps of analysis to determine whether you have a copyright issue in an international text and data mining research project. Assume for the moment that you are trying to decide whether you can locate a particular research activity in another country in which you have a research partner.

I assume here that you might want to undertake the following activities in a TDM project:

- Reproducing whole works to create a database or corpus;
- Sharing a database with other researchers (either in the country or across borders);
- Finding and reporting facts through use of the database;
- Quoting the materials mined for validation and illustration.

One or all of these activities might take place in another country or between researchers in other countries. This section will focus on what kind of laws you can expect to find in different countries.

Exercise: Keep track of what you learn in your own copy of the TDM Activities Worksheet. To use the worksheet, make a copy of it and then add your information directly into your copy.

Scope of protection

Our goal here is to give you information about what aspects of

copyright law are near universal and what the main variations are so you can do what we law profs call issue spotting. That is, be able to spot where there is likely to be or likely not be a real legal issue that you might need to dig more deeply into. To answer a specific question with regard to a specific country you may need to dig a little deeper into the individual context.

As we covered with respect to US law, there are two basic stages to any copyright analysis. First you look to whether the work and intended activity are within the scope of copyright protection. Second, if the work and activity fall within the scope of protection, then you look to whether a limitation or exception to the exclusive rights none-the-less permits the activity.

Is the work protected?

By now you probably all realize that working with resources in the public domain can resolve all of your copyright concerns. However, determining what is in the public domain may be somewhat difficult.

Definition of a protected work

The definition of protected works in every copyright law is incredibly broad, in part because international law requires a broad definition of protected works.

The Berne Convention defines a protected work as "every production in the literary, scientific and artistic domain, whatever may be the mode or form of its expression." The Convention gives an illustrative list:

- books,
- dramatic or choreographic or cinematographic works,

- musical compositions,
- drawing, painting, architecture, sculpture,
- photography;
- applied art;
- maps

What about government works?

Unfortunately, you cannot assume that a work is freely usable because it is a government work – even a law.

The Berne Convention, allows, but does not require, an exemption for official texts, such as laws. The US exempts these texts from copyright. But some countries—including the UK and many commonwealth countries—protect such works.

What about old works?

The Berne Convention states a minimum required term of protection of life of the author plus 50 years. But countries can protect longer, and many do.

Most of the countries in Africa and Asia protect copyright for life plus 50 years, or sometimes less. (Not all countries have signed on to the Berne limits.) And Berne allows countries to apply lower terms to photographs—as few as 25 years.

But about half the countries in the world protects works for longer than life plus 50 years. Mexico tops that list with terms of life plus 100 years.

The result is that some older works may be subject to copyright in the U.S. but in the public domain overseas, and vice versa.

Is the Activity Protected?

If you conclude – or prefer for simplicity to assume – that a materials you want to use is a protected work, then the next question you will have is whether your use of that work is subject to an exclusive right of the copyright holder.

There is a fair amount of uniformity on this question.

Berne requires that copyright laws protect against reproduction "in any manner or form."

Laws normally require that a substantial amount of the work be copied to constitute a reproduction. But there are courts that have held that as few as 11 words from a work can constitute a substantial reproduction (EU).

Countries have generally implemented the reproduction right broadly. German law, for example, excludes all copies by whatever method in whatever quantity.

So here, think about whether any or all of the activities you might want to undertake for TDM involve a reproduction of the work in any method and in any quantity.

There are more rights

The reproduction right—which is the most central and oldest right in copyright—is certainly incredibly broad. But international laws have expanded on the definition over time, adding new exclusive rights for activities that may not involve a technical reproduction at all.

First Berne requires protection against the translation or adaptation of works. Some prominent commenters have opined that translation and adaptation rights may apply not only between

human languages, but also "translations from one computer language to another."[1]

And later treaties require that countries protect the right to "distribute," "communicate," or "make available" a work.

It is generally accepted that a distribution can take place when one transfers the work to another person, whether that be a hard copy or sharing a file.

Exhaustion

Now, some transfers are exempted from the distribution right. Copyright's exclusive right to control the distribution of a work within the same country is "exhausted"—that is, the right ceases to bind – after the first sale of that work. This is why used book stores can occur and why you can gift a book to another person. But in some countries that exhaustion does not apply outside of the country where the first sale occurs. And in very few countries does the exhaustion rule apply to a digital copy.

Also note that making available rights can be infringed through allowing members of the public to access works from a place and at a time individually chosen by them. Can that be sharing a link to a dropbox file? What if you allow any researcher—the broad "public" in other words—to use your research corpus and thereby "access" the works you have made a copy of?

If we end here, the copyright environment looks pretty daunting. There may be limiting interpretations of these concepts in domestic laws or court decisions. But at least on their surface, you may be able to conclude that all of the uses of works we discussed above,

1. Paul Goldstein and P. Bernt Hugenholtz, International Copyright: Principles, Law and Practice, 299 (4th ed. 2019).

and maybe some more you have since thought of, are subject to copyright law on their face. Thus, for a great many text and data mining project activities, you are going to need help from the next section—limitations and exceptions.

Universal exceptions and limitations

Recall the purpose of copyright. Copyright exists to prevent competing uses of protected works. We sometimes think of these as public uses. Uses that can substitute for the original work in a way that harms the market for the work.

Under this general theory, uses of a work that cannot substitute for the work in the market—e.g. because they are confined only to a use in the home, like copying your CD to your hard drive—should not be protected. Why? Because that use does not share the work with anyone in a way that can displace a use.

In the last section we showed that the definitions of exclusive right appear to protect many uses such as private, at home, use. But that use is lawful in probably every country in the world. Why? Because of the presence of exceptions to copyright.

Some of the most important limitations and exceptions to copyright are required by international copyright agreements, such as the Berne Convention and TRIPs. We refer to these as "universal."

Exclusion of facts

The first important exception required by international law—and often via freedom of expression rights—is the exclusion of facts. All copyright laws around the world apply only to original expression, not to the facts conveyed by that expression. The Berne Convention requires this distinction – expressly excluding protection of "news

of the day" and "miscellaneous facts having the character of mere items of press information."[2]

The WTO TRIPS Agreement expands on this aspect, requiring what is often referred to as the "idea-expression distinction." "Copyright protection shall extend to expressions and not to ideas, procedures, methods of operation or mathematical concepts as such."[3]

A basic example of the difference between facts and expression is an article about a sports tournament and the score. The score may be included in the article and may be where you got that information. The newspaper has an exclusive right over the article—the original expression of the sports writer describing the event. But the score is a fact. You can use the fact freely, even if you can't copy the article.

The problem of course arises in how you access that fact without copying the expression. You can read the article. We all admit that. But can you mine it? If you have to copy the work to mine it for its facts you may need more.

Quotation

International law also requires the right of quotation.[4] Berne does

2. Berne Art 2(8), "The protection of this Convention shall not apply to news of the day or to having the character of mere items of press information."
3. TRIPs Art 9(2).
4. Berne Convention Art. 10(1) ("It shall be permissible to make quotations from a work which has already been lawfully made available to the public, provided that their making is compatible with fair practice, and their extent

not go into a lot of detail about what the quotation right means. But we can generally assume that it means only the use of an excerpt of the work, not the whole work. So this exception does not likely give researchers a right to make whole copies of works to create a database to be mined. But it may be useful in communicating and illustrating the results of such research.

Some national copyright laws authorize quotation for any purpose;[5] some explicitly exempt research purposes.[6] The most limited quotation rights require criticism or review of the work quoted. Pause there and ask yourself—and note in your worksheet—whether a quotation exception limited to "criticism and review of the work quoted" would be sufficient to authorize the quotes you want to make for publication and validation purposes of your project.

Review your worksheet now and fill out as much of the third

does not exceed that justified by the purpose, including quotations from newspaper articles and periodicals in the form of press summaries.').

5. See South Africa
6. See Mexico, Federal Law on Copyright (consolidated text published in the Official Journal of the Federation on June 15, 2018), Art. 148 ("Literary and artistic works that have already been disclosed may only be used in the following cases without the consent of the owner of the economic rights and without remuneration, provided that the normal exploitation of the work is not adversely affected thereby and provided also that the source is invariably mentioned and that no alteration is made to the work: . . . III. Reproduce portions of the work, for critical and scientific, literary or artistic research").

column you can through application of these universal exceptions to copyright protection. What do you have left? You will need to fill in the empty spaces in your worksheet in the next session analyzing specific laws in specific counties. Here the law gets a little more complicated.

National approaches to copyright limitations and exceptions

You should have concluded that there are some activities that TDM researchers need to do that should be permitted in every country by virtue of the idea/expression dichotomy and the right of quotation.

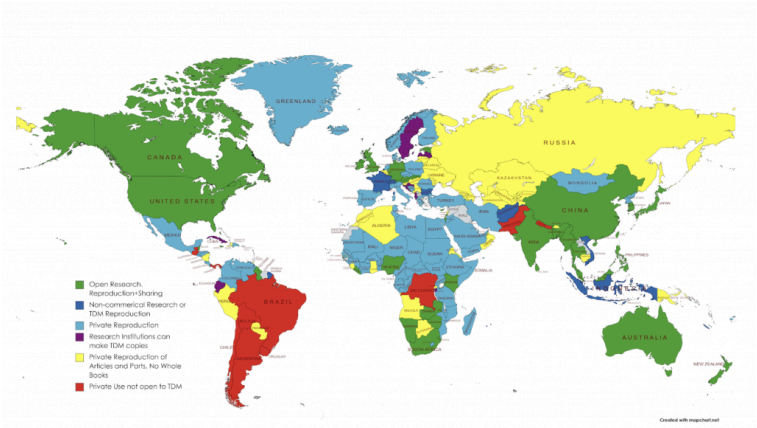
But these universal exceptions are not sufficient to authorize all of the activities that TDM researchers need to do. This may be true even where that activity does not appear to compromise copyright law's core objective of prohibiting the making of copies that can substitute for the work in the market. Unfortunately for us, the manner in which countries protect the interests of users in making non-competitive uses of works varies significantly.

Beyond the mandatory exceptions and limitations, international law leaves countries largely free to craft exceptions for uses that do not harm the interests of copyright protection.[7] The so-called

7. International copyright treaties all contain a basic enabling and limiting principle that "It shall be a matter for legislation in the countries of the Union to permit the reproduction of such works in certain special cases, provided that such reproduction does not conflict with a normal exploitation of the work and does not unreasonably prejudice the legitimate interests of the

three-step test in Berne allows countries to permit any use that "does not conflict with a normal exploitation of the work and does not unreasonably prejudice the legitimate interests of the author." That should sound a lot like the fair use factors you learned about previously. The trick is that some, but not all, countries take full advantage of this flexibility to exempt non-competitive uses from copyright control.

Let's start with the conclusion. A map of the world based on whether you can reproduce and share copyrighted works for sole purpose of research—without sharing those works to the general public—looks like this:



Comparative Copyright Law on Research Exceptions, Sean Flynn, Andres Isquierdo, Mike Palmedo, PIJIP (2020)

I say "law on the books" meaning the copyright statute itself. In application, there may other rights—such as human rights to receive and impart information—that may make the rigid application of the

author." Berne Convention Art 9(2); accord TRIPS Agreement Art. 13.

law in these countries to ban data mining unconstitutional. This seems a likely outcome in Brazil, for example.

And so it appears to be the case that in most countries of the world the law appears open to the interpretation that you could make the necessary copies needed to create a database for a "private" TDM project. But also in most of the world there is a lack of a clear right to share those copies with another researcher.

In the next part we will describe in more depth what the provisions of the law look like that we are interpreting here.

Open and General Exceptions

An exception can be general or specific; open or closed—on a continuum.

By general I mean that a single exception applies one balancing test—e.g. to fairness—to a group of different purposes. Specific exceptions apply to only one (or sometimes a couple of related) purpose of use.

By open I mean that the exception applies to the full scope of protection. It covers all rights, all works, and by any user.

A fully open general exception applies a single balancing test to a use of any work, by any user, for any purpose. Fair use is such an exception. But it is not the only one. And a fully open research exception can be just as useful for a TDM researcher than a fully open general exception.

I am going to use this map to go through the different kinds of exceptions that could authorize the making or sharing of TDM databases.

The general and open exceptions for research are labeled in Green. In those countries, the copyright exceptions on the books are phrased broadly enough to permit both the making, and sharing between researchers, of a TDM database.

Let me start with the fair use and fair dealing countries.

Fair use and fair dealing

The US fair use right is an open general exception. It applies one basic fairness to assess the permissibility of any utilization of a work that implicates any exclusive right, by any user, of any work, for any purpose.

General exceptions are most common in, but not exclusive to, countries from the common law tradition evolving from the United Kingdom. Such exceptions often provide a general defense for "fair use" or "fair dealing."

I want to address what I see as a common misconception about the difference between fair use and fair dealing. The misconception is that fair use is a more open right than fair dealing. That is not universally true.

In the US and some other countries, the term for the utilization permitted by the exception is "fair use." In the UK and many other commonwealth countries, the historical term used for a permitted utilization is "fair dealing." Almost always the word "use" or "dealing" mean to apply to the exercise of any exclusive right.[8]

8. An exception is Malaysia, where a fair dealing right is open to any use, by virtue of inclusion of the word "including" before the list of authorized purposes, but it only applies to reproduction: Malaysia Copyright Act 1987 (2012) Article 9. Copyright in published editions of works . . . (4) Reproduction of the typographical arrangement of a published edition for any purpose including research, private study, criticism, review or the reporting of news or current events does not infringe the copyright subsisting by virtue of this section if such reproduction is compatible with fair dealing

Ireland

Copyright and Related Rights Act, 2000

Article 50.

(1) Fair dealing with a literary, dramatic, musical or artistic work, sound recording, film, broadcast, cable programme, or non- electronic original database, for the purposes of research or private study, shall not infringe any copyright in the work.

Zambia

The Copyright and Performance Rights Act, 1994

Article 21. Acts which do not constitute infringements

. . .

(a) fair dealing with a work for private study or for the purposes of research done by an individual for his personal purposes, otherwise than for profit.

Notice that "use" and "dealing" mean the same thing. They both apply to any type of utilization of the work, that is—a utilization that implicates any exclusive right of the copyright holder.

In this example, the Australian fair dealing right is subject to a closed list of purposes and the US fair use right has an open list. The magic words to look for here are "such as."

But is not true that "fair use" rights are open and fair dealing rights are closed. Look at these two examples.

The Uganda fair use right is not open. And the Malaysia fair dealing right is not closed.

This distinction is unlikely to matter here since most fair use and fair dealing rights explicitly apply to "research" purposes.

Other general exceptions

There are also general exceptions that are not fair use or fair dealing rights. Indonesia has a general exception for any "use" of a work for research or other purposes.

Indonesia

Law of the Republic of Indonesia No. 28 of September 16, 2014

Article 44.

(1) The use, retrieval, duplication, and amendment of a copyright work or a related right in whole or in part is not considered as a violation of copyright if the source is stated or stated in full for the purposes of:

1. education, research, writing scientific papers, preparing reports, writing criticisms or reviewing a problem without harming the reasonable interests of the Creator or Copyright Holder

Thailand simply makes the entire scope of the Berne three-step test a general exception.

Thailand

Section 32. Exceptions to Infringement of Copyright

An act against a copyright work under this Act of another person which does not conflict with normal exploitation of the copyright work by the owner of copyright and does not unreasonably prejudice the legitimate rights of the owner of copyright shall not be deemed an infringement of copyright.[9]

The Republic of Korea combines the Thailand approach to the three-step test with the fair use multi-factor test:

Republic of Korea

9. Accord Namibia Copyright Act, Art, 16 ("General exceptions in respect of reproduction of works: In addition to reproductions permitted in terms of this Act reproduction of a work shall also be permitted in such circumstances as are prescribed, but in such a manner that the reproduction is not in conflict with a normal exploitation of the work and is not unreasonably prejudicial to the legitimate interests of the owner of the copyright.").

Copyright Act (Act No. 432 of January 28, 1957, as amended up to Act No. 14634 of March 21, 2017)

Article 35-3. (Fair Use of Works, etc.)

(1) Except as provided in Articles 23 through 35-2 and 101-3 through 101-5, where a person does not unreasonably prejudice an author's legitimate interest without conflicting with the normal exploitation of works, he/she may use such works.

(2) In determining whether an act of using works, etc. falls under paragraph (1), the following shall be considered:

1. Purposes and characters of use including whether such use is for or not-for nonprofit;
2. Types and natures of works, etc.;
3. Amount and substantiality of portion used in relation to the whole works, etc.;
4. Effect of the use of works, etc. on the current or potential market for or value of such work etc.

Open research exceptions

I have also labeled in green specific exceptions for research that are sufficiently open to apply to the use of all works and apply to both reproduction and sharing rights that we are most concerned with.

Some research rights are open to application to all exclusive rights. E.g.

Liechtenstein

Law on Copyright and Neighboring Rights (Copyright Law) (version as of 1 June 2016)

Article 22. Privileged uses of the work

1) Published works may be used for special purposes. A special purpose is:

1. a) any use of the work in the personal sphere and in the circle of persons who are closely related, such as relatives or friends;
2. b) the use of the work for illustration in class or for scientific research insofar as this is justified for the pursuit of non-commercial purposes and if possible the source and the name of the author are given;

c) the reproduction of the work on paper or a similar medium by means of photomechanical processes or other processes with a similar effect for educational purposes, for scientific research or for internal information and documentation in companies, public administrations, institutes, commissions and similar institutions;

1. d) digital reproduction for educational purposes and for scientific research without any direct or indirect economic or commercial purpose.

Some of the specific exceptions for data mining are also open framed. Japan applies to any "exploitation," including for data analysis.

Japan

Article 30-4. Exploitations not for enjoying the ideas or emotions expressed in a work

It is permissible to exploit work, in any way and to the extent considered necessary, in any of the following cases or other cases where such exploitation is not for enjoying or causing another person to enjoy the ideas or emotions expressed in such work; provided, however that this does not apply if the exploitation would unreasonably prejudice the interests of the copyright owner in light of the natures and purposes of such work, as well as the circumstances of such exploitation:

(i) exploitation for using the work in experiments for the development or practical realization of technologies concerning the recording of sounds and visuals or other exploitations of such work;

(ii) exploitation for using the work in a data analysis (meaning the extraction, comparison, classification, or other statistical analysis of language, sound, or image data, or other elements of which a large number of works or a large volume of data is composed; the same applies in Article 47-5, paragraph (1), item (ii));

(iii) in addition to the cases set forth in the preceding two items, exploitation for using the work in the course of computer data processing or otherwise that does not involve perceiving the expressions in such work through the human sense (in regard of works of computer programming, the execution of such work on a computer shall be excluded).

Other research exceptions, although not open to every "use," nonetheless specifically make provision for both reproduction and sharing. E.g.

Luxembourg

Law of April 18, 2004, amending Law of April 18, 2001 on Copyright, Neighboring Rights and the Databases

Article 10.

When the work has been lawfully made available to the public, the author may not prohibit:

…

2. The reproduction and communication to the public of works by way of illustration of teaching or scientific research and to the extent justified by the aim to be achieved and provided that such use is in accordance with good practice.

Germany makes similar provision in its recent law focused specifically on authorizing text and data mining:

Germany

Section 60d. Text and data mining

(1) In order to enable the automatic analysis of large numbers of works (source material) for scientific research, it shall be permissible:

1. to reproduce the source material, including automatically and systematically, in order to create, particularly by means of normalisation, structuring and categorisation, a corpus which can be analysed and
2. to make the corpus available to the public for a specifically limited circle of persons for their joint scientific research, as well as to individual third persons for the purpose of monitoring the quality of scientific research.

As we discuss below, most current TDM laws in the EU do not make this provision for sharing and the EU directive does not require it.

We have labeled all the laws in this section GREEN. These are laws that, on their face at least, appear to authorize reproduction and limited sharing between researchers of all works by any user for a research purpose.

Non-expressive uses as fair practice

The work in all these exceptions is done in the balancing test used to determine if a particular use is permitted. Sometimes there is a multi-factor test like US fair use. Sometimes it is a single test like "fair practice." In any case, the balancing factor gives an opportunity for calibration of exclusive rights to promote copyright's purposes. A central question in each will be whether the use unfairly competes with the original.

If you are making a copy of works into a private database that will not be released to the public in any way, then the test should

be readily passed. This was the holding in US courts in the *Google Books*, *HathiTrust* and other cases.

Reproduction for research

Now we move to the countries I have marked in Blue in the map. The difference between from the last category is that blue countries only authorize reproduction, not distribution or communication rights. As a result, whether a researcher can copy and transfer a whole database to another researcher in these countries is either very unclear or clearly prohibited.

The simplest of these exceptions provide exceptions for reproduction for research. The key here is that it only allows reproduction, not distributions or communications.[10]

10. Malaysia has an exception that applies only to reproduction, although interestingly it is open to any purpose by virtue of inclusion of the word "including." This is an exception to the general rule that a "dealing" is the same as a "use." Malaysia Copyright Act 1987 (2012) Article 9. Copyright in published editions of works (1) Copyright shall subsist, subject to the provisions of this Act, in every published edition of any one or more literary, artistic or musical work in the case of which either- . . . (4) Reproduction of the typographical arrangement of a published edition for any purpose including research, private study, criticism, review or the reporting of news or current events does not infringe

Morocco

Law No. 2-00 on Copyright and Related Rights (2000))

Article 54. Free Uses (Research)

Notwithstanding the provisions of Articles 50 to 53, the following acts shall be permitted without the authorization of the successors in title mentioned in these articles and without the payment of a fee:

…

(b) reproduction solely for the purposes of scientific research;

Maldives

Section 29.

Section 25,26,27 and 28 shall not apply where the acts referred to in those sections are related to:

….

(b) reproduction solely for scientific research;

Sometimes the research right is included within in a private use or private study right, as in Somoa. What were are looking for in a

the copyright subsisting by virtue of this section if such reproduction is compatible with fair dealing

connector like "or" that makes clear the research right is separate from the private use right.[11]

Samoa

Copyright Act 1998 (as consolidated in 2011)

Section 8A. Reproduction for purposes of research or private study

(l) Despite section 6(1)(a), but subject to subsection (2), a person reproducing a work for the purposes of research or private study is not to be regarded as infringing any of the copyright in that work.

(2) Despite subsection (1), if a person reproducing the work knows or has reason to believe that it will result in copies of substantially the same material being provided to more than one person at substantially the same time,

11. The research right may also be combined with educational rights, as in Vietnam: Vietnam Law No. 50/2005/QH11 of November 29, 2005, on Intellectual Property Article 25. Cases of use of published works where permission and payment of royalties and/or remunerations are not required: 1. Cases of use of published works where permission or payment of royalties and/or remunerations is not required include: a. Duplication of works by authors for scientific research or teaching purpose;

that person will not be regarded as reproducing the work for the purposes of subsection (1).

As we will discuss further below, the EU directive on text and data mining only requires that EU countries have an exception for reproduction, not for distributions and communications even between researchers.[12]

European Union (EU)

12. For an example of an EU domestic law that is restricted to reproductions, see France Intellectual Property Code (amended by Act No. 2016-925 of July 7, 2016) Article L122-5. 1. Private/personal use When the work has been disclosed, the author cannot prohibit: …. 10°. Copies and digital reproductions made from a lawful source for the purposes of mining text and data included in or associated with scientific publications, for public research purposes, excluding all commercial purposes. A decree fixes the conditions under which the exploration of texts and data is implemented, as well as the methods of conservation and communication of the files produced at the end of the research activities for which they were produced; these files constitute research data;

Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market (DSM Directive)

Article 3. Text and data mining for the purposes of scientific research

1. Member States shall provide for an exception to the rights provided for in Article 5(a) and Article 7(1) of Directive 96/9/EC, Article 2 of Directive 2001/29/EC, and Article 15(1) of this Directive for reproductions and extractions made by research organisations and cultural heritage institutions in order to carry out, for the purposes of scientific research, text and data mining of works or other subject matter to which they have lawful access.

Private reproduction

Another category of exception that may be useful in authorizing TDM research activities are private use rights.

These rights generally allow researchers and others to make a copy (often just one) of a work, including for a research purpose. Often these rights apply to making copies of whole works. Where broadly phrased, private use rights may thus permit the making of a database for TDM. E.g.

Malawi

Copyright Act, 2016 (Act No. 26 of 2016), https://wipolex.wipo.int/en/text/446811

Article 38. (Personal or Private)

The reproduction, translation, adaptation, arrangement or other transformation of a work exclusively for the user's own personal or private use of a work which has already been lawfully made available to the public shall be permitted: Provided that it is made on the basis of a representation that the authorized under this Act at the initiative of the user and not for the purpose of gain and only in single copies.

Azerbaijan

Article 17. Free Use of Works and Phonograms for Personal Purposes

1. It shall be permissible to reproduce one copy of works previously published lawfully for personal purposes without the consent of author or other copyright owner and without payment of author's remuneration, on nonprofit base.

There are several common restrictions in private use rights. First, as in the example above, often these rights contain express

prohibitions of commercial or for-profit use. Even where such express limitations are not provided, they may be implied by the definition of "private."

Similarly, the definition of "private" is often expressly limited to a natural or physical person. A corporation, university or research institution cannot normally rely on a private use exception to create a TDM database unless there is a separate right of such institutions.

Private use rights do not generally extend to sharing of the copied work. The rights may limit sharing by extending only to a reproduction – not a distribution or communication of the work. Or sometimes the rights include an internal restriction making clear that sharing is not permitted.[14]

Finally, many private use rights often explicitly forbid making copies of a "database," and sometimes specifically an electronic database. We already assume that private use rights are not sufficient to authorize the copying of a TDM database to share with other researchers. This is sometimes very explicit. E.g.

Burkina Faso

Law No. 032-99/AN of December 22, 1999, on the Protection of Literary and Artistic Property

Article 21. Private/personal use

Where a work has been legally disclosed, the author may not prohibit: ...

– copies or reproductions reserved strictly for the private use of the copier and not intended for collective use, with the exception of: ... the total or substantial reproduction of databases;

Thus, in the best case, private use rights may be sufficient in many

countries to authorize an individual researcher to create a corpus of works for TDM activities. But they are not likely to be sufficient to authorize the sharing of the database between researchers in ways that require reproduction of the database itself.

Restricted private use rights (yellow)

Some private use rights are further restricted in ways that would allow the creation of only some kinds of TDM databases. We have flagged these countries in yellow.

The most prominent example here is the relatively frequent restriction from using private use rights to copy a whole book. E.g.

Russian Federation

Civil Code of the Russian Federation (Part Four, as amended up to Federal Law No. 549-FL of December 27, 2018, and Federal Law No. 177-FL of July 18, 2019)

Article 1273. Free Reproduction for Personal Purposes

1. A citizen may reproduce, if necessary and exclusively for personal purposes a legally promulgated work without the author's or other right holder's consent and without paying a fee, except for the following:

…

2) the reproduction of databases or significant parts thereof, except as provided for by Article 1280 of this Code;

4) the reproduction of books (in full) and musical notation texts (Article 1275), that is the facsimile reproduction with the help of technical facilities for the purposes other than publication;

Excerpts only (red)

Finally, some private use rights are not useful for TDM projects at all because they are limited to the use of excerpts, and therefore function in reality as quotation rights.

My favorite example here is from Argentina, which has the most restrictive copyright exceptions I have ever seen. There is just one exception to copyright and it is only for quotation.

Argentina

Law No. 11.723 of September 28, 1933, on Legal Intellectual Property Regime (Copyright Law, as amended up to Law No. 26.570 of November 25, 2009)

Article 10. Any person may publish, for didactic or scientific purposes, comments, criticisms or notes referring to intellectual works, including up to 1,000 words for literary or scientific works, or eight bars in musical works and, in all cases, only the parts of the text essential for that purpose.

This provision shall cover educational and teaching

works, collections, anthologies and other similar works.

Where inclusions from works by other people are the main part of the new work, the courts may fix, on an equitable basis and in summary judgment, the proportional amount to which holders of the rights in the works included are entitled.

So there you have the world.

There are a number of countries we cannot find or translate the law. They are left in grey.

The number of countries where you cannot make a TDM database at all is relatively small, but clustered in some huge and important countries to our South.

On the other hand, the number of countries where you can both make and share a TDM databases with other researchers is also relatively small, although it includes some very large and important places.

The question for the next section is how to approach the matter when you are in a green country but want to do a project with a colleague in a blue, tallow or red one. Does the law there restrain you here?

Library and research institution exceptions

One final source of copyright exception that may extend to the creation of a text and data mining database is in exceptions for libraries and research institutions. Many national copyright laws contain special exceptions for uses by libraries which may contain rights to make copies for third party research projects. It's possible that such exceptions could be helpful in relation to text data mining

research, but again, we would have to look at these country-by-country to say much more than this.

Temporary reproductions

A significant number of more recently amended national copyright laws allow for temporary reproductions to carry out technical processes. Depending on the technical process being utilized, a limited right to make temporary reproductions may be enough to engage in text data mining research. Storing copyrighted works in a database is not likely to qualify as a temporary reproduction. But an exemption for temporary reproduction should apply where copyrighted works are stored briefly (briefly as in seconds, not weeks), analyzed to derive relevant metadata and then deleted.

Specific exceptions for TDM research

One reason why copyright law treats text data mining research differently in different countries is that some jurisdictions have amended their copyright laws with text data mining in mind, whereas most have not. But even where legislative accommodations have been made, the text and intent of the relevant provisions varies.

Only a handful of countries have specific exceptions for TDM research. In 2009, Japan became the first country to adopt an express exemption for text data mining. Between 2014 and 2018, the

United Kingdom,[13] France,[14] Estonia,[15] and Germany[16] also enacted laws specific to text data mining. In 2019, the European Union adopted the Digital Single Market Directive which includes two separate provisions meant to enable TDM research under different conditions.

None of these laws are exactly the same, and they probably all differ from the legal position in the United States to some degree.

Because of this lack of uniformity, even cross-border research collaborations between jurisdictions that both support TDM research might run into obstacles.

To give you a sense of what these obstacles might be, we are going to summarize some of the key points of differentiation between the law as we understand it in the United States and those jurisdictions that have enacted copyright exceptions meant to enable TDM research.

13. UK Copyright, Designs and Patents Act 1988, § 29A (UK) (amended by Regulation 3 of the Copyright and Rights in Performances (Research, Education, Libraries and Archives) Regulations 2014, No. 1372.
14. Article 38 of Law No. 2016-1231 38 for a Digital Republic added paragraph 10 to article L122-5 and paragraph 5 to article L342-3 of the Intellectual Property Code [Code de la propriété intellectuelle] (Fr.) (providing a TDM exception for works and databases respectively).
15. Estonian Copyright Act art. 19(3).
16. Urheberrechtsgesetz [Law on Copyright and Related Rights] art. 60d (Ger.) (amended on June 30, 2017, effective March 1, 2018)

Exclusion of "commercial" research

There doesn't appear to be any relevant commercial/non-commercial distinction with respect to TDM research and fair use in the United States.[17] In contrast, the UK text mining provision is limited to non-commercial research, and the European DSM Directive takes a bifurcated approach: the robust text mining rights in Article 3 only apply to non-commercial research institutions; whereas the weaker rights in Article 4 are available to all.

It's possible that when other jurisdictions address the question of text data mining and "fair use" or "fair dealing" that they might draw a distinction between commercial and non-commercial users. We don't think that this is how the law should be interpreted, but courts don't always do what we think they should do.

Finally, on this point of commercial use, it's also worth repeating that some of the general research rights we discussed before only apply to non-commercial research.

Exclusion of some exclusive rights

In the United States, the non-expressive use of a work in relation to text mining will not infringe any of the copyright owner's exclusive rights. The situation is not so clear overseas.

The text mining provisions in Articles 3 and Article 4 of the European Union Digital Single Market Directive apply to the reproduction right, but they don't apply to the European right of "communication to the public," the right of "making available to the

17. Sag 2019.

public,"[18] or the right of adaptation.[19] Although the reproduction right will usually be the primary concern of a text mining researcher trying to establish a corpus, these other rights could be triggered by subsequent uses of the corpus.[20]

Lawful access

The EU Directive and some other laws require that TDM databases be made only with works to which the researcher has "lawful access." This is not required by any of the U.S. precedents on text data mining.[21]

18. See Article 3 of the InfoSoc Directive.
19. Adaptation is not harmonized under the EU Copyright Directives, so it is hard to even say authoritatively what it means without consulting the laws of every EU member state.
20. We are particularly concerned about the possibility that when researchers share a research corpus, they might be considered to be making it available to the public. We don't agree with this interpretation, but it is enough to give us pause.
21. The term "lawful access" is defined in Recital 14 DSM Directive ("Lawful access should be understood as covering access to content based on an open access policy or through contractual arrangements between rightholders and research organisations or cultural heritage institutions, such as subscriptions, or through

Overriding contractual and technological restrictions

Article 3 of the DSM Directive does not allow private contracts (e.g. a publisher's license) to override the data mining right. There is no rule like this in the United States. The fact that a US researcher violated a contract that limited her ability to engage in text mining is unlikely to detract from her assertion of fair use; but her fair use argument is equally unlikely to count for much in a breach of contract suit.

We don't yet have any guidance on how the EU contractual override provision interacts with their "lawful access" requirement.[22]

The rights under Article 3 of the DSM Directive are also not subject to the usual restrictions that apply to overcoming "technological protection measures" or "digital rights management" restrictions on access. Again, this is not the law in the US.

Security measures and retention of copies

In the United States, the fair use status of TDM research may be

other lawful means... Lawful access should also cover access to content that is freely available online").

22. It's possible that legislation and court decisions implementing the DSM directive will say that a researcher who violates a condition of access to a database or a website will fall foul of the "lawful access" requirement. But it's possible that they will hold that the contractual override provision renders access lawful.

contingent on taking reasonable security measures to protect the corpus from unauthorized use beyond the parameters of fair use.

Article 3 of the DSM deals with the retention of works copied as part of a text mining process in a similar way. Under the Article 3 exemption, the covered organization must adopt an "appropriate level of security" and may retain the works "for the purposes of scientific research, including for the verification of research results."[23]

However, researchers relying on Article 4 face much more restrictive conditions. Under Article 4, the works may be retained only "for as long as is necessary for the purposes of text and data mining."[24]

Territorial rights in a globally networked world

Determining which territory's law applies

By now it should be clear that although the broad outlines of copyright law are fairly consistent from one country to the next, there are, nonetheless, some important differences that might be relevant to TDM research. The question we need to grapple with now is, to what extent are these differences a problem for TDM research in a world of cross-border data flows and international collaboration?

Copyright law is inherently territorial. United States copyright

23. Article 3(2) of the DSM Directive
24. Article 4(2) of the DSM Directive

law wouldn't take any interest in an unauthorized reproduction or performance that takes place entirely overseas. A pirated DVD sold on the streets of London doesn't violate US copyright law unless and until someone tries to bring it into the US. As far as we know, other countries feel the same way. By the same token, if a movie was in the public domain in the United States, but still subject to copyright in Italy, you couldn't sell pirated DVDs of that movie in the streets of Rome and expect to have US law applied. Indeed, because copyright law is inherently territorial, the advice "When in Rome, do as the Romans do" makes a lot of sense.[25]

However, the problem with global communications networks is that, as far as copyright law is concerned, you might simultaneously be in Rome, Sydney, Chicago, and Beijing.

Because the "harm" of copyright infringement consists simply of trespassing on the copyright owner's exclusive rights in a given jurisdiction, it is possible that simply making a work available on a server in one country could constitute copyright infringement in multiple countries.

Usually, foreign courts won't be interested in trivial or incidental cross-border infringements.[26] Generally courts only take an interest in infringers that intentionally target their jurisdiction in

25. We know that foreign law is often applied to questions of ownership, but that additional level of detail does not seem particularly relevant here.
26. This sentence elides a great deal of complexity. It does not hold true for cross-border actions within the EU, but it's a fair general approximation for a lay audience. Within the EU, there is no targeting requirement for cross-border copyright infringement, but for foreigners outside the EU, you have to look to the laws of individual member states.

the sense that they deliberately engage with an audience there. However, whether courts require intentional targeting of their jurisdiction, and how they interpret that requirement, both vary considerably.

The details of the activity matter

One of the most important things people tend not to understand about copyright law is that the details matter. Copyright is not a general right of exclusive advantage; copyright is a bundle of exclusive rights in relation to specific actions. In the vocabulary of the United States Copyright Act, copyright owners have the exclusive right to reproduce the work, make derivative works, distribute the work, and publicly perform or publicly display the work.

It's important to understand what is not included in the copyright owner's exclusive rights. Unless one of those exclusive rights is triggered, there is nothing wrong with "using" a copyrighted work, "learning" from it, or gaining some other advantage from it.

So, when we are thinking about international and cross-border copyright issues in relation to text data mining, we have to carefully evaluate which technical actions are being performed and what the copyright implications of those actions might be in different jurisdictions. We also need to think about the sometimes strange and metaphysical question of exactly where the action takes place.

We will go over some specific technical acts with respect to copyrighted works and explain their jurisdictional implications. Then we will take these basic principles and apply them to some common scenarios you might encounter in text data mining research.

Reproduction and making available

Reproduction

Reproduction is one of the core exclusive rights of the copyright owner. It is safe to assume that any reproduction made across a communications network can be thought of as taking place at either end. Thus, electronically transferring a file from country A to country B may well infringe the reproduction right at the source, and at the destination.

Making available

In jurisdictions that recognize a "making available to the public" right as part of copyright, simply making a work accessible online constitutes infringement, even if no one actually takes advantage of that accessibility. There is no "making available" right in the US (there is some disagreement here, but we are 99.9% sure) but this right is fairly common overseas.[27] If a copyrighted work is hosted on a server in country A and is accessible in country B, it has been "made available" in country B and could infringe the making available right in country B.

27. Note also that making available a copy may be considered circumstantial proof of actual distribution. See Robert Kasunic, Making Circumstantial Proof of Distribution Available, FORDHAM INTELL. PROP., MEDIA & ENT. L. J. 1145, 1163 (2008).

Distribution, performance and display

Distribution

Technically, a digital download of a copyrighted work is both a reproduction and a distribution. However, the distribution right is essentially redundant in the online context because the reproduction right can do all of the heavy lifting.

The distribution right is also potentially triggered by simply transferring possession of a physical copy of the work from one person to the next. In general, the distribution right is infringed in the place where the work is received.

The distribution rights sounds incredibly broad, but the distribution right is limited by the "first sale doctrine" (other countries call this the doctrine of "exhaustion"). Once the copyright owner has sold or given away a particular copy of the work, she no longer has any right to control any subsequent distribution of that particular copy. She still has the right to control copying, but the copy she just sold should be free from post-sale restrictions.

In some countries, the principle of exhaustion only applies to a sale within that country. The United States takes a much broader view. Under US law, the copyright owner's rights are exhausted by the first sale no matter where it takes place. The European Union takes a regional approach to exhaustion. So, a physical book sold in Paris can be resold in Berlin without further authorization, but a book sold in Pittsburg couldn't be.

In the United States, the right to import and export copies of works is treated as a subset of the distribution right. Importing a work into, or exporting a work from, the U.S. infringes the distribution right if it is done "without the authority of the owner of copyright" under U.S. law and the making of the relevant copies either "constituted an infringement of copyright" under U.S. law or "would have constituted an infringement of copyright" if U.S. law

had applied. It is worth emphasizing that U.S., not foreign law is the benchmark here.

Performance and Display

Even in the absence of a reproduction, copyright can be infringed by transmitting the work as a public performance or a public display. In the EU and many other jurisdictions, this would be a "communication to the public." Streaming video and broadcast radio are both examples of public performance/communication through transmission.

For the purpose of thinking about cross-border issues, it seems safe to assume that a work is performed/communicated either in the place where the transmission was initiated, or in the place where it was received. However, only the person making the transmission violates the performance right. So, if a work is streamed from country A to an audience in country B, the person making the transmission may be liable in both jurisdictions, but the person receiving the transmission wouldn't be liable in either.

The use of data derived from copyrighted works

The distinction between protectable original expression and unprotectable facts and ideas, is one of the universal building blocks of copyright law. The non-expressive metadata the results from text data mining research doesn't, in and of itself, infringe the copyright in any of the underlying works from which it was derived. This is important. Building a research corpus usually involves substantial amounts copying. However, once the corpus has been created, the computational process of querying the database to produce metadata may have no copyright significance.

Derived metadata does not infringe copyright because the derived data is not, in any relevant sense, a copy of the underlying works.

This means that there should be no copyright issue with exporting derived data to another jurisdiction, even if the copying that was necessary to build the research corpus in the first place would not have been allowed there. It also means that there shouldn't be any issue with allowing overseas researchers to query a U.S. corpus, so long as the results of those queries are confined to derived data.

Risk management

By now it should be clear to you that there are some theoretical cross-border copyright risks related to text data mining projects based in the United States that interact with the rest of the world. Our focus is primarily on how to identify and minimize those risks.

We can distinguish between theoretical risk and practical risk.

Here we use theoretical risk to refer to the technical application of the law on the books to the action in question to determine whether—if litigated—a court would likely find liability. We use the term practical risk to refer to the chance that the issue in question might actually be litigated. The two risks can operate separately from each other.

Sometimes there might be a high theoretical risk, but very low practical risk. Imagine a colleague emails you a copy of an article that you were missing from your database. There are countries where that appears illegal. But is the rule ever enforced?

On the other hand, there may be cases where the theoretical risk is very low but the practical risk is very high. The Google Books Project was a new, very public, and very large scale use of copyrighted works. Google knew its design of the project was compliant with fair use. But it surely also knew that if it wanted to

carry the project through, it would have to budget in substantial litigation costs.

At the end of the day, you need to make your own judgment about practical risk, based on what we can tell you about theoretical risk. How you want to balance these risks and what you think is an acceptable level of risk are questions we can't answer for you.

The distinction between theoretical risk and practical risk is quite important in the cross-border copyright context. Even if a US institution was judged to have violated copyright law in some overseas jurisdiction, the practical risk of litigation may be incredibly low. Assuming that the US defendant has no assets in the foreign jurisdiction, the foreign plaintiff would need to take legal action in their own jurisdiction, and then undertake a separate action in the United States to have the judgment enforced.

This might be especially challenging if the conduct complained of would be fair use under U.S. law because of the quasi-constitutional status of fair use. The Supreme Court has indicated that at least some aspects of the fair use doctrine and the idea-expression distinction are critical to the constitutionality of copyright law in light of the First Amendment. If a foreign judgment condemns activity that would be permissible under the fair use doctrine, the US defendant would be well placed to argue that the final judgment should not be enforced due to its conflict with public policy, namely the First Amendment.[28]

The outcome here is far from certain: the defendant would have to show much more than the simple fact that an American court would have come to a different conclusion, it would have to show that a finding in favor of the plaintiff would be repugnant to the

28. See Sarl Louis Feraud Int'l v. Viewfinder, Inc., 489 F.3d 474 (2nd Cir. 2007).

First Amendment.[29] Nonetheless, this is a significant obstacle for a foreign plaintiff to overcome.

Scenarios

In this section we will work our way through TDM scenarios with the potential to raise cross-border issues. Our aim is to identify when overseas copyright law would be relevant and when it wouldn't, and to address potential best practices in risk identification and mitigation.

We will also identify where there is potential to lobby for changes to copyright law at a national or international level that would improve research opportunities without undermining the legitimate interests of copyright owners.

We will try to focus here on use cases that are arguably within the boundaries of United States copyright law but might raise questions in other jurisdictions, or at least require us to know something about the law in other countries.

29. See eg Yahoo!, Inc. v. La Ligue Contre Le Racisme et L'Antisemitisme, 169 F.Supp.2d 1181, 1189-90 (N.D.Cal.2001) (holding unenforceable French judgment rendered under law prohibiting Nazi propaganda because such law would violate the First Amendment), rev'd on other grounds, 433 F.3d 1199 (9th Cir.2006) (in banc).

Building a corpus

Reproducing copyrighted works for the purpose of TDM in the US

Reproducing copyrighted works for the purpose of text data mining will be treated as fair use in the United States. As long as the reproduction takes place in the United States, there are no international or cross-border issues, even if the copyright is held by a foreign author or a foreign corporation. Foreign copyright owners have at least the same rights as American copyright owners under our system, but if they are objecting to something that happened in this country they are, in effect, asserting their United States rights and thus, US law will apply.

Receiving physical copies from abroad

Suppose an institution in the United States receives physical copies of works from overseas. For example, someone might send TextPot (our Hypothetical academic text mining institution) a box full of old science fiction books or a box of French sitcoms recorded on DVD.

If these copies were made legally overseas, then under the first sale doctrine, there should be no problem under U.S. law with importing them into the US. Because of the way the import/export provisions of the Copyright Act (Section 602) are written, the relevant question is with respect to the making of the copies to be imported "would have constituted an infringement of copyright" if U.S. law had applied. If it would have, then importing those copies without the authority of the copyright owner infringes their US rights. If not, there is no U.S. infringement.

Suppose the copies were specifically made for the purpose of

inclusion in a text mining corpus in a country where that would violate copyright law. Clearly this has legal significance for the person(s) who made those copies overseas, but importing those copies would not violate the US Copyright Act because the relevant question is whether the making of the copies to be imported "would have constituted an infringement of copyright" if U.S. law had applied. This makes sense because the right to distribute the work, like all of the copyright owner's exclusive rights, is subject to the fair use doctrine as well as other more specific limitations and exceptions.

However, the export from the foreign source might infringe the overseas jurisdiction's distribution right: it depends on how that jurisdiction implements its own first sale doctrine (i.e. whether it has national or international exhaustion).

If the relevant copies were not lawfully made overseas, exporting them would most likely violate the foreign equivalent of the distribution right in the sending country.

From a U.S. perspective, the law is reasonably clear that there is no domestic liability for acts of infringement that occur overseas.[30] Nor is there domestic liability for "authorizing" within the territorial boundaries of the United States of acts of infringement that occur entirely abroad.[31]

The final question is whether simply importing a copy that would be legal in the U.S. but unlawful in the source jurisdiction triggers liability for the U.S. receiver in the jurisdiction from whence the works came? The answer depends on the US receiver's degree of involvement in the initial copying. If the US receiver explicitly or implicitly encouraged the making of the unlawful copies, it would

30. Subafilms, Ltd. v. MGM-Pathe Communications Co., 24 F.3d 1088 (9th Cir. 1994).
31. Subafilms, Ltd. v. MGM-Pathe Communications Co., 24 F.3d 1088 (9th Cir. 1994).

quite probably be liable for the overseas infringement. On the other hand, if the receiver did not play an active part in the making of the unlawful copy in the first place, liability should only attach to the exporter.

Receiving/obtaining electronic copies from abroad via a computer network (i.e., a download, not a CD or DVD)

This scenario is the same as the one above, except that the works are not imported in fixed copies, they are transmitted over the Internet. However, this difference in mechanism changes the legal analysis quite significantly.

The single action of transmitting an electronic file from a country such as Australia to the United States without the authorization of the copyright owner would implicate the reproduction right in both jurisdictions. The sending party would clearly be liable in both jurisdictions and there is a reasonable prospect that the receiver would be liable in the US as well.[32]

There would be no liability under US law for either party if the

32. In the US, one could argue that the receiver had not "made" the copy and thus the requirement of a volitional act is missing. However, if the receiver was sufficiently involved with the reproduction it might be seen as the party "making" it, or it could still be liable under a theory of contributory liability, the carrier's liability, or inducement. If the receiver did not ask for the material and did not know that it was coming, secondary liability would be unlikely to attach.

action is deemed to be fair use, applying US standards. Clearly, if the reproduction violated Australian law the sending party would be liable for copyright infringement there. What is less clear is whether an Australian court would also hold that the American receiver had violated Australian copyright law.

Retention of copies and security

Suppose Search Corp Italia (a for profit entity) scans an archive of Italian poetry from the 1950s for text mining purposes and transmits the archive to the University of Evanston in the United States on the understanding that the works will only be used consistent with the U.S. fair use doctrine. Search Corp Italia then deletes its copies of the files. What does the University of Evanston need to know about the storage and retention of those files?

The University of Evanston would need to store the files with appropriate security to maintain its fair use status in the U.S.

How an institution manages file storage, retention, and security can have important legal implications, but it is important to understand that once a file has been copied onto a particular server, the failure to delete it does not have any independent copyright significance in the U.S. There is no exclusive right to retain copyrighted works, and keeping something is not the same as reproducing it, distributing it, performing it, or displaying it. The same goes for security measures: failure to take adequate security measures can change how the initial copying is characterized, but simply having bad security does not trigger any of the exclusive rights of the copyright owner.

The fact that the University of Evanston has retained the files might take Search Corp Italia outside the scope of Article 4 of the DSM Directive. This is a problem for Search Corp Italia, but not for the University of Evanston.

Why would this raise an issue under the DSM Directive? If the

EU text miner is not a non-profit research organization or cultural heritage institution, then it will have to rely on the more limited provisions of Article 4 of the DSM. One of the limitations of Article 4 is that the works may be retained only "for as long as is necessary for the purposes of text and data mining."[33]

Generating and sharing data

Analytical processing by overseas researchers

Suppose that TextPot allows affiliated researchers from the EU to query the corpus? There are no copyright implications here as long as the process of turning text into data does not involve making a substantial copy of the underlying works, distributing those works, or performing or displaying them.

As we explained in previous chapter on copyright, the distinction between protectable original expression and unprotectable facts and ideas is one of the universal building blocks of copyright law. Not just in the United States, but around the world. The non-expressive metadata the results from text data mining research doesn't, by itself, infringe the copyright in any of the underlying works from which it was derived.

This is important. Building a research corpus usually involves substantial amounts copying. However, once the corpus has been created, the computational process of querying the database to produce metadata has no copyright significance. The derived data is not in any relevant sense a copy of the underlying works.

Accordingly, there should be no cross-border problem with giving

33. Article 4(2) of the DSM Directive

anyone the ability to query the corpus as long as the result of that query is on the right side of the idea-expression distinction.

What if the overseas researcher is getting access to more than just derived data? For example, text snippets, illustrative examples, replication subsets? We'll come to these questions shortly, but for now it's important to understand they are different to the data-only scenario.

Sharing and using the data

For the reasons we just discussed, there shouldn't be any cross-border issues with publishing derived data or making it available internationally.

Adjunct uses of original expression (snippets, verification, and validation)

Sometimes metadata is not enough.

It is very unlikely that the initial results of an academic text mining process could be taken at face value without some reference to the underlying works as validation. Our understanding of US law is that limited display uses for the purpose of the verification and validation of results would be well within the parameters of fair use. In addition, as the Google Books case illustrates, some limited expressive uses are also allowed if they are made for purposes, such as presenting results in context or allowing third parties to verify the accuracy or relevance of results. Classic transformative uses of this kind will be fair use so long as the amount displayed is reasonable in light of the underlying purpose and is unlikely to disrupt any cognizable market for the original work.

As discussed above, there should be no copyright law

impediments to transferring data derived from an American text mining corpus overseas, but it's possible that adjunct uses of original expression that would be considered non-infringing in the United States may violate copyright law in at least some overseas jurisdictions.

We are pretty confident that such adjunct uses would qualify as fair dealing in countries like Canada and Australia, but they seem to be beyond the scope of the TDM provisions of the new EU DSM Directive. Such adjunct uses may be allowed under the German text mining law. The German law permits the making the corpus available only to a "specifically limited circle of persons for their joint scientific research, as well as to individual third persons" for quality assurance. However, other exceptions and limitations may allow for similar results in other EU countries.

Recommendations: We think that the risk that making limited display uses for the purpose of the verification and validation of results violates copyright law is actually quite low in many overseas jurisdictions. A text mining project seeking to eliminate this risk would have to obtain jurisdiction-specific advice or simply limit the scope of access to persons within the United States through site access restrictions or geo-blocking.

Special issues relating to machine learning and AI

Can the contents of a machine learning algorithm infringe copyright in the training data?

Suppose researchers at TextPot train a machine learning algorithm on a corpus consisting of copyrighted works. In most cases, any features derived from the training set that become embedded in

the machine learning algorithm won't look anything like the original expression in the corpus itself. Accordingly, in the run-of-the-mill scenario, machine learning algorithms and their AI cousins don't raise any new copyright issues. As discussed above, the data derived from a corpus is not a copy of any particular work in the corpus, it can be used for any purpose without fear of copyright liability. That analysis doesn't change if the derived data is embedded in a machine learning algorithm.

Nonetheless, it's worth considering a low probability scenario in which a machine learning algorithm did actually embody enough of the original expression from the training data that it constituted either an infringing reproduction, or an infringing adaptation.

This scenario is unlikely under United States copyright law given current thresholds of what it takes to conclude that one work is too similar to another work and our current understanding of the minimum amount of expression required to cross the threshold of copyrightability. Both of these thresholds appear to be somewhat lower in the EU, consequently the risk may be slightly greater outside the United States.

In the United States, even if the content of a machine learning/AI program did constitute a prima facie reproduction or adaptation of some underlying copyrighted work, that use would be just as protected by the fair use doctrine as the initial copying of the primary works into a database. However, the same machine learning algorithm might fall outside the narrower protections for TDM in some overseas jurisdictions.[34]

Recommendation: machine learning algorithms which embody

34. One of the problems with the EU directive is that it does not apply to the right to make an adaptation. Presumably, this is because the adaptation right itself is not harmonized across the EU. Add cites to discussion of this issue...

non-trivial amounts of the original expression from copyright works should not be exported to a given jurisdiction without first ascertaining whether the algorithm might itself constitute an infringing adaptation of those works in that jurisdiction.

Works created by AI and machine learning techniques based on data derived from copyrighted works.

If the output of a machine learning algorithm is too similar to one or more of the underlying works in the algorithm's training set, that new work will infringe copyright under traditional copyright law principles.

Imagine an AI program that uses songs by Taylor Swift as a training set and produces songs that are very similar to Taylor Swift songs as the output.

If the t-AI-lor Swift songs are too similar to works of Taylor Swift, the fact that an AI was used to create them is largely beside the point. But the much more likely scenario is that the AI would produce works that are in the same genre and share features in common with the works in its training set, but that the new works don't actually meet any of the traditional tests of infringement.

In this much more plausible example, the mere fact that a work was created using data derived from a set of copyrighted works does not make the new work itself a violation of copyright.

Sharing the corpus

Access to the works that constitute the corpus

Making the entire research corpus available to the general public would be inconsistent with the fair use rationale for text data mining articulated in HathiTrust and reiterated in Google Books. However, an institution might give qualified researchers access to the corpus for research purposes related to text mining and still fall comfortably within the parameters of fair use in the United States. The more difficult question for our purposes is whether that kind of access needs to be limited to people within the United States.

Giving overseas researchers direct access to the corpus might violate the reproduction right in their home jurisdiction, and even if nothing is downloaded, it could violate the foreign equivalent of the public display right in addition to the "making available" right. It is possible that the foreign researcher's actions would be covered by limitations and exceptions in their own jurisdiction, but that is something that would have to be reviewed on a country by country basis. If we assume for the sake of argument that no such limitation or exception applies, the US institution would violate foreign copyright law in this particular cross-border scenario.

Recommendations: Unless the risk of that limited research access would violate copyright law in a particular overseas jurisdiction has been assessed and is regarded as sufficiently unlikely, overseas researchers should only be given direct access to the corpus from within the United States (this seemed less problematic in the pre-coronavirus era). We suggest making this a condition of access and also using geo-blocking as a backstop.

Reproducing the corpus overseas

There may be legal, technical, and policy reasons to want to reproduce or mirror a research corpus in a second location. Assuming that the corpus was built in the United States for TDM purposes, we are confident that reproducing it at a second location within the United States for a similar TDM purpose would also be fair use.[35] The US fair use analysis would not change if the second location was in a foreign jurisdiction, even if this violated foreign law.

Conversely, the fact that the original corpus was constructed within the parameters of American fair use would not prevent the reproduction of the corpus in some foreign country being characterized as infringement if that country has not made any accommodation for the practice within its copyright law.

The legal rules and standards applicable to text data mining outside the United States are in a state of flux. Relatively few jurisdictions have passed relevant legislation or addressed the issue through case law or administrative regulation. Members of the European Union are required to enact legislation implementing the Digital Single Market Directive by June 7, 2021[36] and it is not yet clear how broadly or narrowly the individual EU members will choose to follow that directive.

Article 3 and article 4 of the DSM Directive require "lawful access" to the underlying work. Our position would be that lawful access means that the particular copy used as source material was not created unlawfully under the laws of the jurisdiction where it was created. However, we can easily imagine a more restrictive

35. HathiTrust makes this explicit.
36. The Directive entered into force on June 7, 2019. Member states will then have until June 7, 2021 to implement the Directive.

interpretation that limits the right to research under the Directive to copies made with the actual authorization of the copyright owner.

There is an opportunity here for positive action at the international level. We faced a similar situation with the provision of accessible works to people with visual disabilities in the Marrakesh Treaty of 2013.[37] The Marrakesh Treaty established some essential minimum standards for copyright exceptions to allow accessible works to be produced for people with visual disabilities. A major question dealt with the recent Marrakesh Treaty for the Blind[38] was similarly whether an accessible format copy lawfully made in one country (e.g. the USA under fair use) could be lawfully transferred to countries that lack clear rights to make similar copies locally. The Marrakesh Treaty solved the problem with a new international rule requiring contracting parties to allow the import and export of accessible format copies under certain conditions. The World Intellectual Property Organization (WIPO) is set to discuss research-related international limitations and exceptions at an

37. Marrakesh Treaty to Facilitate Access to Published Works for Persons Who Are Blind, Visually Impaired, or Otherwise Print Disabled
38. See Treaty on Education and Research Activities https://www.wcl.american.edu/impact/initiatives-programs/pijip/impact/global-network-on-copyright-user-rights/treaty-on-educational-and-research-activities/ (the treaty text was developed through an academic research project and endorsed by 39 organizations representing tens of millions of teachers and researchers around the world).

upcoming meeting.[39] An import/export provision modeled on the Marrakesh Treaty should be part of that discussion.

39. In light of the coronavirus pandemic we cannot say for certain when that will be.

3. Technological protection measures

SEAN FLYNN AND MATTHEW SAG

Introduction

Sometimes the works you would like to analyze using text data mining tools are already available in a high-quality digital form. You may be able to get what you need in e-books acquired from Amazon, or journal articles downloaded from a publisher's website, or you might simply be able to scrape user generated content from a social media site. Or, even better, perhaps someone else has already done this work and is happy to share their source materials.

These modes of acquisition all sound very promising, but they raise a new set of questions that take us beyond the parameters of traditional copyright law. Some of the actions I have just described might involve circumventing technological protection measures or possibly illegally gaining unauthorized access to someone else's computer.

In the following sections, we are going to take a look at the issues raised by the anti-circumvention provisions of a law known as the Digital Millennium Copyright Act and the application of the Computer Fraud and Abuse Act and similar "anti-hacking" laws.

The problem of digital locks

What are technological protection measures and

digital rights management?

Works in digital form may be protected by technological protection measures (often just called, "TPMs") that control access to copyrighted works. These technological protection measures are also referred to as digital rights management (or "DRM"). We will use the terms TPM and DRM interchangeably here, but the simplest way to think about them is as digital locks. Like physical locks, digital locks can be used to control access to a thing or to limit what can be done with it.

Such digital locks are a potential problem for text data mining initiatives because often the cleanest and simplest way to build a corpus is to get access to authorized copies of the original works in digital form.

In the world of books, for example, cracking the encryption on an ebook sold by Amazon would give the researcher access to a much cleaner copy than could be achieved through OCR (optical character recognition). This mode of acquisition is also preferable in some cases because it overcomes coverage limitations in existing repositories. For those of you working with large volumes of audiovisual material, defeating encryption may be the only option to get content into a text mining database that wouldn't take decades.

Breaking digital locks is generally illegal in the United States

However, in spite of its attractions, building a research corpus by breaking DRM has at least one very significant disadvantage, in the United States at least, it's illegal.

(a) The anti-circumvention provisions of the DMCA

In 1998 Congress added some special provisions to the Copyright

Act which made breaking digital locks that protect copyrighted works a civil, and potentially also a criminal, offense.

These "anti-circumvention rules" apply separate and independent of any underlying copyright infringement. In the Digital Millennium Copyright Act, Congress added section 1201 to the Copyright Act. Section 1201 prohibits the circumvention of technological measures that restrict access to, or copying of, copyrighted works. It also prohibits the creation or distribution of tools that facilitate circumvention.[1] The various parts of Section 1201 are generally

1. The DMCA contains three provisions targeted at the circumvention of technological protections. The first is subsection 1201(a)(1)(A), the anticircumvention provision. The second and third provisions are subsections 1201(a)(2) and 1201(b)(1) the anti-trafficking provisions. Subsection 1201(a)(1) differs from both of these anti-trafficking subsections in that it targets the use of a circumvention technology, not the trafficking in such a technology. The anti-trafficking provisions are targeted to both access and copy control, but it is important to note that the DMCA does not contain a ban on the act of circumventing copy controls themselves. The DMCA makes it unlawful to circumvent a TPM that "effectively controls access" to a copyrighted work. 17 U.S.C. § 1201(a)(1)(A) (2012). The law does not prohibit circumventing a TPM that controls specific uses of a work without denying access altogether. However, it is unlawful to distribute any tool or device that would be primarily used for either of these purposes--i.e., circumventing access or use TPMs. Id. § 1201(a)(2), (b)(1).

referred to together as the "anti-circumvention" provisions of the DMCA. The DMCA creates civil remedies and criminal sanctions for violations of the anti-circumvention provisions.[2]

(b) There is no fair use exemption under the anti-circumvention laws in the United States

The hardest thing to accept about the anti-circumvention provisions of the DMCA is that they make breaking digital locks illegal, even when the copying/access that this allows would be covered by the fair use doctrine.

Arguably, this shouldn't be the case, but to date, courts in the United States have not been convinced. Thus, although the anti-circumvention provisions of the DMCA were not intended to limit or restrict fair use, courts have not treated fair use as a defense to the anti-circumvention provisions either.[3]

2. See §1203 (civil), §1204 (criminal). It also authorizes a court to grant temporary and permanent injunctions on such terms as it deems reasonable to prevent or restrain a violation of the anti-circumvention provisions. See §1203(b)(1)(injunctions).
3. Id. § 1201(c)(1) ("Nothing in this section shall affect rights, remedies, limitations, or defenses to copyright infringement, including fair use, under this title."). See Universal City Studios v. Reimerdes, 273 F.3d 429 (2d Cir. 2001); MDY Indus., LLC v. Blizzard Entm't, Inc., 629 F.3d 928 (9th Cir. 2010). The Federal Circuit requires that the act of circumvention has some potential nexus to copyright infringement, but does not go so far as to make fair use a defense to the anti-circumvention rules. See Chamberlain Grp., Inc. v. Skylink Techs., Inc., 381 F.3d 1178, 1203 (Fed. Cir. 2004); Storage Tech. Corp. v.

This means that although copying e-books for the purpose of text data mining research would be protected by the fair use doctrine, breaking the DRM on those e-books to make that copying possible would still be unlawful.[4]

(c) Possible future exemptions to the DMCA

The DMCA contains exceptions for reverse engineering and encryption research, but there are no similar provisions for text mining.[5] This may change. The Copyright Act authorizes an administrative procedure whereby the Librarian of Congress may grant temporary, three-year exemptions to the DMCA anti-circumvention rules.

At the time of recording, a group based in the Samuelson Law, Technology & Public Policy Clinic at UC Berkeley is currently pursuing this, but they have a lot of work to do. To make the case for a text mining exception they will have to show that the underlying use is non-infringing, that the absence of an exemption adversely affects users or is likely to do so in the near future.[6]

Custom Hardware Eng'g & Consulting, Inc., 421 F.3d 1307 (Fed. Cir. 2005).

4. I am assuming here the e-book DRM "effectively controls access" to a copyrighted work.
5. 17 U.S.C § 1201(f) (2012) (discussing reverse engineering); id. § 1201(g) (discussing encryption research).
6. U.S. Copyright Office, Section 1201 of Title 17 114-15 (2017), https://www.copyright.gov/policy/1201/section-1201-full-report.pdf. While temporary exemptions must be renewed every three years, the Copyright Office has instituted streamlined procedures to allow for the renewal of previously granted

(d) Text data mining by research organizations and cultural heritage institutions appears to be exempt from anti-circumvention rules in the European Union.

In April 2019, the European Union adopted the Digital Single Market Directive ("DSM Directive") featuring two mandatory exceptions for text and data mining. EU members have until June 7, 2021 to implement the directive in national legislation and our current assessment of the impact of the EU directive may change once we see exactly how that implementation proceeds.

It appears that the mandatory exception for text data mining by "research organisations and cultural heritage institutions" under Article 3 of the EU Digital Single Market Directive ("DSM Directive") seems to preempt otherwise applicable anti-circumvention laws, and also overrides contract or license terms that otherwise would restrict the ability to circumvent digital locks.[7]

Individuals and organizations relying on the narrower exemption under Article 4 — i.e., anyone who is not a "non-profit educational institution or cultural heritage institution" — remain subject to European anti-circumvention laws and do not get the benefit of contractual override.

But note, we have yet to see how the members of the EU plan to implement the DSM Directive, so the analysis above is preliminary.

exemptions on the existing evidentiary record. Id. at 143-46.

7. Article 3(3) provides that "Rightholders shall be allowed to apply measures to ensure the security and integrity of the networks and databases where the works or other subject matter are hosted. Such measures shall not go beyond what is necessary to achieve that objective."

Recommendations

Researchers in the United States need to make their own assessment as to whether the risks of potential civil and criminal penalties under the DMCA for violating the anti-circumvention rules are worth the rewards. We are aware that this practice is relatively common and that in many contexts the chances of enforcement action being taken are fairly low, but we are not in a position to recommend it.

Dealing with liberated works

Is DRM an issue for those who receive unlawfully "liberated" copies of works that were once protected by DRM?

(a) Lawful access in the United States

Many TDM researchers face the issue of whether they should take advantage of access to copyrighted works that have been initially copied illegally, or have had their digital locks broken in violation of the applicable rules under the DMCA.

There is no United States case law directly on point and none of the precedents confirming the fair use status of reproduction for the purpose of TDM suggests that lawful access is a precondition to fair use. Consequently, we can only address this difficult question by reasoning from first principles.

The overwhelming weight of authority rejects any notion that

lawful access is an absolute per se precondition to fair use,[8] and the more persuasive view is that the question of whether the work was subject to prior unlawful acts by third parties is irrelevant to the fair use analysis.

Furthermore, although there is mixed authority on the question, it is doubtful that the defendant's own morality and propriety should influence the question of fair use.[9] The fair use doctrine does not come down to questions of individual moral or artistic virtue, it defines the outer boundary of copyright protection. Case law suggesting that fair use is presupposed on "good faith" conflates the fair use doctrine with the rules developed by English courts of equity but this is erroneous. The fair use doctrine began as a matter of statutory interpretation,[10] not an equitable doctrine. Thus, although it is not beyond argument, the better view is that "a user's good faith is irrelevant to the fair use analysis."[11] Moreover, even

8. The only court to hold to the contrary is the Federal Circuit in Atari Games Corp. v. Nintendo of America Inc., 975 F.2d 832, 843 (Fed. Cir. 1992). However, that court's reasoning is inconsistent with later Supreme Court precedent and has been expressly rejected by subsequent courts. See NXIVM Corp. v. Ross Institute 364 F.3d 471 (2d Cir. 2004).
9. The Supreme Court has recently reiterated its "skepticism about whether bad faith has any role in a fair use analysis." See Google LLC v. Oracle America, Inc., 141 S. Ct. 1183 (2021)
10. Predicting Fair Use, 73 OHIO STATE LAW JOURNAL 47–91 (2012)
11. Michael Carroll, Copyright and the Progress of Science: Why Text and Data Mining Is Lawful, 53 UC Davis L. Rev.,

if good faith is relevant in some circumstances, we believe that (1) it would be simplistic to equate good faith to access to a legally made copy and (2) even if good faith is relevant under some circumstances, it likely has no real significance in the face of an otherwise compelling fair use argument.

However, caution is warranted. It is entirely plausible that US courts will be influenced by the prevalence of a lawful access precondition to the right to engage in TDM research in other jurisdictions (see below) and adopt that requirement here.

(b) Lawful access to the work in the EU and elsewhere

Article 3 of the DSM is limited to "reproductions and extractions … of works or other subject matter to which they have lawful access;" Article 4 is likewise limited to "reproductions and extractions of lawfully accessible works and other subject matter." There is some ambiguity about the scope of this requirement, but it seems likely that otherwise lawful text data mining would be rendered unlawful in Europe if the source material was copied illegally. At this point, we can only speculate as to how the lawful access requirement is meant to interact with the provision in Article 3 that appears to preempt otherwise applicable anti-circumvention laws.

It's also worth noting that several other jurisdictions have adopted a similar "lawful access" requirement.[12]

(c) Risk assessment and mitigation

This is an area where the applicable law may change and where specific factual permutations may be highly relevant. We recommend seeking advice before relying on these materials to design a research program. With those caveats in place, we believe

893, 898 (2019). See also, Mark A. Lemley, The Fruit of the Poisonous Tree in IP Law, 103 Iowa L. Rev. 245, 248 (2017)

12. Singapore, for one.

that the better view, on balance, is that the fact that a third party illegally copied a work, or illegally circumvented a technological protection measure relating to the work should not alter the fair use analysis. We also believe for similar reasons that even if a researcher unlawfully bypassed DRM herself, that should not affect the fair use analysis. However, because there is no US authority directly on point, we can only express a moderate level of confidence about these conclusions and we should note that US courts may be influenced by the law in other jurisdictions that goes in the opposite direction.

In terms of a hierarchy of risk, we think that the risk of merely obtaining works from a third party who broke DRM is low; and that the risk of obtaining works from a third party who obtained them in violation of the copyright owner's exclusive rights is moderate. In contrast, the risk of breaking DRM oneself (or encouraging someone to do it for you) is moderate in terms of how it might affect the fair use analysis, but relatively high in terms of potential liability under the anti-circumvention provisions of the DMCA.

In the European Union (and other countries outside the U.S.) the hierarchy of risk may be different. Researchers who benefit from Article 3 may be able to break DRM without liability (depending on how the DSM Directive is implemented), but it's unclear whether obtaining a work from a third party who had broken DRM would be regarded as violating the "lawful access requirement." Researchers in Europe would also be prohibited from conducting text mining research using source material that had been copied illegally.

Liability under anti-hacking laws for violating terms of service

Text data mining researchers often want to analyze texts and other primary materials that are available online in one sense, but are not

necessarily "available" to them, or at least, not for the purpose of text data mining. We might be talking about journal articles hosted by commercial publishers, or social media content hosted by Facebook, classified ads hosted by Craigslist, or even company press releases on a corporate website.

This content may be hidden behind a paywall, not shared at all, or access may be subject to terms and conditions that do not permit text data mining. The basic contract law issues in this scenario have been/will be dealt with elsewhere, but in addition to those issues, researchers also need some familiarity with anti-hacking laws such as the Computer Fraud and Abuse Act ("CFAA").

These laws make it illegal to "access" someone else's computer system without authorization. I'm sure that we can all imagine some scenarios where access is clearly authorized, or clearly unauthorized, but there is a substantial gray area in between that we need to address.

Websites protected by a password, a paywall or similar devices

The Computer Fraud and Abuse Act (or "CFAA")[13] is a pre-Internet law aimed at preventing computer hacking. The CFAA has been around for a while, but there is still some ambiguity about the scope of conduct it prohibits. As the Supreme Court has explained, the statute "provides two ways of committing the crime of improperly accessing a protected computer: (1) obtaining access without authorization; and (2) obtaining access with authorization but then using that access improperly."[14]

13. 18 U.S.C. § 1030 (2012).
14. Musacchio v. United States, 136 S. Ct. 709, 713 (2016).

Let's start with something simple: accessing a password-protected computer system without authorization, or when authorization has been specifically revoked, violates the CFAA.[15]

Working around authentication controls or permission requirements (such as usernames and passwords), using stolen usernames and passwords, or somehow defeating payment requirements, are all examples of conduct that would violate the CFAA in most circumstances.

Such conduct should be strictly avoided.

There is a distinction between violating terms and conditions and computer hacking

Most courts recognize that there is a critical distinction between the violating terms and conditions of access and accessing a computer system without authorization.

Whether merely violating conditions of access to a computer system that is not open to the public triggers CFAA liability is a matter of contention. The better view, adopted in the Ninth Circuit and the Fourth Circuit, is that it does not.[16] However, the First,

15. Facebook, Inc. v. Power Ventures, Inc., 844 F.3d 1058, 1067 (9th Cir. 2016).
16. See Facebook, Inc. v. Power Ventures, Inc., 844 F.3d 1058, 1067 (9th Cir. 2016)("[A] violation of the terms of use of a website—without more—cannot establish liability under the CFAA."); Nosal I, 676 F.3d at 862 ("We remain unpersuaded by the decisions of our sister circuits that interpret the CFAA broadly to cover violations of corporate computer use restrictions or violations of a

Seventh and Eleventh, take a broader view of what it means to "exceed authorized access" under the CFAA.[17]

The difference largely comes down to whether the court sees the CFAA as an anti-intrusion statute, or embraces a more expansive contract-based interpretation of the CFAA's "without authorization" provisions.

The emerging consensus appears to favor interpreting the CFAA

duty of loyalty.") See also Oracle USA, Inc. v. Rimini Street, Inc., 879 F.3d 948, 962 (9th Cir. 2018) (interpreting an analogous state law the Ninth Circuit held that "taking data using a method prohibited by the applicable terms of use, when the taking itself generally is permitted, does not violate the CDAFA"). For cases rejecting a broader interpretation of "exceeds authorized access" under the CFAA, see United States v. Valle, 807 F.3d 508, 528 (2d Cir. 2015); United States v. Nosal, 676 F.3d 854, 862-63 (9th Cir. 2012); WEC Carolina Energy Sols. LLC v. Miller, 687 F.3d 199, 207 (4th Cir. 2012).

17. See Brown Jordan Int'l, Inc. v. Carmicle, 846 F.3d 1167, 1174-75 (11th Cir. 2017); United States v. John, 597 F.3d 263, 272 (5th Cir. 2010); Int'l Airport Ctrs., L.L.C. v. Citrin, 440 F.3d 418, 420-21 (7th Cir. 2006); EF Cultural Travel BV v. Explorica, Inc., 274 F.3d 577, 583-84 (1st Cir. 2001). The Eleventh Circuit has acknowledged criticism of its decision in Rodriguez in a way that clearly invites Supreme Court review, but continues to adhere to it nevertheless. See EarthCam, Inc. v. OxBlue Corp., No. 15-11893, 2017 WL 3188453, at *9 n.2 (11th Cir. July 27, 2017).

as an anti-intrusion statute. This interpretation is particularly favored in cases where the computer system is available to the public at large without registration or password protection.

In the recent case of hiQ Labs, Inc. v. LinkedIn Corp., the Ninth Circuit court of appeals held that accessing a computer system that is available to the public at large does not trigger liability under the CFAA, even if permission to access has been specifically revoked.[18] The Ninth Circuit reasoned that "the CFAA is best understood as an anti-intrusion statute and not as a misappropriation statute," and thus obtaining information by scraping that was "available to anyone with a web browser" fell outside the scope of the CFAA.[19]

In early 2020 the District Court for the District of Columbia addressed potential liability under the CFAA in a research context. The court in Sandvig v. Barr held that accessing online hiring websites for the purpose of conducting academic research would not violate the access provisions of the CFAA, even though such access would clearly violate the websites' terms of service.

The researchers were conducting audit testing on employment websites by submitting fake resumes in order to determine whether the algorithms used by the websites were racially biased. This deception clearly violated the applicable terms of service. Nonetheless, the court concluded that "the CFAA does not

18. hiQ Labs, Inc. v. LinkedIn Corp., 938 F.3d 985, 1001 (9th Cir. 2019) (Concluding for the purpose of a preliminary injunction that the hiQ Labs had "raised a serious question as to whether the reference to access 'without authorization' limits the scope of the statutory coverage to computer information for which authorization or access permission, such as password authentication, is generally required.)
19. Id.

criminalize mere terms-of-service violations on consumer websites and, thus, that plaintiffs' proposed research plans are not criminal under the CFAA."

At the time of recording, the US Supreme Court had agreed to hear a case addressing these issues, but the hearing date has not yet been set.[20]

Recommendations

To avoid civil and criminal liability under the CFAA, researchers should not defeat access controls to non-public computer systems.

Researchers in the First, Seventh and Eleventh Circuits (i.e. the states of Maine, Massachusetts, New Hampshire, Puerto Rico, Rhode Island, Illinois, Indiana, Wisconsin, Alabama, Florida, and Georgia) should also refrain from violating the terms of service will govern access to non-public computer systems to avoid liability under the CFAA. Researchers in those jurisdictions planning to violate the terms of service for access to computer systems open

20. The Question Presented in Van Buren v. United States, is "Whether a person who is authorized to access information on a computer for certain purposes violates Section 1030(a)(2) of the Computer Fraud and Abuse Act if he accesses the same information for an improper purpose." The case was argued in December 2020 and had not been decided as of May 14, 2021. For a review of the argument, see https://www.scotusblog.com/2020/12/argument-analysis-justices-seem-wary-of-breadth-of-federal-computer-fraud-statute/

to the public are in a slightly better position, but they still face considerable risk.

Outside the First, Seventh and Eleventh Circuits, we believe that the view that the CFAA is an anti-intrusion statute should hold sway, and that mere violations of terms of service will not trigger liability under the CFAA. Of course, the Supreme Court may hold otherwise and we will be watching the case of Van Buren v. United States with great interest.

At the moment, (at the time of recording) this is clearly the law in the Ninth Circuit (Alaska, Arizona, California, Hawaii, Idaho, Montana, Nevada, Oregon, Washington), the Fourth Circuit (Maryland, North Carolina, South Carolina, Virginia, West Virginia), and the District of Columbia.

4. Licensing

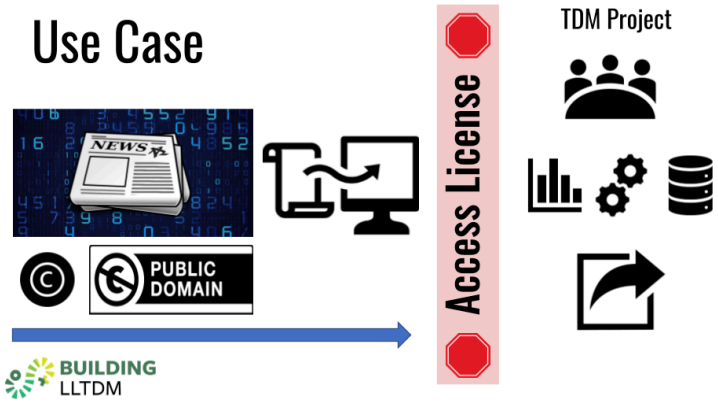
SCOTT ALTHAUS; BRANDON BUTLER; KYLE K. COURTNEY; AND GLEN WORTHEY

Licensing use case

Let's identify a use case for you to keep in mind as you are moving through the Licensing chapter.

Let's say a TDM scholar has a new critical interdisciplinary research project on women's roles in top 50 corporations in the United States from 1920-2020. For their analysis, they want to create a data file that includes both the text and basic metadata for selected articles focusing on women's roles from key national newspapers, which also includes coding work developed for the project to mine and index all the information. The scholar believes that the unique dataset they will create via the TDM process, which includes both the metadata and full text of the articles, would be incredibly useful to other researchers studying this same interdisciplinary topic. Because of the scope and range of the timeline, the 1920s through 2020, some of these articles are in the public domain and some are in-copyright.

For those that are in copyright, the TDM scholar can access and hopefully mine some of these articles through their University Library's licensed resource. These types of databases are commonly sold to a higher education market by a vendor that has a relationship with major newspapers and collects, indexes, and provides full text access to these newspaper articles that are historical in nature.



Both copyright law and contractual agreements affect how TDM may be conducted.

However, some of the articles will be accessed and mined via a subscription agreement. The TDM scholar has an online subscription to this newspaper, and receives daily emails featuring the day's articles. Additionally, one national paper in particular that was the subject of some of the scholar's research provides access to the historical newspaper articles via a separate agreement, which is not part of the daily circulation agreement.

The scholar wants to mine these newspapers, develop the coding, and publish this dataset, with the selected articles in full text, as a public-use file.

This might be a very familiar scenario. As we will review, a license or contract is an agreement between two parties to specific terms. A license or contract can modify, change, or alter rights.

And while the licensing of digital content exists in a legal realm that is separate from copyright law, they do interact—as they most likely will in this scenario or any TDM scenario.

Now, please take a moment now to reflect on this scenario. What issues do you see arising? What are some factors identified that will benefit this work? What might require more explanation or

negotiation? How would you begin to strategize in understanding any risk on the use, input, and output of the data and full text?

Contract & licensing basics

While we read about copyright and fair use in the earlier chapter on copyright, the second step is to determine the details of access to the materials. Many different contracts and agreements govern access to copyrighted materials and define what particular uses a researcher may make of these materials.

As we explored a bit in the use case, a strategic question to ask when you are beginning to be involved in a TDM project is: How will you access this material? The answers will vary. Some access will be via a library-licensed resource, which is sometimes part of institutional wide access. Some access might be through public facing websites featuring terms of use, and other access might be through an individual subscription which had an agreement that you clicked on to access.

Contract law is about enforcing promises. A contract is a promise or a set of promises for the breach of which the law gives a remedy, or the performance of which the law in some way recognizes as a duty. Licenses are most often granted within the context of a contractual relationship and often the same words used to create the license are also contained in the same instrument that also memorializes a contract. A license is a "contract not to sue." For our discussion, then, a license or contract is a legal interest created by a titleholder granting use-privileges to some non-titleholder. We will use the terms "license" and "contract" interchangeably.

So, as you can imagine, contract and licensing agreements can determine what a TDM researcher can do within legal bounds. Many of us will never have to write a contract from scratch. Trust us, this is a good thing! However, we do want to explore the underlying

contract and licensing system so that you have some context for the parts in the legal process that makes a contract or license valid.

The first is the **offer**. The offer is where one of the parties made a promise to do some specified action in the future.

Second is **consideration**. This is where something of value is promised in exchange for the specified action or non-action. This can take the form of a significant expenditure of money or effort, a promise to perform some service, or an agreement not to do something. Consideration is the value that induces the parties to enter into the contract.

Third, we have **acceptance**. The trick is that the offer has to be clearly accepted. Acceptance may be expressed through words, deeds, or performance as called for in the contract.

And last is **mutuality** or "meeting of the minds." It is necessary that the contracting parties had "a meeting of the minds" regarding the agreement. This means the parties understood and agreed to the basic substance and terms of the contract.

Beyond the legal requirements, there are also several contract provisions that are standard.

- **The Parties**. Definitely be sure you are naming the correct parties. And, this is a good area to look for in case you take the permission route or need to contact the right person or party. The publisher, vendor, or database might have one name, but the legal party to the contract—the corporation or person that has the rights—might be listed there, with a different name.

- **The Overview**. It is a mistake not to at least consider drafting, asking, or including an overview or purpose. Think of the overview as a chance to tell parties (and third parties viewing the contract) what the contract is about in a few paragraphs. This could help other users down the road that have to interpret this contract or license.

- **Payment section**. As stated before, consideration in the

formation of a contract can be simple—a payment, for example. If it is complex, you can refer to the contract section that sets forth other consideration: scheduling, quarterly payments, or per-use payments might be listed here.

- **The Date**. This is often overlooked. Be sure the date of execution by each party is included so that there will be a time at which the parties became bound to the contract. And, this may be related to when the agreement "starts the clock" if it is a limited timeline or subject to renewal based on this date.

- **The signature**. Print or digital is acceptable.

Boilerplate clauses are often standard, and most are not typically heavily negotiated. But they are important. Many contract disputes depend on the drafting of boilerplate clauses such as termination, force majeure, and entire agreement.

Why are they important? Most likely, the TDM project you are dealing with will have boilerplate language—even if it's a closed or open license!

Types of licenses & contracts

In the last section we mentioned boilerplate. The opposite of a boilerplate clause is one which is written and expressly addresses the desired outcome.

For the next few examples, we will look at TDM contact language utilized in the NERL Consortium Generic License Agreement and the Liblicense Model Agreement. These are drafted as ready-to-apply provisions that could work with a variety of licences and could be incorporated into a standard authorized agreement with a vendor.

> Authorized Users may use the Licensed Materials to perform and engage in text and/or data mining activities for academic research, scholarship, and other educational purposes and may utilize and share the results of text and/or data mining in their scholarly work and make the results available for use by others, so long as the purpose is not to create a product for use by third parties that would substitute for the Licensed Materials.

These clauses specify uses that are familiar to most TDM work and directly address the needs and issues that arise.

Note that this selected language is integrated into the document using the same uniform language as the original contract, including defined and capitalized terms such as Authorized Users, Licensor, and Materials.

Note also how the clause outlines the limits, defining the purpose of the use as different from the protected commercial market.

Occasionally, you will get pushback in proposing these clauses. Always be sure to have a backup clause or justification for TDM or related clauses. For example, the fees provision that is listed at the top of the example below is rejected, you might, as suggested by this model Liblicense Agreement, limit or categorize the fees with the bullet points listed below. Always be ready with another clause if you can:

> Licensor shall provide to Licensee, upon request,

copies of the Licensed Materials for text and data mining purposes without any extra fees.

- OR: If the licensor insists on referencing fees, they should not exceed the cost of preparation and delivery
- OR: If Licensee or Authorized Users request the Licensor to deliver or otherwise prepare copies of the Licensed Materials for text and data mining purposes, any fees charged by Licensor shall be solely for preparing and delivering such copies on a time and materials basis.

And that's some of the difference with boilerplate and negotiated clauses. While you can't change boilerplate, you can negotiate with these TDM specific clauses.

Now we focus on some of the most common types of contracts.

Non-negotiated licenses

Non-negotiated licenses are typically associated with major publishers and online resources. They are filled with the generic boilerplate terms, and, additionally, as the title states, do not typically accept any negotiated terms. In easy terms, this license is called "take it or leave it." The non-negotiated licenses default uses license terms that are biased in favor of the licensor. Again, they offer little room for changes or addendums to attach to the contract. TDM is certainly new enough of a field to have been completely left out of any previous access or purchase licenses,

although we will discuss some places they do exist in other sections or language.

Non-negotiated licenses can also come in the form of a common mass market license (like in software or vendor products) and click-wrap or browse-wrap. Sometimes they are part of a more generalized public license, which will be covered in a later section of this chapter.

A librarian or researcher is forced to weigh the non-negotiated license provisions as part of the cost-benefit analysis of assenting to the agreement. The key question is what may be forbidden under this document that I actually need to do for my scholarship or project?

Click wrap licenses

Click licenses have many names: click-through, clickwrap, splash screen, or even click-to-accept contracts. But many of us are well-familiar with this type—we all have probably downloaded an app and checked "I agree" without reading the license. All of these are a type of license where a user must expressly assent to a non-negotiable unilateral agreement by clicking a button displayed next to or below a statement. The button does most of the work here: it asks the user to accept or agree to the proposed contract terms. In some cases a licensor will use a checkbox and/or scrolling mechanism to let the user view or browse through the entire agreement and to make sure you have scrolled to the bottom before clicking the button. A quick side note: this scroll through method does not ensure that the user actually read the agreement —it is just one method to get the user to at least scroll through it.

Despite the fact that many users do not read the text, these agreements have been upheld by both state and federal courts, provided that the text preceding the acceptance button makes it clear that a user is accepting the terms of a contract and not merely

signifying readiness to proceed to the next screen, at least where it is clear about the terms. The user consents to these conditions by clicking on a dialog box on the screen, which then proceeds with the transaction.

Two factor authentication (for example, texting a code in response to a click) is used as well. This is called incorporation by reference. It shores up the legal argument that the actions were sufficient to establish express assent to the Terms and Conditions in the agreement.

Occasionally, there is a basic link to the terms which reside elsewhere. Either way, the check or click is the assent to the terms of the agreement.

However, if you are concerned about certain clauses or terms affecting TDM and you do want to read and not click right away—and we'd highly recommend that—there are some key sections to look for where TDM related clauses may reside. One is certainly a section on "Authorized Uses" or "Permitted Uses." Note that occasionally there will be a section on non-permitted uses or restrictions. Moreover, the definitions section occasionally even defines TDM right there. And finally, TDM-related clauses may be found in any sections listed as "Intellectual property" or "copyright." The TDM-related clauses are typically found in some or all of these sections.

Browse-wrap licenses

Browse-wrap licenses are another type of non-negotiable, unilateral contract where express assent is not obtained. These licenses are typically a static display of the terms and conditions (or "Terms of Service") for the resource. And usually it is presented through a hyperlink or language in the footer. This indicates to the user that by using the resource, you are bound by those terms.

These browsewrap agreements may be enforceable, but only if

assent or a "meeting of the minds" may be fairly implied based on the conduct after a user is put on actual or reasonable notice that access or use is subject to these terms and conditions. Courts have even looked to see if the conduct could be continued use of or access to the website, database, or service. Or the conduct can be identified that the user downloaded the product.

Interestingly enough, a study by two law professors in 2019 found that 99% of the 500 most popular U.S. websites had terms of service written as equally complex as an academic journal article, which makes them, possibly, inaccessible to most humans.

Here's a quick negotiation strategy: If you are creating or negotiating parts of a license with a TDM project, seek to include language that the license agreement has precedence and prevails over any click-through license on the licensor's site and that any proposed language for a click-through license is approved by the licensee prior to implementation. Some licensors give you a license, but then link to other terms in some other URL somewhere else that you are also bound to—and sometimes these terms are different or confusing because they may be generic and not specific to the TDM License.

Again, if you have concerns, look for the sections on "Authorized Uses," the definitions section, sections listed as intellectual property or copyright. The TDM-related clauses usually live in there, or in parts in all of the sections.

Open and public licenses

This section covers open and public licenses.

Public licenses are "boilerplate"—a term introduced you to in an earlier section of this chapter—meaning (very roughly) that they're non-negotiated. These are licenses under which copyright holders may choose to release their works for use by the public without requiring special permission.

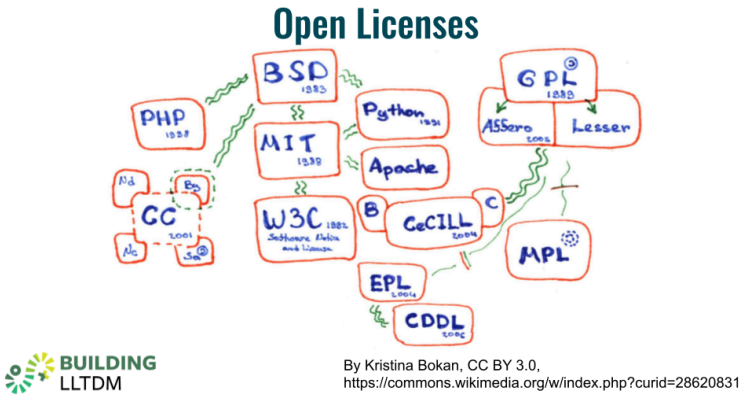
Probably the most famous public license—really, a suite of licenses—is the famous Creative Commons licenses.

Some people believe that Creative Commons is somehow the "opposite" of copyright, or that it somehow negates copyright. That is not the case: copyright in a work is generally an automatic right (as long as that work meets a few very specific criteria). Copyright doesn't require registration, doesn't require a little "C" in a circle, and it remains in force until its term ends.

But what an open public license does is to provide a mechanism for copyright holders to grant to "the public" permissions to use their work. The copyright holder relinquishes some of the rights to which copyright law entitles her. It's a license—as we've put it before, a "contract not to sue"—between the copyright holder and the public, for particular uses of a work that would otherwise be restricted, and violations of which could be litigated.

To reiterate what we've learned in previous segments, licenses (and other contracts) operate in a separate legal realm from copyright. The don't undo or modify copyright, but rather (in an interesting sort of turnabout) they actually rely on the copyright holder's exclusive economic rights in order for others to do something interesting with their works: for example, to choose not to make money from their creations, or to choose not to prevent redistribution or derivative works.

The world of public licenses is immense! Even an important subset of that world, the realm of open licenses, is immense! The chart below illustrates the tremendous variety of the "network of open licenses," and a sort of genealogy and chronology of their development. We won't spend much time on their myriad flavors and nuances, but let's talk briefly about a few of them, starting with the example of the very image you're looking at now.



Various open licenses.

This illustration is a copyrighted work. It has an author, Kristina Bokan, who holds the copyright to her creation. But she has granted the public a license to use it—and look how helpfully she's done so, highlighting (in broken green outline, near the left-hand side of the chart) the specific license she has chosen for it. It is under these terms that she is allowing all of us—including you!—to use her work. This license is her "contract not to sue" us for our reuse of her work—without us even having to ask her—and it's our contract with her to acknowledge her as its creator. We do this by including her name on the slide.

These two agreements represent the totality of our mutual agreement. We didn't negotiate it: she set the terms herself, and we accepted them simply by using her image in this chapter. With this interaction, our contract is settled and enforceable. Kristina can't take away our right to use her cool chart, and we are obliged always to give her credit for it—which of course, as responsible users, we would always do anyway.

"Copyleft" and software licenses

Let's focus briefly on one corner of this chart, highlighting one family of licenses sometimes called "copyleft."

Many readers will be able to decipher many of the words denoting the licenses you see here (listed in roughly decreasing order of decipherability): Python (the currently popular programming language), Apache (the software that runs most of the world's web servers), and W3C (the World Wide Web Consortium).

MIT stands for precisely the great university that you probably think it stands for. And the B in BSD stands for the publisher of this resource, UC Berkeley; the SD in BSD stands for "Software Distribution"—which is the key to what unites these open licenses: they're all generally used for software, as you might have guessed.

By the way, GPL here stands for the "General Public License of the Free Software Foundation's GNU Project. (And GNU stands, recursively, for "Gnu's Not Unix.")

In addition to the self-referential GNU acronym, note also the hard-to-translate pun implied in the term "copyleft," which implies that it's the opposite of copyright. In fact, it's not: it's just another license, very much dependent on copyright law (as we've noted repeatedly). A creator's exclusive right to define and determine the terms of use for her creation—even these very permissive terms—is the product of copyright, even when her intent is to manipulate (and in many respects, to undo) many of those terms.

Software licenses are pretty much like licenses for other kinds of content, like texts and images. But one of the distinguishing features of "copyleft" licenses is that they tend to include terms that explicitly allow derivative works, and they tend to require those derivative works themselves to be distributed with the same "share-alike" terms—which makes some sense given their origins. It was precisely the peculiarities of working with software that inspired this particular community to be so activist in creating and promoting open licenses in the late 1970s and 1980s: unlike a book

or a record, software begs to be tinkered with: debugged, modified, copied, and so forth. And the copyright regime, together with the restrictive licenses that made billionaires out of many copyright holders, seemed an impediment.

Creative Commons licenses

But maybe it shouldn't be said that software is "unlike a book or a record" in its desire to be modified and remixed. One of the hallmarks of the next major phase of open licensing, Creative Commons, is the idea of "remix culture" popular in the 1990s and 2000s, and trumpeted by the founder of Creative Commons himself, Lawrence Lessig. That's our next topic.

Lessig created and began promoting these licenses through the Creative Commons Foundation just over 20 years ago. He has written prolifically (and highly readably) about the origins and philosophy of Creative Commons; about its particular importance in the Internet era; about "remix culture" and "free culture"; and about other aspects of this contract-turned-movement too many to discuss here. It's interesting and important work, even for non-lawyers.

In practical terms, Creative Commons licenses have a veritable smorgasbord of options that modify the blanket permissions granted by the licensor (that is, the copyright holder). These options can include the requirement that every use be accompanied by an attribution, as with CC-BY; or that only non-commercial uses are allowed; or that disallow derivative works (translations, etc.); and a number of others. Strictly speaking, because of the possibility of these restrictions, these licenses are considered by some people not to be open. For others, that interpretation is a little too fundamentalist: these licenses are specifically designed to make copyright-protected content more open, even if that openness may have some terms attached to it. Even more importantly, these

licenses are considered open by many because, by their design, any use not strictly prohibited is actually allowed—no questions asked, no lawsuits threatened, no money changing hands.

Thinking back to Kristina Bokan's helpful whiteboard chart, has she granted us any other permissions with her CC-BY license? Certainly! With this license, we not only can use her image in this chapter, we could also include it in other presentations and publications without asking permission—we could even just republish and re-distribute it by itself, without adding anything of our own! Likewise, Kristina is allowing us to translate her chart into another language, or set it to music, or use it in a collage; but if she had chosen a Creative Commons "No Derivatives" license, we would not have the right to do any of those things under the license.

Could we put it on a bunch of t-shirts and coffee mugs and sell them? Absolutely! Do we have to ask her permission? Would we owe her any royalties? No: she has already decided this question just by choosing this license. And if Kristina were opposed to that sort of thing (and many reasonable people might be), she could simply have chosen a Creative Commons "Non-Commercial" license and refused to have her creativity feed our crass commercialism. But she didn't do that.

Open licenses in TDM

What do open licenses have to do with Text Data Mining? Like any other contract, these licenses imply at least two parties, a licensor and a licensee—and people in the TDM community, whether as librarians or other practitioners, have opportunities to act in both roles.

So far we've mainly discussed the licensee role: what we're allowed, or not allowed to do with materials that have some sort of open license applied to them.

But it's also critical to understand that we ourselves often act

as creators—and thus as copyright holders—and therefore that we also have the right to determine the license and the terms under which we allow our work to be used. It won't surprise any reader of the present that we authors would encourage you, to the extent possible, to choose open licenses. Our libraries can and should do this with the research materials that we create; and scholars can do the same.

Let's look briefly at an important example of open licenses in the TDM community: the HathiTrust Research Center, and its "Extracted Features" dataset, a staple of TDM work. Although it's used in text mining, here's a remarkable thing: it doesn't actually contain any "text" as we normally define it! Instead, it consists solely of metadata about the texts contained in the massive HathiTrust Digital Library, all 17+ million volumes of them—including a substantial number (about ⅔ of the total) of in-copyright books.

This metadata, naturally, includes descriptions of each book, like any library catalog. But much more significantly (and radically, and intelligently, in our view) is that it includes metadata about each page, each line, each word, and even each letter and each number in those texts. This is all many text-miners need to do their work.

Even though many uses of the texts described by the Extracted Features data may be restricted or proscribed by copyright, because this metadata consists only of facts about these texts, it is not a violation of copyright for the HathiTrust Research Center to extract them or to share them with others: in fact, as we've already heard hopefully more than once (but it always bears repeating), the courts have found that precisely this sort of use is a fair use.

But this dataset itself, as a compilation of information with a particular arrangement, an apparatus, and documentation, all of which HathiTrust devised itself, is in itself a newly authored work whose copyright belongs to HathiTrust—which could, in theory, claim all sorts of rights for itself and restrictions for other people. But of course they don't do that! In order to promote its adoption and use (and reuse, and experimentation, and so forth), they've published it under a Creative Commons license.

This is sort of like alchemy: they've turned massive numbers of copyright-restricted (or at the very least, ambiguously protected) texts into free and open research materials—and then of course have shared that treasure by being free and open with its presentation of this dataset. This is not some legal loophole, but rather a conscious and conscientious legal innovation based on a solid understanding of the law. Amazing, isn't it?

Examples and case studies: Library e-resource licenses

Having revelled a bit in the glories of open licensing, let us now turn to some specific examples of private, non-open contracts and licenses that are often highly relevant to TDM practitioners: library e-resource licenses.

This section is based largely on some practical, real-life examples of library licenses as they relate to TDM. They're not all pretty, but we hope they'll be instructive.

The world of library licensing for e-resources can seem both complicated and shrouded in mystery. Often this is just a matter of the complexities of back-office library acquisitions processes (selecting, negotiating, signing, paying, getting access, setting up authentication and proxy servers, etc.), and the related feeling that nobody except those directly involved really needs to know how this particular sausage is made.

Non-disclosure agreements

But sometimes this mystery is intentional: many licenses, and the negotiations leading up to their signing, are specifically subject to

non-disclosure agreements ("NDAs"). These NDAs are imposed by vendors who don't want libraries to compare the supposedly "great deals" they're being offered with deals offered to other libraries. (Some might find secretive price-setting to be a legitimate business practice, although in this particular practice, the prices paid for the very same product by different libraries, and the discounts and "great deals" offered to them, can vary to such an extent, and be so irregular, that the entire pricing regime seems to border on the fictional.)

However, many people find NDAs fairly pernicious for other reasons as well, especially where the non-price terms are concerned. As we've seen, license terms can often drastically curtail some very important rights in areas of scholarship like TDM. There has been a movement in universities to ban the entering into contracts subject to NDAs, and several of us authors have been made proud and happy when our universities have done that. The particular example license terms about to be used as illustrations may have come before the NDA ban, but vendors' identities have been obscured just in case.

Adventures in commercial licensing to libraries

As described above, licenses are a form of contract: in particular, a "contract not to sue." But library licenses for electronic resources generally have a substantial set of terms—terms that are consequential to TDM work—long before anyone gets to the point of suing!

We should care about these licenses for at least two very important reasons: one is that they are, broadly speaking, licenses governing our right to read (including the specific type of reading to which this workshop is dedicated: text data mining as reading). Another is that we are all bound by these licenses when we read (or

use in any way) the texts that make up the e-resource—whether we know it or not.

And yet, how many non-librarian scholars or students do you think have ever seen, or thought about, or even know of the existence of these licenses? In our experience, not very many: even among library workers, it's rare that someone outside of a very few acquisitions people, or a special licensing librarian, has ever seen them!

How many people in the campus community generally even know that they're bound by the terms of a license signed secretly on their behalf by some librarian they don't know, a license that they didn't agree to, and haven't even seen?

To engage in some stereotyping, we've seen several different categories of reactions among users of library-licensed e-resources. The vast majority is largely apathetic: they don't know and don't care, and that's generally okay: their use of e-resources is pretty well covered by pretty much any license that the library may have signed on their behalf.

> **A bit of a digression**: Actually, there is one occasion when almost everyone on campus bumps into e-resource licenses, and it's often a deeply frustrating one: when trying to access licensed resources from off campus. What a system: libraries pay many hundreds of thousands of dollars for paywalled digital content, and then spend hundreds of thousands more to set up, maintain, and troubleshoot systems to unlock that content for our authorized users—after which these users spend countless precious scholar-hours trying to make those systems actually work for them.

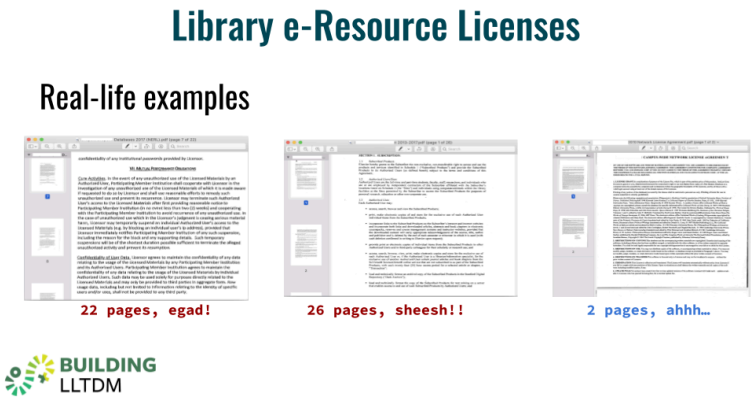
Among readers who are interested in more than simply reading these licensed works—say, those engaging in TDM—there is

probably a broader (but no less problematic) range of knowledge about their licenses, for example:

- Bold ignorance: the savvy grad student who knows how to script and how to scrape, and generally believes that, if whoever put this stuff "on the web" didn't want them to scrape it, they wouldn't have made it so easy.
- Or fear: scholars who don't even bother asking for TDM access, because "what if someone gets into trouble?"
- Or even outrage that we librarians have agreed to a license that forbids them from doing TDM-based research.

Difficulties

Some people, both practitioners and scholars, may experience a deep dreariness as part of the license reading experience. But in case you're tempted to escape from an e-resource license back into the comfort of volume 3 of War and Peace or the multivolume classic of your choice, we'd like to posit that there are some truly important and not at all uninteresting bits of text here, at least not uninteresting to readers of this text. In spite of the difficulties experienced by non-lawyers reading legalese, we would encourage you to ask around your libraries to find and talk to the people who negotiate and maintain licenses, because they're so important to what we all do in the TDM community.



Examples of library e-resource licenses.

You might think, with so many lengthy, carefully crafted pages of strict legalese, that library licenses would be water-tight bastions of hard-and-fast terms and conditions. But far from it! Every license is a product of imperfect human authors, sometimes a long line of them inheriting prose from predecessors, on both sides of a complex purchasing and licensing transaction, and they merit multiple close readings because, as we've said above, they have real consequences.

Here are few passages from an actual e-resource license (underlining in the original; bold added here for emphasis):

> 6. **Data Mining**. Subject to any content-specific restrictions, Customer and its Authorized **Users may extract and compile data** from locally-loaded copies of the Purchased Content for Customer's teaching, learning, and research purposes
>
> […]
>
> 9. **Restrictions**. Except as expressly permitted above, Customer and its Authorized **Users shall not:**
>
> […]

i) **Text mine, data mine or harvest metadata** from the Service…

This license, from a major library vendor, seems to include two distinctly contradictory terms on the very same page! The first one, allowing that "Users may extract and compile data" from the resource, seems expressly to permit data mining. But the second one seems expressly to prohibit it.

Not only are many of these licenses dense and difficult to read, they're also, often, a real mess. But with practice, even non-lawyers can easily learn to spot problematic terms and try to eliminate them through negotiation. There's a sample of problematic licenses in the Readings section of this text, along with a good model license (which we'll touch on in a few minutes) for comparison.

Negotiation

Once we understand that a license is a voluntary contract, there are some important aspects to the licensing process that can play in our favor: complementary interests. For example, the vendor has a commercial interest (it wants to make a sale), and the library and its scholars have an academic interest (they want access to the vendor's content). So although we may have competing interests in the price, we really have common interests in finding agreeable terms.

Unfortunately, many library vendors either don't yet understand TDM practices, or overestimate their importance to an extent that leads them to believe libraries might be willing and able to pay a premium for TDM rights. Likewise, many of these vendors are third parties, selling content that they themselves may have licensed from an actual copyright holder, which obviously complicates matters. And naturally, the easiest and safest position for a vendor to take is a restrictive one.

But it's essential to push back. While we don't have any special

tricks to offer for negotiating licenses, we do strongly believe in a couple of principles: first, the right to read is the right to text-mine, and it's a right we should never willingly sign away. Some have advocated for the inclusion of a simple escape clause in our licenses, along the lines of, "notwithstanding any of the foregoing, nothing in this license should be interpreted to prohibit fair use of the licensed materials." Since the courts have ruled that TDM is generally a fair use, this clause should, in theory, provide blanket permission for TDM activities.

The second principle is to maintain the clear position that one of the primary affordances of electronic texts is, in fact, the ability to read them with a computer—that is, to do TDM. If the only allowable uses of a digital text are basically the same uses that we could make with print books (many of which we have in our collections anyway), why on earth would we spend these huge sums of money for an electronic copy? Mere convenience of access is not worth the premium that some vendors put on their electronic resources.

Finally, for all of these reasons, it's crucial to be prepared to walk away from negotiations and decline a purchase if the terms aren't right.

Model licenses

But these days, there's no need for anyone—vendor or library—to draft a license completely from scratch. In fact, it's better if they don't! One important innovation in recent years is the "model license," which various research library consortia have developed and adapted as an expression of what the library research community considers reasonable expectations for licensing terms. The Center for Research Libraries, NERL (the NorthEast Research Libraries consortium), and the California Digital Library all offer model licenses that are available to all—vendor and librarian

alike—to use as references, sources for terms, or even straight-out adoption.

The California Digital Library's model license has—no surprise—particularly good terms for TDM, including both explicit mention of TDM as an authorized use, and a fair use "escape clause." Here's a snippet of these simple, powerful terms (underlining in the original; bold added here for emphasis):

> Text and Data Mining. Authorized Users may use the Licensed Materials to perform and engage in **text and/or data mining activities** for academic research…
>
> […]
>
> Licensee and Authorized Users **may make all use of the Licensed Materials as is consistent with United States copyright law, including its Fair Use Provisions**.

These model licenses are important for several reasons: not only do they lighten the load of drafting from scratch, but even more importantly, they set general expectations that are broadly shared by the TDM research community. For example, the CDL model license presents as a given that research libraries expect to have text data mining rights—and particular kinds of terms—in their vendor licenses. In this way, vendors (many of whom have historically been quite unfriendly to the whole idea of text mining) are put on notice that academic expectations with regard to TDM rights are now clear, and that these are terms that our community, in growing numbers, expects and will demand.

This and the other model licenses we've mentioned are incredibly important resources, both tactically and strategically. Because they originate in the academy, they're favorable to academic uses—unlike commercial licenses, which are generally written from a strong protectionist instinct and with commercial interests foremost.

Although we advocate taking a tough TDM stance with vendors in the negotiation of licenses, we should emphasize that there's real value in establishing and maintaining good relations with them: vendors have something that we want and need, and they exercise control over it, whether we like it or not. It's worth remembering an

influential 2018 blog post by our co-author Brandon Butler, whose title says most of what you need to know: "For Text and Data Mining, Fair Use Is Powerful, but Possession Is Still 9/10 of the Law."

Breaches and consequences

In our experience, the consequence of not having good TDM license terms—or not exercising them if we have them, or not informing our communities about them—is that scholars inevitably find ways to get, or to attempt to get, the data they want by web-scraping or by some other systematic means that are often explicitly prohibited, and can have unpleasant consequences for both the vendor and the offending library (and beyond). This has happened frequently enough in our collective library experience that we suspect it's a fairly widespread occurrence—but it doesn't have to be that way.

The most immediate consequence of a vendor discovering what it considers to be illegal downloading is to shut off access to the entire campus. With good vendor relationships, these consequences have been temporary: librarians have been able to track down the offending (and often unsuspecting and well-intentioned) party, and offer an explanation of why a particular activity is prohibited. In an ideal situation, the librarian can propose a license- or fair use-enabled alternative to the prohibited methods. Given a solid relationship, the library is able to reassure the vendor that the prohibited activity has ceased, and the vendor will generally open things up again. (Remember that even the most rigidly license-enforcing vendors actually want us, above all, to resubscribe to their products.) This is a real hassle, but relatively minor in the scheme of things. It's better to negotiate clear terms up front.

Another reason to establish and maintain good relations with our vendors, aside from simple human decency, is so that we can confidently approach them with requests for special access or data deliveries for use by our researchers. It has been our experience

that vendors will do their best, against tradition and their protectionist instincts, to honor the request, and to give their customers what they need.

There's obviously much more to be said about library licenses, but we hope these examples and this discussion will encourage you to approach licensing thoughtfully, boldly, and without too much fear or loathing.

Websites and terms of use

The CFAA: Is scraping a public website illegal hacking?

One concern that may arise in connection with scraping public websites is whether there are any legal repercussions in addition to potential breach of contract when scraping is inconsistent with website policies. Website operators have tried to use federal anti-hacking law—in particular the Computer Fraud and Abuse Act—to add teeth to their terms of use. The CFAA bars any "unauthorized" access to any "protected computer," which courts have said means essentially any machine connected to the internet. The most high-profile CFAA prosecution in recent years was brought against the free culture activist Aaron Swartz, who downloaded millions of research articles from JSTOR by circumventing security measures at MIT. Federal prosecutors charged him criminally for violating the CFAA, but were roundly criticized (along with JSTOR and MIT) for their aggressive pursuit of the case. Nevertheless, website operators have argued that any access to a site that exceeds the site's terms of use is "unauthorized," which should trigger CFAA liability.

Luckily, the clear trend in the courts in recent years has been to reject this argument, at least for public websites. Two recent

cases illustrate the point. In hiQ Labs v. LinkedIn, the data analytics company hiQ was accused of violating the CFAA by scraping public LinkedIn profiles after being ordered directly by LinkedIn to cease and desist from scraping. The Ninth Circuit ruled that "authorization is only required for password-protected sites or sites that otherwise prevent the general public from viewing the information." The case has been appealed to the Supreme Court, which hasn't yet agreed to hear it as of the time of this writing.

In Sandvig v. Barr, the ACLU brought a challenge to the CFAA on behalf of journalists and researchers who planned to use scraping as well as fake profiles and other deceptive practices to probe whether employment websites were discriminating against some users. This is a well-established way for journalists and investigators to uncover discrimination, but the terms of use of these sites prohibit providing false information. Can site proprietors use federal anti-hacking laws to insulate themselves from discrimination probes simply by changing their terms of use? Citing hiQ, the district court found that CFAA does not apply to scraping public websites (among other behaviors), and should only apply when a user bypasses an authentication mechanism, such as a password restriction, designed to ensure that only certain, authorized individuals have access to the site.

Use case: The Twitter API

The Twitter Developer policy, agreement, and terms, which govern access to data via the Twitter API, are a good example of a robust, enforceable contract governing a commonly-used source of research data. The Twitter API makes it easy to retrieve massive amounts of data from the Twitter ecosystem, but Twitter tightly regulates how that data can be used and, especially, how it can be shared. The Twitter API Terms create a strong, enforceable contract by ensuring that anyone who participates is required to clearly

signal their assent, and only permitting access to those who have created an account and assented. Twitter makes special allowances for scholarly use, but even academics are prohibited from sharing large corpora of full-text tweets. The detailed provisions in the Twitter API, including distinctions between "Tweet IDs" and full-text content, warrant a close read by any researcher working with the API. It's clear that Twitter takes these terms seriously, and violating them could land you in hot water with the company, a political problem that could be very damaging for a researcher who relies on Twitter data for their work.

Use case: Digitized library materials

Even material digitized from library collections—even public domain material!—can be governed by tricky terms of service. For example, much of the digitized collection in the HathiTrust corpus was created in partnership with Google, and limitations on reuse were part of that arrangement. Accordingly, HathiTrust (and member libraries) uses an Access and Use policy to ensure that users don't do anything that would place them in breach of their agreement with Google (or otherwise create liability for HathiTrust or its members). Additional terms of use govern the HathiTrust Research Center's TDM tools. These terms are designed to ensure that HathiTrust and its users remain within the bounds of what fair use permits.

Another example of a context where library materials may be governed by terms of use is collections digitized in partnership with a vendor like Adam Matthew or ProQuest. It is very common for these materials to be in the public domain, but because they are rare and may not exist in digital form anywhere else, it's possible to keep them behind paywalls and monetize access. To make that model work, vendors typically require users to agree not to download collections in bulk, or share them publicly, among other things.

Some libraries, museums, and special collections impose their own terms of use on materials they post online. Sometimes the goal of these terms is just to ensure that the library or archives receives credit as the source of collections material. Other times, the institution is trying to guard against liability (or political embarrassment) for itself by ensuring users don't do anything untoward, or at least documenting that it took steps to warn or constrain users. As libraries move to make their collections more accessible and useful online, more and more are removing all restrictions on public domain materials.

Beyond the terms of the license

So far you've learned how licenses work as contracts, and you've seen some different kinds of licenses you may encounter in the wild. You know that if you're accessing content subject to a license agreement, the terms of that license may affect your ability to do TDM research, even though copyright itself is TDM-friendly, thanks to fair use. Now we're going to look at some of the legal questions you can ask about a license, other than "What's in it?" These questions include:

- Am I bound?
- How does this license affect fair use?
- What happens if I breach?
- What on Earth is "trespass to chattels"?
- And finally, how to manage risk.

Bound by (contract) law: Privity

The word for someone bound by a contract is "privity"—if you're "in

privity" with the other parties to a contract, you're bound by it. If not, you're not bound. How do you know if you're "in privity"?

As you learned at the beginning of this series, a contract requires both offer and acceptance. And to accept a contract, you need adequate notice of its terms.

If a contract mechanism fails, you won't be in privity. With non-negotiable contracts, especially online and digital ones, there is still substantial controversy about when and how these agreements can bind users. Some "browsewrap" licenses (where the terms of an agreement are linked from a notice on a website, often in small print at the bottom of the page) have been ruled unenforceable in court because users didn't have adequate notice of the terms, or a meaningful opportunity to affirmatively accept (or reject) them.

Other contexts where a user may not be bound by a contract include "downstream" users of resources subject to license. Consider a second-hand user who obtains data not directly from the publisher but through a colleague or intermediary. It seems unlikely that someone in that scenario can be bound by terms they never saw and never had any opportunity to accept. Similarly, someone who acquires a copy of a work on the second-hand market—used software or other media, for example—may never be presented with adequate opportunity to accept the relevant terms.

Licenses and fair use (and other user rights)

Some people who work with licensed materials, including lawyers (unfortunately), come to believe that the license is all that matters when it comes to figuring out whether and how licensed collections can be used. A license is "private law" that the parties make for themselves, after all, and the parties can (and often do) agree to abridge the default legal rights they bring to the table, as part of the bargain. If a contract is a legally enforceable promise, it's easy to see how someone could promise not to exercise fair use, for

example. But depending on the contract, you might NOT have made that promise, in which case, fair use (or another default legal right) will survive.

Instead of thinking of the presence of a contract as necessarily nullifying fair use, you should imagine contract law and fair use rights as separate sources of authority. You can seek permission (a license) to use a covered work, OR you can exercise your own rights under the law. If the copyright holder withholds permission, that doesn't necessarily undermine fair use. Indeed, it had better not, because fair use JUST IS the right to make certain uses without permission. Whether fair use survives a license will depend on the specifics of the contract.

Here are some common types of provisions that can occur in license agreements, and their likely effects on fair use. As you can see, far from always nullifying fair use, there are many circumstances in which fair use survives a license.

Does Fair Use Survive?

Language type	Example	Fair use consequence
Clear prohibition	*"User shall not…" or "User shall not...without additional permission"*	License trumps
Limited license	*"This license is for personal use only…"*	Fair use survives
Silence (nothing in the license explicitly or implicitly addresses your use)	...	Fair use survives
Savings clause	*"Notwithstanding any other provision in this contract, nothing shall bar lawful uses...."*	Fair use survives
Ambiguity	...	Fair use favored

License language and its effect on fair use.

Language of clear prohibition or a promise not to engage in certain uses is most likely sufficient to surrender fair use rights. An example of clearly prohibitory language is "User agrees not to..." or "User

shall not…" This is a promise by the user not to exercise her fair use rights. Licensors commonly use this kind of language to ensure users do not engage in bulk downloading or redistribution.

Language describing the limits of a license, such as a statement that a particular license is "for XYZ use only," (e.g., "for personal use only"), should be read to leave fair use intact. That language tells you how far the license goes, but it does not tell you that you may not rely on fair use to go further. It may be that the licensor would be unpleasantly surprised by uses that exceed the license, and you may factor that into your risk calculus. However, fair use is by definition a use that the rights holder cannot control simply by withholding their consent.

Contractual silence about a particular fair use activity should also generally leave fair use rights intact, by the same logic. But be careful: if you promise not to do certain things that are necessary predicates to your fair use (e.g., large-scale downloading from a database), that promise will effectively prevent you from engaging in fair use.

The best case scenario is a fair use "savings clause," which is increasingly popular as a strategy for libraries negotiating licenses. These clauses will typically say something quite broad, like, "Nothing in this agreement shall bar users from making lawful/fair uses of licensed materials." An agreement with this kind of clear, broad savings language lets you ignore contrary language elsewhere in the agreement as long as your use is otherwise lawful and fair.

When a contract is ambiguous, there are several reasons a court or other interpreter might favor fair use. First, fair use is a right with constitutional underpinnings; waiver of such rights must typically be clear and unambiguous. Second, contracts, especially non-negotiated ones, are typically interpreted "against" the author of the agreement. This is because these contracts place so much power in the hands of the contract drafter, courts are wary of permitting them to use ambiguity to their advantage. Instead, they force licensors to be as clear as possible to place other parties on

adequate notice of the terms, or else risk losing any dispute over the terms' meaning.

The stakes: Remedies and consequences of breach

Remedies for breach of contract are typically much less severe than the toughest copyright penalties. Licenses present a mix of copyright and contract issues, and violating a license can trigger copyright liability. But remember: failing to abide by a license isn't copyright infringement unless your use requires a license. In other words, if your use is a fair use, then breaching a contract is only a breach of contract, and nothing more.

The most likely negative outcome is one the licensor can impose unilaterally on your institution: shutting off access to the resource. Licensors don't have to go to a court to enforce the terms privately by terminating access in this way. And because some TDM research can resemble a serious security breach, vendors may be more likely to quickly shut down access in response to unexpected TDM-related activity. If your institution disagrees with the vendor, they could threaten to sue the vendor to get access restored, but that's an expensive proposition. The more likely outcome is that you and your institution will have to negotiate with the vendor to have access restored. In the meantime, other researchers who need access to the resource will be frustrated.

Trespass to chattels, or, why you should scrape nicely

One last issue to consider, especially when scraping public websites, is trespass to chattels. Trespass may be more familiar in the context

of land, but trespass to chattels is unreasonable interference with the ordinary use of someone's personal property.

A paradigm case of trespass to chattels online is a DDOS attack, which barrages a server with so many inquiries that the server becomes unusable for its ordinary purpose. Automated scraping or web harvesting activity could trigger a trespass to chattels claim if it took place in a time or manner that interfered with the vendor's ordinary use of the server. Event promoters like Ticketmaster have brought trespass claims successfully against scalpers who overburdened their servers by using bots to buy tickets.

The best way to avoid this kind of claim is to be polite when you scrape. Don't hit servers hard, especially during normal business hours.

Risk management

How can you lower the likelihood of something going wrong, and how can you lower the stakes and reduce the impact in case something does go wrong?

One thing to consider is reaching out to the copyright holder/licensor and getting additional or more specific permissions. Experiences diverge wildly, but vendors are increasingly familiar with TDM and may well be amenable to negotiating specific terms to permit it, even if their standard contract does not. As you may have learned in the copyright chapter of this book, being told "no" doesn't hurt your fair use argument—and may even help you.

Another way of controlling risk is to be polite in your use of licensed resources. As we mentioned in discussing trespass to chattels, a lot of good will can be won by scraping, downloading, or otherwise accessing content in ways that don't interfere with the ordinary use of a licensed resource. Ill will and risk, however, go up quickly when your TDM-related activity looks like a security breach or piracy.

Finally, be available and responsive when folks have concerns. If you share your data, include a way to be in touch with you. If someone reaches out, don't ignore them. Do what you can to make it easy to channel any objections or concerns quickly and easily into a low-impact resolution.

Creative ways to work within licensing boundaries

Despite the challenges of navigating the range of licensing issues that ethical TDM researchers need to traverse, many researchers have found creative ways to work within the boundaries of what is allowed that open up more opportunities for ethical research than might be apparent after a first glance at licensing terms. What follows summarizes a recent publication on this topic co-authored by one of the authors of this chapter with the unusual title of "The Trouble with Sharing Your Privates". If you enter that title into Google Scholar there is only one that will come up—it's easy to find.

We start with the standard way of thinking about legal boundaries, of which the main umbrella categories are copyright law and contract (or licensing) law. Copyright law can be thought of as a national-level entity and contract (or licensing) law can be thought of as an organizational-level constraint.

Collaborating with researchers in other places presents a special set of compliance challenges. What happens when you have collaborators in different countries: which set of copyright provisions apply? Or what about collaborators who are in different universities: how to navigate the licensing issues for team members who may be bound by different licensing restrictions? Raising these difficult questions with legal authorities in your campus often results in a discouraging answer, given the scale or institutional risk for such a collaboration. Yet there are a number of ideas that

those legal authorities may be unfamiliar with that might be legally compliant with many licensing restrictions.

Short-term solutions

There are some short-term solutions that might respect legal boundaries. We want to underscore "might" here because every license is different and researchers will have to check which among these might be compliant with the legal boundaries in each particular situation. The first is using non-consumptive or non-expressive research modes. The Hathi Trust Research Center (HTRC) provides extracted feature access to the entire HathiTrust book corpus. HTRC allows for access to extracted features such as entities, sentiment scores, token counts, and verb counts. All of this is pure information that exists apart from expressive uses of text, so working with extracted features violates no expressive use and may be compliant with many licenses.

A second possibility is publishing metadata and extracted features that allow your collaborative team to actually find the full-text content on their own, through their own licensing regime. Metadata for a typical newspaper article includes the title, the author, the date of publication, the source of publication, and so on. Often collaborators in other places can use that metadata to track down where they can get access to the full-text content within their current licensing regime. And sometimes it's easier than that. It is possible to construct Lexis-Nexis metadata into URLs that include unique 16-digit identifiers for specific pieces of Lexis-Nexis content. If a researcher is in an institutional setting with licensing that is compliant with access to that content, dropping that URL into an authenticated web browser will magically reveal the full-text content. Researchers in institutions that don't have proper licensing access will get a "404: File Not Found". This is just one way to share full text data by exchanging only metadata.

A third possibility is providing remote access to compliant computer systems. A researcher might set up a virtualized server that resides in an institution that's bound by its licensing agreements, and ensure that any licensed data always stays on the hard drive of that compliant physical server. What's different in this model is that users can be brought to the data from other countries and other institutions just as easily as users from across your own campus. If the user remotes-in and accesses licensed data that always stays on that server, that is a model that might comply with an otherwise restrictive license.

The fourth possibility is publishing or sharing small validation data sets. Random samples of larger corpora published under fair use provisions (if that would apply) would allow collaborators to develop and refine their algorithms that they want to run on the larger corpus. When they've got their algorithms up to speed and producing the kind of output that they want, they send those algorithms over to the researcher with licensed access to the larger corpus, and in a compliant manner that algorithm could be over the entire corpus. So long as the resulting extracted features or similar output violates terms of copyright or licensing, it should be possible to deliver the resulting set of extracted features back to the originating collaborators.

A last possibility is bringing collaborators from other locations physically to the campus or institution that holds the licensed data in order to work together face to face. Most campus licensing provisions for library materials have a cutout for visiting scholars. Bringing somebody physically to your location often gets them full access temporarily to the same content that any researcher at that location has licensed access to.

Longer-term solutions

There are also some longer-term solutions that the TDM

community would do well to explore. One is building more collaborative open data sets like the Linguistic Data Consortium at Penn that allows—for a very small and reasonable licensing fee—access to full-text data that can be shared across national jurisdictions. There's also Amazon's AWS Common Crawl, which is freely-available web content at very large scale. Licensing might apply even to freely-available data, so it is important to check that such data can be appropriately used for a given research context. Another model is Wikipedia: a lot of content out there can be mined and shared within Wikipedia's permissive licensing.

A second longer-term solution is to advocate both within our institutions and within our professional associations for better data agreements that have clearer terms, that have more expansive allowable uses for research purposes, that give us clearer boundaries so that we can know what we can and what we can't do but that also respect the important need for researchers to have relatively free and broad access to sensitive, in-copyright materials.

A third option is solving a local problem. When researchers want to do something that's outside of what everybody locally already knows how to do legally, they often end up talking with somebody in front of a desk where the answer's going to be "No" because nobody's quite sure exactly who's got the final authority to make the call. Encouraging our campuses to develop a "buck stops here" position—call this a data ombudsperson—who is empowered to make that final decision can simplify the process of getting research done in a timely and efficient manner. A good data ombudsperson would know what you're allowed to do with text and what you're not allowed to do, would know the legal landscape, would understand the licensing, and could calm people down who might be a little bit concerned about what a researcher wants to do. Empowering such positions will open up broader opportunities for scholars and students to do expansive and innovative text data mining research in more reasonable timelines than they might otherwise be up against.

5. Privacy

BETH CATE AND RACHAEL SAMBERG

Introduction

Digital humanities scholars are often surprised that TDM questions seeming to present problems of legal privacy often wind up not being governed by U.S. privacy laws, but by professional or disciplinary ethical norms. Because of the specific scope of U.S. federal privacy laws and the strong privacy exceptions under state privacy laws, when TDM digital humanities researchers face privacy concerns, they are often matters of "privacy," but not "legal privacy."

The Gamergate case study highlights this phenomenon well. In Applying an Ethics of Care to Internet Research: Gamergate and Digital Humanities,[1] authors Suomela et al. overview the Gamergate scandal involving the harassment of women who spoke out on Twitter on the topic of misogyny within video game development culture. The women who shared their views received rape and death threats. In collecting Tweets from the women as well as their harassers, Suomela and team needed to consider whether their analysis and republication of such materials violated posters' privacy. What they discovered was that privacy concerns related to the ethical issue of reamplifying hate messages, but not legal privacy

1. Suomela, T., Chee, F., Berendt, B., & Rockwell, G. (2019). Applying an Ethics of Care to Internet Research: Gamergate and Digital Humanities. Digital Studies/le Champ Numérique, 9(1), 4. DOI: http://doi.org/10.16995/dscn.302

because the voluntary disclosure of personal information—such as in someone's own public postings—waives any legal privacy rights even if the subject content had been protected by laws.

In the next chapter, we'll address the ethical challenges embedded in TDM research. But here, we'll detail the kinds of privacy laws that scholars confront in the U.S., and the very powerful exceptions that often render concerns as ethical rather than legal.

What is "private" under the law?

When we think of privacy law and TDM, we often think about cleaning our data so as not to reveal personal information about individuals. But what personal information is actually protected by privacy law, and what are we allowed to publish? And specifically, how do privacy law challenges come up in the context of text data mining? In other words, what do we mean when we say "privacy"?

In the U.S., and unlike with copyright law which is basically just a matter of federal statute, there are actually multiple sources of privacy law.

Constitutional privacy

First, there's the constitutional right of privacy, which protects personal privacy against unlawful government invasion. Note that the Constitution does not explicitly include the right to privacy, but the Supreme Court has found that it implicitly grants a right to privacy against governmental intrusion — and it does this through the First, Third, Fourth Amendment, and Fifth Amendments. These Constitutional rights to privacy, however, are typically not what we're dealing with in the context of humanities text data mining. If you, a researcher or TDM professional, are doing the work, you

are not a government actor, so you're not likely violating someone's constitutional right to privacy with the research you're doing. You may be violating their privacy rights, but not those privacy rights arising under the Constitution— because the Constitution protects against government intrusions. Largely, individual researchers' TDM research does not invoke Constitutional privacy rights.

Federal statutes

There are also federal statutes—laws made by the U.S. Congress—that provide protections for certain types of information, or certain types of individuals. And there are a bunch of them. Federal privacy statutes include statutes like the Children's Online Privacy Protection Act (COPPA), the Fair Credit Reporting Act, the Family Educational Rights and Privacy Act (FERPA), the Financial Services Modernization Act (Gramm-Leach-Bliley Act), the Health Insurance Portability and Accountability Act (HIPAA), the Privacy Act, the Right to Financial Privacy Act, the Stored Communications Act, and more. These statutes impose obligations on how such data should be collected, managed, and disclosed or not.

Likely you will already be complying with institutional review board requirements for human subjects research, and because you will already be adhering to federal privacy laws when you're using financial, medical, and other federally-protected materials. There is little particularly unique to TDM in this type of research, other than that potential privacy problems are exacerbated by the volume of data you might be collecting. If your data set contains covered information from thousands of individuals, you could violate these statutes at much greater scale than with other types of research. This is why robust research data management plans covering the data for the entire lifecycle of your research are critical.

Because overall, because these federal laws cover very specific types of research not often implicated in digital humanities

research, and because institutional review boards already provide oversight whenever your research does happen to involve federally-protected information, what we want to focus on instead is a third source of privacy law—the one most likely to have particular relevance to humanities TDM research because of the sources of information you may be using.

State statutes or common law (i.e. "torts")

That third source of law is: general privacy laws created by states. These can either be creatures of state statutes, or what is considered "common law"— that is, law derived from the court opinions apart from any statute that might exist. These statutes and common law create what is called a tort cause of action resulting from an unlawful invasion of privacy. Torts are essentially a wrongful act or an infringement of a right (other than under contract) leading to civil legal liability. So a tort is basically a civil (as opposed to criminal) wrong that you could do to someone, and something that's an infringement of some non-contract based right that either statutes or common law have created. For instance, if you have trespassed on someone's property, you may have committed a tort. If you have interfered with their livelihood, you may have committed a tort. If you have defamed someone and caused them harm, you may have committed a tort. These are civil wrongs that infringe various personal rights that people hold.

In the context of privacy, there are typically four torts we need to be aware of as TDM researchers. Now, the existence and recognition of these privacy torts varies by state–making these waters very murky for cross-border research. There will always be questions of which state's law applies, and in turn, what privacy torts are at issue. Typically, tort issues are determined by the local law of the state which has the most significant relationship to the occurrence of the invaston and the parties. But generally speaking anyway, these are

the four privacy torts that most states recognize in some form or another—whether through statute or common law rights.

Although recognition of these four harms goes back much further, the torts were articulated by William Prosser in his California Law Review article titled "Privacy" in 1960, and they include:

1. Intrusion upon seclusion or solitude, or into private affairs;
2. Public disclosure of embarrassing private facts;
3. Painting someone in a person in a false light in the public eye; and
4. Appropriation of name or likeness.

As set forth in the legal encyclopedia *American Jurisprudence*,[2] the rationale behind recognizing these four torts is that:

One has a public persona, exposed and active, and a private persona, guarded and preserved, and the heart of our liberty is choosing which parts of our lives will become public and which parts we hold close…Courts have a unique and essential role in protecting the individual's private life and 'space' from well-intentioned but ultimately oppressive, insulting, degrading, and demeaning intrusions, whether these intrusions come from the omnipresent forces of the state, or from the equally omnipresent and inescapable forces of the market.

To understand the Prosser torts, it's also important to know that the right protected by a tort action for invasion of privacy is a personal right, specific to the individual whose privacy is invaded. In the absence of a state statute providing otherwise, the cause of action is not assignable, and it cannot be maintained by other persons, such as members of the individual's family. This is why, as we'll see, a person's death typically extinguishes their right to privacy. We may feel bad—ethically—about disclosing the private affairs of deceased people, but the deceased people typically do

2. 62A Am. Jur. 2d Privacy § 1

not bear a privacy right in that information anymore under state statutes.

Digging into Prosser torts

So let's talk about what these four torts protect.

1. **Both intrusion upon seclusion and public disclosure of embarrassing private facts** require the invasion of something secret, secluded, or private. For there to be a tort on these grounds, a person must have had an objectively reasonable expectation of seclusion or solitude in the particular invaded place or as to the particular topic or matter intruded upon. In order for a defendant to be considered to have intruded into a place, conversation, or matter as to which the plaintiff has a reasonable expectation of privacy, the defendant must have penetrated some zone of physical or sensory privacy or obtained unwanted access to data by electronic or other covert means, in violation of the law or social norms. A defendant is not liable for invasion of privacy under the theory of intrusion upon seclusion if the plaintiff is already in public view at the time of the alleged invasion. This set up reveals that community standards are often important for gauging privacy invasions. Intrusion into private matters is not binary; there are nuances to societal recognition of expectations of privacy. By the same token, the fact that the privacy one expects in a given setting is not complete or absolute does not render the person's expectation unreasonable as a matter of law. Notably, the law does not recognize a right of privacy in connection with further publication or amplification of information that is already public, or known to many people, or a matter of public record, or otherwise open to the public eye. For a fact to be considered private, someone must demonstrate an actual

expectation that the disclosed fact remain private, and that society would recognize this expectation of privacy as reasonable and be willing to respect it. So again we see that community standards are important for gauging whether a privacy violation has occurred under these first two Prosser torts.

2. **Painting someone in a false light.** This privacy tort is similar to the tort of defamation but there are different standards of proof. You've painted someone in a false light if you've published the information widely (i.e., not to just a single person, as in defamation); the publication identifies the plaintiff; there is an element of fiction or falsity; that falsity would be highly offensive to a reasonable person, and you were at fault in publishing the information.
3. **Appropriation of name or likeness** protects a person's exclusive use of his or her own identity. The phrase "name or likeness" embraces the concept of a person's character. The tort does not protect one's name per se but rather the value associated with that name, and typically only when done for commercial gain. We're instead typically talking about leveraging someone's name or likeness for your personal gain—to try to obtain for yourself the reputation, prestige, social or commercial standing, public interest, or other values of the underlying subject. You're unlikely to have any such intentionality in non-profit research.

So we can see that mostly the two torts you'd be concerned about in the type of TDM research you're doing are the first two Prosser torts: Intrusion upon seclusion, and public disclosure of embarrassing private facts. And further, intrusion upon seclusion requires some kind of invading of someone's space where they have a realm of privacy to capture content; this is possible, but frankly unlikely in digital humanities research. **This leaves "public disclosure of embarrassing private facts" as being the most likely Prosser tort to be at issue in TDM digital humanities research.**

So the question becomes: Now that we know what we do about the common privacy torts (and specifically the public disclosure tort) that can arise in TDM research, what do we need to know about exceptions to that tort when making research choices?

Safeguards supporting TDM research

There are some inherent protections built into the nature of what a plaintiff must show to sustain a claim for Prosser torts that insulate you from some risk. Some protections are in the nature of burden of proof, and others are express exceptions

Regarding burden of proof, typically in order to succeed on a claim for intrusion on seclusion or public disclosure of private facts, under state statute or common law, plaintiffs must usually show:

- That a reasonable person would have been offended or injured (not just that they are hypersensitive).
- In turn, a determination of whether a defendant's actions were reasonable is made by balancing the interests of the plaintiff in protecting his or her privacy from serious invasions with a defendant's interest in pursuing its course of conduct.
- And further, to sustain a claim, a plaintiff must show they actually suffered harm, such as mental distress or embarrassment.

These required showings provide important risk mitigations for researchers, as they are an impediment to lawsuits being filed or moving forward.

Exceptions supporting TDM research

Perhaps more importantly, though there are critical exceptions to various Prosser Privacy torts that are very favorable to TDM researchers:

1. **Public Interest:** When it comes to public disclosure of private facts, the right of privacy is not violated by comment or disclosures as to matters of legitimate public interest. Relatedly, tort liability might also be inconsistent with the free speech and free press provisions of the First Amendment to the U.S. Constitution, as applied to state law through the Fourteenth Amendment. In these cases, courts often have to balance a person's right to keep information private with your First Amendment right to disseminate information to the public. In achieving this balance, courts sometimes look to whether the facts you're seeking to disclose are of legitimate public concern and/or would be highly offensive to a reasonable person.
2. **Death:** As we said earlier, a person's death ends their right of privacy, though not necessarily their commercial right of publicity—that depends on state statute. However, you're likely not doing your research for commercial gain anyway
3. **Unidentifiability:** There are no privacy concerns if the people are not identifiable
4. **Consent:** And finally, if someone has released the info themselves—such as on social media sites—or given you permission, they cannot sustain a privacy tort claim

An approach to mitigating risk

We want to take a moment to highlight here for you a potential

practical approach to integrating consideration of these privacy torts and exceptions into your TDM research

It may come as no surprise to you that the same legal literacies researchers and professionals need to understand to navigate TDM research are critical for libraries to understand in determining what collections or corpora to make available for TDM research to begin with. At the UC Berkeley Library, we have launched what we call a Digital Lifecycle Program through which we digitized certain of our collections and make them available for free online for TDM and other research.

We have to answer the same copyright, contracts, privacy, and ethics questions in making the content available that you have to answer in using and publishing with it. And when it comes to the four privacy torts, we rely on similar exceptions that you as researchers would do. You can see from our own "Responsible Access Workflows" that if the subject matter of the collections is no longer living, or the subject matter is newsworthy or of public interest, from a state tort privacy perspective, digitization can proceed through the remaining workflows. We hope our workflows can be a practical way to help you work through privacy and other questions as your research proceeds.

International intersections

So far we've covered only U.S. law. What about international collaboration or if and how international privacy standards bleed into U.S. research? We know that researchers are not guaranteed to be insulated from international privacy regulation simply because their data collection is conducted within the United States. Data that is collected solely within the US may be produced, say, in France, or created by French citizens. The data may have been originally provided with the expectation and under the terms of use that appropriate local data protections would be followed. Many

of these factors that should be taken into consideration may not be documented or readily accessible to a diligent researcher who inspects information prior to collection.

Ethically, legally, and practically, it is not safe to assume that the US definition of privacy is the sole relevant consideration. The contexts in which individuals share information online should play an important role in the sharing and use of information—even if U.S. or state privacy law doesn't cover it.

GDPR

Our guidance for Building LLTDM focuses mainly on U.S. law, but there are two international intersections that bear some attention regardless. The first is the General Data Protection Regulation, or "GDPR." The GDPR was adopted in April 2016 and became enforceable beginning May 2018, and it deals with the protection of privacy and the collection and management of data. Basically, if a business doesn't process an individual's data in the correct way, it can be fined by the EU regulator.

At the core of GDPR is personal data as defined under European law. This is the type of information that allows a living person to be directly or indirectly identified from data that's available, and it is much broader than under U.S. law. Personal data for purposes of GDPR can include something obvious, such as a person's name, location data, or an online username, or it can be something that may be less apparent, such as an IP address. There are also a few special categories of personal data that are given greater protections, including information about racial or ethnic origin, political opinions, religious beliefs, membership of trade unions, genetic and biometric data, health information and data around a person's sex life or orientation.

The GDPR aims to give individuals better control over their personal data. It enacts technical measures that dictate how

businesses and other entities process personal data of EU citizens. Businesses and data controllers are required to enable safeguards to protect user data so that datasets are not publicly available by default, and can't be used to identify subjects.

Even though GDPR is focused on the protection of EU citizens, it can also apply to entities that are based outside of the EU. So, if a business located in the US does business or has users in the EU, then the GDPR could apply to it. In turn, TDM researchers should care about regulations such as the GDPR because social media companies and other organizations that provide products and services to EU citizens is directly affected by these data protection rules.

First, let's take a brief look at how the GDPR applies to data processors. These processors must follow seven protection and accountability principles when dealing with personal data.

1. Processing must be lawful, fair, and transparent to the data subject.
2. Processing must only be for the legitimate purposes specified explicitly to the data subject at the time of collection.
3. It should collect and process only as much data as absolutely necessary for the purposes specified.
4. Processors must keep personal data accurate and up to date.
5. Processors may only store personally identifying data for as long as necessary for the specified purpose.
6. Processing must be done in such a way as to ensure appropriate security, integrity, and confidentiality of the data.
7. And finally, the data controller is responsible for being able to demonstrate GDPR compliance with all of these principles.

Next, let's briefly look at the rights that must be provided to the individuals who are subject of the data collection. They have:

1. The right to be informed
2. The right of access

3. The right to rectification
4. The right to erasure
5. The right to restrict processing
6. The right to data portability
7. The right to object
8. And other rights in relation to automated decision making and profiling.

As we saw from the previous list, one of the user rights is the right to erasure, otherwise known as "the right to be forgotten." Article 17 of the GDPR states, "The data subject shall have the right to obtain from the controller the erasure of personal data concerning him or her without undue delay and the controller shall have the obligation to erase personal data without undue delay." This right can be invoked when a particular situation arises. Some of these include:

- When the personal data are no longer necessary in relation to the purposes for which they were collected or processed;
- When the data subject withdraws consent;
- When the personal data have been unlawfully processed;
- And when the personal data must be erased to comply with a legal obligation in the EU or a Member State law
- As well as a few other reasons.

We can see that under the GDPR there are powerful mechanisms for the protection of personal data, and also ways that users can demand that personal data be redacted from the holdings of data processors.

So, what does this mean for you as a TDM researcher? How could complicated regulations like GDPR affect the utility of particular data sets for research? If a dataset was thought to have been processed appropriately and a researcher wishes to use it to conduct TDM, what are the effects later if the dataset begins to develop holes since some of the information has been removed due

to the right to be forgotten, or another redaction? It's clear from reading parts of the GDPR that if personal data are being processed for scientific research purposes, the regulation indeed applies to that processing.

But, as under U.S. law, there are important limitations and exceptions to the rules that can provide a safety valve for particular types of activities. For example, we were just talking about Article 17, the right to be forgotten. This right is not an absolute user right. Article 17, as well as Article 89 delve more deeply into the safeguards relating to processing for archiving purposes that are in the public interest, as well as scientific, historical, and statistical research purposes.

These safeguards say that the GDPR provisions will not apply when certain circumstances arise. For example,

- for exercising the right of freedom of expression and information;
- for reasons of public interest in the area of public health;
- for archiving purposes related to scientific, historical, and statistical research,
- for the establishment, exercise or defence of legal claims
- And for a few other reasons.

So, while GDPR has some strong protections for privacy rights of EU citizens, it also has some strong limitations and exceptions that support applicable research, including text and data mining. These limitations can give TDM researchers some flexibility in conducting their research without violating the law.

Chapter summary

We've seen the fairly circumscribed intersections between state tort "legal privacy" and digital humanities TDM research. Those

risk junctures are so limited largely because of important research and public-interest related exceptions to state privacy law. But that doesn't mean that we feel great about collecting, analyzing, and disseminating this content even if it is not technically "private" from a legal perspective. In the next chapter, we'll address privacy from an ethical perspective.

6. Ethics

STACY REARDON; RACHAEL SAMBERG; AND TIMOTHY VOLLMER

An overview of ethical considerations in TDM

How can digital humanities researchers incorporate an ethical framework into text data mining? Privacy protections, which are primarily legal mechanisms, will only get us so far. As Todd Suomela and colleagues wrote in Applying an Ethics of Care to Internet Research: Gamergate and Digital Humanities,[1]

> "Humanists are not used to thinking ethically about research subjects because we mostly deal with either subjects who are dead or subjects that are public figures like authors, politicians or other humanists, whose roles and activities are open to scrutiny and debate. The real or potential harms inflicted by research methods on these subjects are often intangible and hard to measure."

We will explore two key ethics questions by surfacing theory through case studies and scholarly literature.

First, we learned in the chapter on privacy that when "public" data is being collected and republished for TDM purposes, it isn't

1. Suomela, T., Chee, F., Berendt, B., & Rockwell, G. (2019). Applying an Ethics of Care to Internet Research: Gamergate and Digital Humanities. Digital Studies/le Champ Numérique, 9(1), 4. DOI: http://doi.org/10.16995/dscn.302

necessarily protected by privacy statutes. So, in the absence of privacy law requirements, to what degree of care should we treat TDM data?

Second, if the current regulatory framework for research involving human subjects is set up to protect privacy, how do we know when to impose an ethical framework? And when and how do we balance the application of a framework with truth-seeking, the public interest, and free expression?

Public, but sensitive

So what do we do with data that is not technically private, but we feel might be sensitive?

As we saw in the Gamergate research mentioned above, the bringing together of public social media messages can serve as a signal boost for hate messages and misinformation. The mere act of compiling these Tweets makes these messages and the targets of the messages more easily discoverable. Yet in many cases, there is nothing private in this data, at least as far as privacy is defined by state and federal laws. But aggregating these hateful messages can also boost the signal concerning views about which an author later has changed their mind. So we have to ask, what is our responsibility as researchers?

Second, should we somehow account for the fact that creators of data might not understand what is protected by privacy law, but might think that data they make available online should essentially still be treated as private? After all, it has been repeatedly demonstrated that the average user's expectations of "privacy" don't necessarily match with what the law says. So do researchers bear an obligation to adhere to social norms or ethics to protect it?

Mismatched expectations are a particular problem when data crosses international boundaries. Many big data research initiatives are international, and protections vary substantially depending on

which national data protection regulation applies. Research subjects may believe that the regulations of their home country protect their personal data, when in fact the requirements of another jurisdiction could apply once the data crosses a border.

And third, how do we approach secondary uses of data that are not intended or predicted by their creators? For example, novelists did not expect for their words to be converted into data. But as Effy Vayena, et al. write in Elements of a New Ethical Framework for Big Data Research,[2] many individuals do not understand the permissible secondary uses of information deemed to be public. In addition, website terms of service do not necessarily help inform people about potential secondary uses. So even setting aside the fact that authors may have had a different expectation of their audience or the ultimate uses of their writing, they also might not realize what they consented to in posting to a public platform.

2. Effy Vayena et al., Elements of a New Ethical Framework for Big Data Research, 72 Wash. & Lee L. Rev. Online 420 (2016), https://scholarlycommons.law.wlu.edu/wlulr-online/vol72/iss3/5

"Please Rob Me" website screenshot

Take the Please Rob Me[3] project as an example of all these harms. Please Rob Me collected information about people's locations from both FourSquare and other social media posts. In 2010, the website demonstrated how users can inadvertently share information that compromises the security of their home by aggregating public tweets from users. The content in the posts suggested that the user was not at home. In this case, the purpose of the website was to raise awareness of the potential for real-world harms, but it is easy to see how this concept could be exploited by bad actors. Even though the information about where people were at the time was public, people's expectation of "privacy" were colored by how obscure they viewed their social media account to be. If another

3. Please Rob Me. (2010). https://pleaserobme.com/

individual mines Twitter accounts for a certain type of information and aggregates and links the information to the accounts of these users, then search cost has been dramatically reduced. The question arises: How should we account for this in TDM research design and publication?

Decontextualization

One contour to consider in deciding how to protect sensitive information is the notion of decontextualization. Below is a photo an Instagrammer took at a restaurant when she ordered avocado toast. The picture and the accompanying article are meant to be a tongue-in-cheek commentary on the value of contemporary restaurant cuisine. But if we were to only look at the photo itself, which includes a piece of toast, half an avocado, lime, and cheese placed separately on a platter, what do we need to be able to understand that this Instagram post is supposed to represent avocado toast, rather than its deconstructed parts?

"Deconstructed" avocado toast.

This image highlights part of the problem with the use of public yet sensitive data: the use of the information for research purposes is potentially stripped of important context or narrative. Not only can this cause personal harm to the author, but in some cases can perpetuate harm to historically marginalized populations. As Kimberly Christen has explained, discovery replays a colonial paradigm, where content is imagined as unhinged from peoples and cultures and free for the taking.

She writes,

> "One can quite easily get content from a Google image search, scrape it from a website, or download it from an academic digital archive. The process is imagined as a neutral act—one of taking something that is already offered up for consumption. But this notion of 'data mining' offers a telling example of how colonial legacies of collecting physical materials from local places and peoples are grafted onto

> digital content. Content is imagined as open, reusable, [and] disconnected from communities, individuals, or families who may have intimate ties to the materials."[4]

Again, the law does not stipulate a way to account for decontextualization.

Structural racism and power imbalances

In addition to decontextualization, another factor to consider in deciding how to protect sensitive information in TDM research is the unequal power structures that enabled its creation or collection. Here, we can observe that data collectors and researchers may be in a greater position of power than the data generators—that is, the people who actually create the information, or from whom the information is collected.

This is a problem for a consent-based ethics framework because underprivileged groups may lack either the knowledge of how information about them will be used, or the ability to intervene in that usage. The World Intellectual Property Organization, or WIPO for short, has tried to develop international frameworks to protect communities not just from having their traditional knowledge exploited, but also to protect them from overstudy and from not receiving the benefits of the research in some meaningful way.[5]

4. Christen, K. (2018). Relationships, Not Records: Digital Heritage and the Ethics of Sharing Indigenous Knowledge Online. In The Routledge Companion to Media Studies and Digital Humanities (1st Edition, pp. 403–412). Routledge. https://www.kimchristen.com/wp-content/uploads/2018/05/41christenKimberly.pdf
5. See the WIPO policy subject area of Traditional

There is a tension, here, though, in that intellectual property laws and human rights frameworks are focused on individual rights, and not group rights. In an article, legal scholar Ruth Okediji explains that absent a fundamental shift, the current types of rules will not facilitate realization of the economic, social, and cultural benefits envisaged and guaranteed by group rights. And without a move toward group rights, it is not really possible for marginalized communities to have real freedom to create, use, and enjoy knowledge assets. Okediji argues that a move toward group rights "strip[s] away any pretense of neutrality and permit[s] scrutiny of, or legal challenges to, private laws with distributive implications that undercut the ideals of human progress and development."[6]

In the absence of a group rights legal framework, we exist in a universe of determining whether and how to seek individual consent for research. Now we'll turn to what a consent-based ethics framework means for TDM research.

Knowledge (TK), which is "a living body of knowledge passed on from generation to generation within a community. It often forms part of a people's cultural and spiritual identity. WIPO's program on TK also addresses traditional cultural expressions (TCEs) and genetic resources (GRs)." https://www.wipo.int/tk/en/

6. Okediji, Ruth, Does Intellectual Property Need Human Rights? (June 25, 2018). New York University Journal of International Law and Politics (JILP), Vol. 50, No. 1, 2018, Harvard Public Law Working Paper No. 18-46, Available at SSRN: https://ssrn.com/abstract=3202478

The research consent framework

As explained in the previous section, to answer some of our questions about whether and how to protect sensitive but not private content, we have to dig into the research consent framework.

Let's imagine that our TDM researchers want to feel like they're doing the right thing by obtaining consent from the creators of the TDM content they're using. And let's also imagine that our TDM researchers are in luck, because they're mining data from a platform like Twitter, through which content creators have already expressly consented to their content being used downstream—merely by using the site.

As Vayena et al 2016 found, even if consent is given for re-use in the terms of service (TOS) for a social media site, because the details are often buried within the lengthy text, users are likely unaware that they have consented to human subjects research through their use of a mobile or social networking platform alone. For instance, Twitter's TOS permits individuals to distribute, Retweet, promote or republish tweets on other media and services.[7]

7. "By submitting, posting or displaying Content on or through the Services, you grant us a worldwide, non-exclusive, royalty-free license (with the right to sublicense) to use, copy, reproduce, process, adapt, modify, publish, transmit, display and distribute such Content in any and all media or distribution methods now known or later developed (for clarity, these rights include, for example, curating, transforming, and translating). This license authorizes us to make your Content available to the rest of the world and to let

Users' reliance on TOS that are vague, complex, and subject to modification without notice leaves users with an incomplete understanding of how their personal information will be used and shared, and arguably fall short of the informed consent requirements intended by research ethics and regulatory frameworks that were developed for clinical research.

Research by legal scholars Kate Crawford and Jason Schultz raises procedural due process considerations. They ask,

> "[H]ow does one give notice and get consent for innumerable and perhaps even yet-to-be-determined queries that one might run that create "personal data"? How does one provide consumers with individual control, context, and accountability over such processes?"

Some uses—including the retention of data for longer than originally envisioned, or for use under a different purpose—may be unforeseen at the time of collection. So, how can it actually be considered consent—and thus how can due process have been served—if one cannot have even conceived of the queries that one

others do the same. You agree that this license includes the right for Twitter to provide, promote, and improve the Services and to make Content submitted to or through the Services available to other companies, organizations or individuals for the syndication, broadcast, distribution, Retweet, promotion or publication of such Content on other media and services, subject to our terms and conditions for such Content use." Twitter Terms of Service. (n.d.). https://twitter.com/en/tos

might run relying on personal data?[8] This is the very reason some scholars are moving away from the consent-based research paradigm, which emerged in the 1970s, to a harms-based paradigm.

From obtaining consent to avoiding harm

Another reason scholars are advocating for a move from a consent framework to one based on treating "harm" is because a consent-based regulatory framework for human subjects research is not conducive to TDM. The current regulatory framework is based on the Common Rule, which outlines ethical rules regarding research involving human subjects.[9] The Common Rule was heavily influenced by the Belmont Report, written in 1979 by the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research.[10] In 1981, with this report as foundational background, the Department of Health and Human Services (HHS) and the Food and Drug Administration (FDA) revised, and made as

8. Kate Crawford & Jason Schultz, Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms, 55 B.C.L. Rev. 93 (2014), http://lawdigitalcommons.bc.edu/bclr/vol55/iss1/4
9. Wikipedia editors. (n.d.). Common Rule. Wikipedia. https://en.wikipedia.org/wiki/Common_Rule
10. National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (1979). Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research. https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/index.html

compatible as possible under their respective statutory authorities, their existing human subjects regulations.

The Common Rule is implemented in Title 45 of the Code of Federal Regulations, pertaining to Public Welfare, and falls within the Department of Health and Human Services.

For all participating federal departments and agencies, the Common Rule outlines the basic provisions around informed consent and assurances of compliance that are subsequently implemented by Institutional Review Boards (IRBs). Human subject research conducted or supported by each federal department/agency also needs to conform to additional regulations of that department or agency.

The Common Rule is porous

The general problem is that the Common Rule—which requires obtaining informed consent—is often inapplicable to common TDM methodology. Typical human subjects research must go through an institutional review board. But many TDM studies in the humanities fall outside the Common Rule's reach where there's no direct interaction with subjects or studies that involve subjects' private, identifiable information. Therefore, institutions are not required to oversee the research at all, even if you may feel there are ethical concerns.

Let's get even more specific about why the informed consent framework creates gaps in research oversight. What qualifies as human subject research—and thus subject to purview of Common Rule and IRB review—is perhaps too narrowly defined because it is concerned with situations where 1) a researcher is collecting information through an intervention or interaction with a subject or 2) identifiable private information is involved. We often don't have either of those two conditions in TDM. First, conducting TDM doesn't necessarily occur with an intervention or interaction with

a human subject. Second, TDM doesn't necessarily involve private information. In addition, de-identification can nominally render the Common Rule/IRB inapplicable, even though doing so has been shown to be an ineffective means of preserving privacy; a research dataset that has been de-identified can in many cases be used in combination with other data to re-identify someone.

Professional guidelines don't necessarily answer the questions about ethical conduct for TDM. The British Psychological Association and Association of Internet Researchers recommend careful consideration of ethical issues when using social media data with particular regard to privacy, but the guidelines do not take an overt stance on the matter of consent for publication.[11]

Even assuming TDM researchers want to apply Common Rule standards and gain informed consent, this isn't always feasible when the data collection is from inordinately large numbers of people.

Also, obtaining consent may not be possible because the very notion of "authorship" may not be aligned with a particular proposed use. As Kimberly Christen explains: "In western settings (and legal contexts), the author is seen as the sole creator of a work [...] In many Indigenous communities, however, the notion of a single creator of a song or author of a narrative is undone by value placed on community production, ancestral creation of stories, or other forms of cultural and artistic content [...] No one person can

11. Webb, H., Jirotka, M., Stahl, B.C., Housley, W., Edwards, A., Williams, M. L., Procter, R., Rana, O. F., & Burnap, P. (2017). The Ethical Challenges of Publishing Twitter Data for Research Dissemination. WebSci '17: Proceedings of the 2017 ACM on Web Science Conference, 339–348. https://doi.org/10.1145/3091478.3091489

or would assert authorship or ownership of the materials."[12] It may flow from this that no one person can give consent.

Given what we know of the consent-based framework, and the gaps in oversight it leaves for much TDM research, how should we proceed with an ethical theory and practice? We'll begin to explore that in the next section.

Ethics theoretical frameworks

As we have discussed, even public data that we wish to use for TDM research might include sensitive information, data decontextualized from the context in which it was created, or data created or used through methods enabled by structural racism or power imbalances. Obtaining explicit consent for such data is one response to negotiating such concerns, even though consent may be unnecessary from a regulatory perspective since the Common Rule typically does not apply to TDM research. Even so, consent may not always be feasible for a variety of reasons. In those cases, we might consider a move from a consent-based ethics framework to a feminist ethics of care.

Here, we briefly consider philosophical underpinnings of three ethical frameworks for conducting research.[13] Imagine you have

12. Christen, K. (2018). Relationships, Not Records: Digital Heritage and the Ethics of Sharing Indigenous Knowledge Online. In The Routledge Companion to Media Studies and Digital Humanities (1st Edition, pp. 403–412). Routledge. https://www.kimchristen.com/wp-content/uploads/2018/05/41christenKimberly.pdf
13. See Lor, P.J., & Britz, J.J (2012). An Ethical Perspective on

the capacity to help someone in need. Furthermore, helping them would not diminish your own capacity. Should you provide this help?

- A **deontologist** would recognize an obligation to help in accordance with a moral rule such as "Do unto others as you would have them do unto you."
- A **virtue ethicist** would act based on the fact that helping the person would be charitable or benevolent.
- And a **utilitarian** will point to the fact that the consequences of doing so will maximize well-being for the greatest number of people.

Each of these normative ethical frameworks places emphasis on moral responsibility and the agency of the individual. Although moral agency assumes free will, power imbalances necessarily complicate the notion of choice or free will. Unequal power structures shape the creation or collection of data. Data collectors and researchers may be in a greater position of power than those of their subjects or of content creators. (This is why, for example, The World Intellectual Property Organization has tried to develop international frameworks to protect communities—not just from having their traditional knowledge exploited, but also to protect them from overstudy and from not receiving the benefits of the

Political-Economic Issues on the Long-Term Preservation of Digital Heritage. Journal of the Association for Information Science and Technology, 63(11), 2153–2164. https://doi.org/10.1002/asi.22725 and Hursthouse, R., Pettigrove, G., & Zalta, E. N. (Ed.) (Winter 2018 Edition). Virtue Ethics. Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/archives/win2018/entries/ethics-virtue/

research in some meaningful way.) This is potentially a problem for TDM research because those represented in the data may not have had free will in how the content was originally created or collected, or they may not have agency in determining how that data is used downstream (such as being used in TDM research). Furthermore, the "distributed morality" of big data—also referred to as "dependent agency"—means that the ethics of data use in a networked framework may be dependent on the morality of other actors in that network, or even on the structure and limitations of the technological infrastructure itself.[14] For these reasons, an individual consent-based framework may not always be enough in guiding ethical decisions.

Ethics of care

An alternative ethics framework might help here. Ethics of Care—also known as Feminist Ethics—is premised on relationships and care as a virtue. This framework recognizes uneven power relationships. Projects adopting an Ethics of Care approach build into their research design an accounting for who possesses power or authority in a given situation.[15] Through its focus on

14. Leonelli, S. (2016). Locating ethics in data science: responsibility and accountability in global and distributed knowledge production systems. Philosophical Transactions of the Royal Society A: Mathematical, Physical & Engineering Sciences, 374(2083), 1–12.
15. Suomela, et al note, "Unlike previous ethical theories that start from the position of an independent rational

relationships, an Ethics of Care framework also enables a progression from accounting for the rights and obligations of individuals, to the rights and obligations of groups.

Like utilitarianism, Ethics of Care endeavors to avoid or at least minimize harm. In "What's the Harm? The Coverage of Ethics and Harm Avoidance in Research Methods Textbooks," Dixon and Quirke[16] identify four categories of harm:

- **Psychological harms** (referring to participants' well-being, and inclusive of things like distress, embarrassment, stress, and betrayal of trust)
- **Physical harms** (this would include physical pain, injury, and death)
- **Legal harms** (this includes legal implications from exposure — imagine here photos of underage drinking, being seen at a protest against a tyrannical government and facing potential action, or depiction as a migrant subjecting one to potential deportation.); and
- **Social harms** (these include damage to relationships, social

subject thinking about how to treat other equally independent rational subjects, the Ethics of Care starts with the real experience of being embedded in relationships with uneven power relations." Suomela, T., Chee, F., Berendt, B., & Rockwell, G. (2019). Applying an Ethics of Care to Internet Research: Gamergate and Digital Humanities. Digital Studies/le Champ Numérique, 9(1), 4. https://doi.org/10.16995/dscn.302.

16. Dixon, S., & Quirke, L. (2017). What's the Harm? The Coverage of Ethics and Harm Avoidance in Research Methods Textbooks. Teaching Sociology, 46(1), 12–24. https://doi.org/10.1177%2F0092055X17711230

> standing, or reputation — and would include impacts on personal and employment relationships through the disclosure of information)

Dixon & Quirke observe that the research ethics textbooks they reviewed failed to treat ethics continually or holistically throughout all stages of the research process. Instead, they approached ethics as a one-time consideration, with a focus on avoiding harm during data collection. However, as they note, "ethical issues permeate and unfold beyond the research design stage and throughout the entire research process." While textbooks may focus on ethics at key moments, such as obtaining informed consent, we might advocate for ethics to be considered throughout the research lifecycle. Moreover, taking into account the network of relationships that compose any project as well as the lifecycle of the project, we might well ask if we need to expand our idea of whose wellbeing beyond that of the research subject should be of concern to us. The Belmont report is set up to protect research subjects, but as we saw from Suomela's Gamergate case study, his project also considered potential harm to the research team. Accordingly, we may wish to apply our ethics framework to all research stakeholders, including researchers and readers.

When considering potential for harm, we might implement a "do no harm" approach or one that seeks to minimize rather than eliminate the potential for harm. The latter may require a risk-benefit analysis of possible harm. In "Elements of a New Ethical Framework for Big Data Research," Vayena et al. (2016)[17] advocate for

17. Vayena, E., Gasser, U., Wood, A., O'Brien, D. R., & Altman, M. (2016). Elements of a New Ethical Framework for Big Data Research. Washington and Lee Law Review Online, 72(3), 420–441. https://scholarlycommons.law.wlu.edu/wlulr-online/vol72/iss3/5

big data researchers and review boards to incorporate systematic risk-benefit assessments. These assessments would evaluate:

- the benefits that would accrue to society as a result of a research activity,
- the intended uses of the data involved,
- the privacy threats and vulnerabilities associated with the research activity,
- and the potential harms to human subjects as a result of the inclusion of their information in the data.

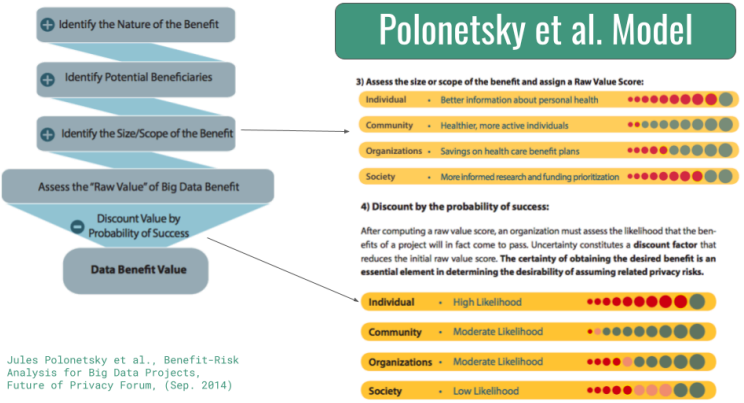
The decision about whether to proceed with the research based on these balanced factors is not binary. Researchers will have to make informed but difficult choices about the best way to proceed. We will explore examples of researchers who attempt to apply a risk-benefit analysis in our next section.

To summarize, we noted that while there is a lack of established best practices when approaching ethical considerations in TDM projects, we can contribute to this evolving discussion. Different ethical frameworks for approaching these issues include deontological, virtue, or utilitarian models, or a feminist Ethics of Care. We might consider different types of harm, such as psychological, physical, legal, and social, and we might consider the different groups in the research lifecycle who could experience such harm, whether those be subjects, fellow researchers, or consumers. Ethical considerations are not just one-time judgments, but extend throughout the research process. Our ethical framework may lead us to adopt an approach that prioritizes doing no harm or one that seeks to weigh harm through a risk-benefit analysis.

Research applications of ethics

We will now review examples of a few research teams who attempted to apply ethical considerations to their TDM projects.

Polonetsky et al.'s (2014) "Benefit-Risk Analysis for Big Data Projects" is one model for operationalizing such a risk-benefit analysis.[18] The Polonetsky model identifies the panoply of benefits of the proposed data project, along with all potential beneficiaries.



Jules Polonetsky et al., Benefit-Risk Analysis for Big Data Projects, Future of Privacy Forum, (Sep. 2014)

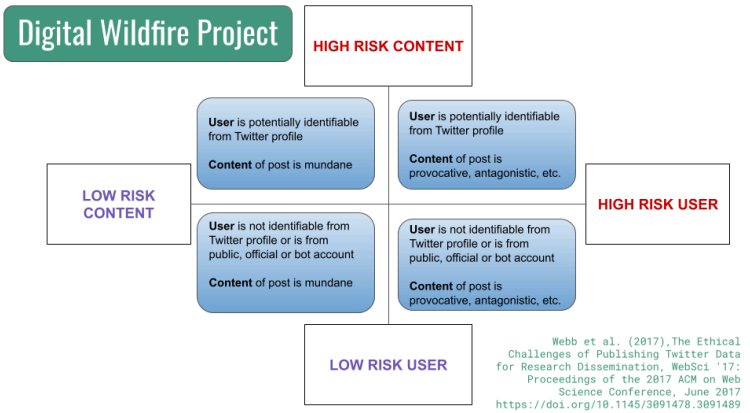
They then determine the size and scope of potential benefit to each of the beneficiaries to weigh the raw value benefit of the project.

18. Polonetsky, J., Tene, O., & Jerome, J. (2014). Benefit-Risk Analysis for Big Data Projects. Future of Privacy Forum. https://dataanalytics.report/Resources/Whitepapers/aa942e84-9174-4dbe-b4cc-911bff14daf8_FPF_DataBenefitAnalysis_FINAL.pdf

This is represented at the top right of the image, with higher value for a stakeholder measured in increasing numbers of red bubbles. That raw value benefit then gets discounted by the probability of success for each of those beneficiaries — the bubbles you see on the lower right. This yields a discounted data benefit value.

Another approach to addressing harm according to a risk-benefit assessment is illustrated by Webb et al. (2017) in their paper on the ethical challenges associated with the Digital Wildfire Project.[19] The Digital Wildfire Project sought to identify opportunities for the responsible governance of digital social spaces by tracking how social media platforms such as Twitter offer the capacity for inflammatory, antagonistic, or provocative digital content to spread on a broad and rapid scale.

19. Webb, H., Jirotka, M., Stahl, B.C., Housley, W., Edwards, A., Williams, M. L., Procter, R., Rana, O. F., & Burnap, P. (2017). The Ethical Challenges of Publishing Twitter Data for Research Dissemination. WebSci '17: Proceedings of the 2017 ACM on Web Science Conference, 339–348. https://doi.org/10.1145/3091478.3091489



Ethics evaluation of the Digital Wildfire Project in Webb et al. (2017), The Ethical Challenges of Publishing Twitter Data for Research Dissemination, WebSci '17: Proceedings of the 2017 ACM on Web Science Conference, June 2017. https://doi.org/10.1145/3091478.3091489

Given that the project included examination of hate speech, there was concern that harm may come to identifiable users who posted content considered to be hateful or inflammatory. The researchers queried whether, from an ethical perspective, they needed to contact the user and solicit informed consent to republish the tweets. Further, there was concern that the re-publication of tweets might cause victims of hate speech harm in addition to the harm they experienced when the content was originally posted.

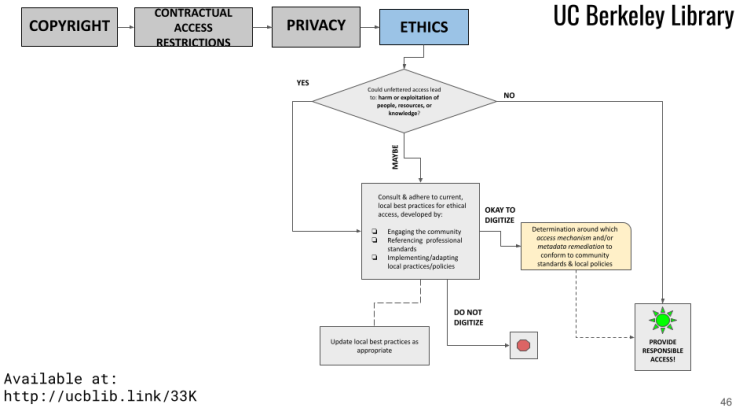
To determine how to proceed, the researchers reviewed relevant guidance, expert opinion, and current practice. There was no consensus from any of these sources. Accordingly, the research team attempted to develop a consistent way to balance value and risks. The graphic depicts the resulting risk grid mapping high/low risk users and high/low risk content.

Nevertheless, they were unable to reach consensus about how to address potential harms or apply a risk-benefit analysis, nor could they draw firm conclusions about best practices with regard to republishing the tweets. Since no one had any answers, they felt the

"best practices" for navigating these concerns might revert to local researcher ethical considerations.

This lack of community guidance could be seen as a challenge, or an opportunity for us all to shape the landscape of legal ethics in this area.

We'd like to highlight one final example of how we at the UC Berkeley Library are trying to implement a harm balancing approach for materials we digitize and make available for TDM projects. Generally speaking, our local ethics best practices would be implemented when providing unfettered access to a collection or materials could potentially lead to harm. They may also be invoked for already digitized content when the Library considers requests pursuant to the community engagement policy, or otherwise becomes aware of situations in which materials in our digitized collections may create harm.



Ethics evaluation within the UC Berkeley Library's Responsible Access Workflows. Available at http://ucblib.link/33K

To develop these policies, the Digital Lifecycle Steering

Committee[20] charged a group of librarians and archivists with expounding upon the definitions and underlying actions that would constitute our local best practices for ethical access to digital content. Our working group conducted a review of relevant literature addressing ethical approaches to digitization in libraries and archives; definitions and treatment of harm and exploitation in law, international policy, and professional literature; empathy and human rights; indigenous knowledge and sovereignty; and the European concept of the right to be forgotten. We found The American Philosophical Society's Protocols for the Treatment of Indigenous Materials to be particularly inspiring and instructive, and much of the format (and some language) for our protocols is based on APS's blueprint.

Our working group expects the local ethics best practices to be finalized in the fall of 2021 following additional community engagement work. As currently conceived, our local ethics best practices ask whether the value to cultural communities, researchers, or the public outweighs the potential for harm or exploitation of people, resources, or knowledge.

1. When referencing *objects*, *materials*, *or resources*: We intend **"harm"** or **"exploitation"** to encompass the following:
    1. economic disadvantage to the interests of a cultural community (such as unfair competition, or commercial appropriation);
    2. violation of customary or national laws, or the established practices of a cultural community; or
    3. risk of looting or defiling of cultural sites or resources.
2. When referencing *people*: We intend **"harm"** or **"exploitation"**

20. https://www.lib.berkeley.edu/about/digital-lifecycle-program-steering-committee

to encompass:

1. A deprivation or violation of, or credible threat to, a person's liberty, body, or well-being.

These definitions were informed by the four types of harms Dixon and Quirke recognized. We then developed a set of principles for how to assess both value and potential harm, similar in intent to what Polonetsky et al recommended, but focused on guidelines rather than formulas. For instance:

- We give added weight to potential value where there is a strong public interest in the material, considering factors like: the content is about public figures; information is about communities, society, or political issues; content is self-authored; the content is composed of government documents or journalistic documents.
- We give added weight to the potential for harm where

    1. Content impacts cultural communities historically disadvantaged by power structures
    2. Material is about the community/creator rather than by the community/creator
    3. Community/creator had or has less ability to control the information
    4. A takedown request was made

This approach makes use of an ethics of care framework that seeks to minimize harm. Less prescriptive, it establishes general guidelines that allow local decision makers to weigh benefit vs. harm, ideally in consultation with community representatives and local experts.

In this section, we have looked at a few examples from research teams who have wrestled with ethical considerations in big data. Polonetsky, et al developed a formula for a risk-benefit analysis based on the scope of benefit to different groups and the likelihood

that each group would receive those benefits. The Digital Wildfire Project attempts a risk-benefit analysis for Twitter data, but instead of proposing their own best practices, they recommend that local practices inform judgment. The UC Berkeley Library is working on ethical guidelines for the provision of digitized materials that ask whether the value to cultural communities, researchers, or the public, outweighs the potential for harm or exploitation of people, resources, or knowledge.

Strategies to address ethical concerns

Over the previous sections, we've come across examples of strategies to approach ethics within TDM research. We characterize these s strategies because there are no actual best practices yet for dealing with sensitive information that is not technically "private" under the law. We hope that you'll begin to think about TDM ethics within your own situation, and start to develop a set of norms and risk management strategies that will allow you to proceed with your research with confidence and relative clarity.

We have loosely organized the following strategies in ascending order of the effort or difficulty in undertaking them. This collection is not meant to be exhaustive; you might have other ideas for strategies that will be applicable in your research situation.

1. Consult journal publications or professional association guidelines. But as discussed above, these may not get you all the way to the question you're trying to answer.
2. Develop local best practices (for instance, you could conduct decision-making within your research group, as the Gamergate research team did,[21] or we did at UC Berkeley for digitizing our collections).[22]
3. You could impose access controls (e.g. user registration to view; publish only data visualizations or extractions), but you'd

need to consider the intersection with any publisher open data requirements.

4. Undertake community engagement to consult with affected populations, and ensure that benefit reverts back to the communities.
5. Seek IRB involvement/approval, even if none is technically required. Of course getting IRB review & approval for research that ordinarily doesn't need approval can slow down the research process (and overwhelm IRBs), so some fundamental structural changes at your institution might be needed.
6. Adopt a new ethics/privacy paradigm (for example, moving from consent-based to harm-avoidance)
    1. Unless you adopt a strict ethics of care and "do no harm" approach, you may need to develop a balancing test that you like. Polonetsky and colleagues have their risk-assessment approach;[23] above we mentioned the UC

21. Suomela, T., Chee, F., Berendt, B., & Rockwell, G. (2019). Applying an Ethics of Care to Internet Research: Gamergate and Digital Humanities. Digital Studies/le Champ Numérique, 9(1), 4. https://doi.org/10.16995/dscn.302
22. UC Berkeley Library. (2020). Ethics Local Practices for Digitization of and Online Access to Collections Materials. https://docs.google.com/document/d/10Ux--7GgrOvoYzTAlbTRtdVAbENYRsCr9HmUNGw9LC4/edit?usp=sharing
23. Polonetsky, J., Tene, O., & Jerome, J. (2014). Benefit-Risk Analysis for Big Data Projects. Future of Privacy Forum. https://dataanalytics.report/Resources/Whitepapers/aa942e84-9174-4dbe-

Berkeley Library's guidelines.

2. Some benefits may not be computable, but efforts to measure value can nevertheless produce useful insights, and the same holds true with big data projects.

Oversight and advocacy

Implementing any of these strategies requires oversight and advocacy in varying degrees. For instance, regulations might need to be changed, or the policies of review boards revised to adopt definitions for terms such as privacy, confidentiality, security, and sensitivity.

As part of the Building LLTDM Institute, we can't necessarily achieve either regulatory change or change to review boards on the spot, but you can bring strategies back to our institutions if you wish to pursue them.

What we can also do as part of this institute is begin considering the development of guidance on research community norms and best practices.

b4cc-911bff14daf8_FPF_DataBenefitAnalysis_FINAL.pdf

PART II

TEACHING THE LITERACIES

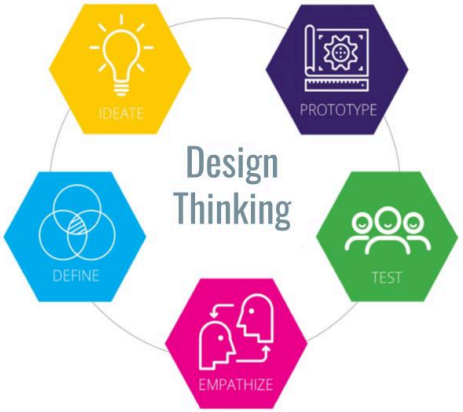
7. Institute development

RACHAEL SAMBERG AND TIMOTHY VOLLMER

This chapter explores how we developed the Institute. First, we explain our overall design thinking approach to the Institute instruction. Second, we discuss our process for recruiting faculty and soliciting applicants. Third, we detail how we selected participants. Fourth, we explain how we financially supported both participants and instructors for taking part in the weeklong Institute. Fifth, we outlined our approach to internal and public communications. Finally, we talk about the pre-Institute tasks required of the participants.

Design thinking approach

To help TDM scholars and digital humanities (DH) professionals build skills tailored for their own DH research agendas, the Institute incorporated a "design thinking" structure reliant upon experiential methodologies. Building LLTDM modeled five stages in design thinking: empathize, define, ideate, prototype, and test.



Elements of design thinking approach.

While we will discuss in detail the day-to-day activities in the next chapter, from a high level we started the first day of the Institute by building out our understanding of participants' experiences with TDM. This helped expand upon what the project team learned from applicants through their applications and questionnaire responses. Day one's "empathize" activities served as an opportunity for participants to get to know each other and to start learning from each other early in our time together. We believed that building trust and common understanding across the cohort led to more robust discussion sessions and collaborative inquiry throughout the Institute.

For days two and three, we cycled iteratively through the "define" and "ideate" phases of the design thinking rubric. The instructors framed and defined a range of different topical TDM issues and literacies through asynchronous videos, and then we used our synchronous time to work through case studies and "putting it together" exercises. This was intended to help strategize how participants could apply their learning to real-word challenges they faced in conducting or supporting TDM research within their home institutions.

On the final day of the Institute, we spent time together prototyping the implementation plans. This was the hands-on time to apply the week's sessions to the participants own work and situated local contexts. We'll be staying connected post-Institute to learn from each other's outcomes as implementation plans evolve.

Recruiting faculty and participants

Faculty

Our Institute project team were composed of legal experts, librarians, faculty, and scholars immersed in digital humanities and research literacies. They were recruited through professional connections and networks. This set of 15 faculty hailed from more than a dozen North American universities and institutions. Faculty contributed to Institute administration and curricular design, and served as instructors during the Institute. Faculty were designated as: humanities researchers ("HR"), librarians ("L"), or legal experts ("LE"). Their real-world roles straddled these boundaries (e.g. some legal experts are also librarians); yet, the divisions ensured that Institute sessions are led by a set of experts who collectively offer a full range of relevant DH expertise.

The project team was led by a Project Director who oversaw curricular design and execution, administrative and operational aspects of Building LLTDM, and also served as one of the instructors during the Institute. The project team was supported by a Project Manager who coordinated design and execution of the Institute, streamlined administrative and operational aspects, and also served as an Institute instructor.

We had a legal expert on call via e-mail during the Institute to

field any questions that instructors were unable to answer in real time.

Participants

We developed a project website to host information about the Institute application process, timeline, and criteria. We advertised the Institute opportunity on the Building LLTDM blog, via digital humanities and library-related email lists, and via social media.

The Institute supported 32 participants, which offered a reasonable instructor-to-attendee ratio to accommodate the highly immersive and discursive aspects of a design thinking approach. We sought an equal number of **digital humanities researchers** and **digital humanities professionals**. We clarified that digital humanities professionals were people like librarians, consultants, and other institutional staff who conduct digital humanities text data mining or aid researchers in their text data mining research. We aimed for the same number of DH research and DH professionals because these two groups were differently situated in their organizations to provide future advocacy and support. We also anticipated the two groups will have mutually beneficial insights and experiences to share. For instance, DH researchers benefitted from LLTDM training that can be both applied to their own research projects and publications, as well as integrated into their teaching and advising, thereby broadening downstream community impact. Conversely, DH professionals are often the first contact point for DH researchers with law-related TDM questions; handle licensing and negotiate access to datasets and digital collections for TDM; and provide training and documentation for DH researchers on workflows and tools. Educating DH professionals enables ongoing Institute impact as they bring the skills they have gained back to their own campuses and professional communities. Finally, we encourage participation from pairs of participants (e.g. one digital

humanities researcher and one professional affiliated with that same institution, organization, or digital humanities project).

We kept the application process as simple as possible. We asked applications to submit two documents via email: 1) a current CV, and 2) a 2-page (maximum) letter of interest that addressed their experience with or interest in: the intersection of text data mining in digital humanities research and the law; your goals for applying knowledge and skills to be acquired at the Institute to your own activities; your goals for sharing knowledge and skills with others at your home institutions/affiliations; and, how you might support the Institute's commitment to diversity and equity.

Participant selection criteria and process

We communicated our selection criteria on the Building LLTDM website. The call for applications was open for two months.

The project team believed that the Institute will work best when it reflects the race and gender demographics of the broader population, and not just those of higher education—and we strived to achieve equity by reflecting these more representative demographics. Additionally, we worked to develop a participant group that is representative of different institution types, research advising and support experience, professional roles, levels of experience with digital humanities text data mining research career stages, and disciplinary perspectives.

The selection process took place over two rounds. First, a subset of the project yeam conducted an initial screen of applications giving preference for the criteria identified below:

- Digital humanities researcher or professional
- Experience working with at least one digital humanities text data mining project
- Articulated interest in the relationship between text data

mining and the law
- Articulated reason for participating in the Institute
- Clear post-Institute goals or ideas for using and sharing knowledge and skills gained
- Application as part of a researcher/professional pair
- Demonstrated commitment to diversity and equity

The project team then performed a second review of applications, making final selections based on the selection criteria and diversity principles identified above.

Since Building LLTDM was made possible through a federal grant (National Endowment for the Humanities Institute for Advanced Topics in the Digital Humanities), we were only able to accept participants based in the United States.

Financial support for participants & instructors

Participant stipends

We offered participant stipends that were distributed to them in advance of the Institute, rather than requiring participants to incur all travel, lodging, and meal expenses and then wait for reimbursement. Our aim was for participants to have zero out-of-pocket costs to attend the Institute. We issued comprehensive stipends because of the social justice implication, as prospective diverse participants may be dissuaded from applying if they know that travel and lodging costs must be charged to a credit card several months in advance of attendance. As a preliminary matter, potential participants may not have credit cards to use for such expenses and, even if they do, they further may not be able to afford

accruing high rates of interest while awaiting reimbursement until after the Institute. To ensure a diverse applicant pool and establish participatory equity for all prospective applicants, we believe it is critical to offer realistic stipends from which participants can cover their costs so they do not have to pay for the Institute out-of-pocket.

We structured the stipends as the equivalent of what we anticipate the participants' actual travel, lodging, and meal expenses will be. We have estimated participant costs based on potential geographic zones from which they would have been traveling. Due to the Covid-19 pandemic, Building LLTDM was conducted entirely online. As the participant stipends were already being processed, and with permission from our NEH program officer, we decided it was fair and efficient to simply deliver the original agreed-upon amount to each participant, even though there was no travel and lodging costs incurred.

Instructor honoraria

We offered instructor honoraria to be distributed to the project team. We awarded honoraria to serve two functions: (1) to recognize the personal (non-work) contributions being made by the project team, and (2) to provide compensation for travel, food, and lodging for the project team members traveling to the Building LLTDM Institute. As with participant stipends, we structured the instructor honoraria as the equivalent of what we anticipate the instructors' actual travel, lodging, and meal expenses will be—plus some compensation to reward their efforts in preparing Institute educational materials, offering instruction during the Institute, and creating the open educational resource following the Institute.

Due to the Covid-19 pandemic, Building LLTDM was conducted entirely online, and the instructors decided to each receive an equal

share of the honoraria allotment, since there was no travel or lodging incurred.

Institute communications

We leveraged several different communication methods, some focusing on internal communications to and between faculty and participants, and some focusing on public communications about the Institute.

Internal communications

For internal communication between faculty to plan the Institute content and delivery preparation, we used a Google Group email.

We set up a separate email group so prospective and accepted participants could ask questions to the LLTDM organizers (those responsible for viewing and answering the email were the Project Director and Project Manager).

For internal communication to participants, we at first used email by cc-ing all participants. For the actual delivery of the online Institute, we relied on a combination of Slack and email. We used slack for announcements, information sharing, and reminding participants and faculty of upcoming sessions. Faculty and participants created additional Slack channels separate from the #general channel to discuss specific TDM research areas, such as #social-media ad #oral-histories. In the weeks leading up to the Institute, we asked both faculty and participants to introduce themselves in an #introductions channel on Slack.

In order to orient faculty about how we would deliver the Institute together, we developed and shared a Faculty Facilitation Guide (we called it the "Faculty Packet"). This Google doc contained

faculty and participant contact information, information about how to use Zoom effectively, and both participant and faculty expectations about contributing and interacting during the online Institute.

We also created a comprehensive guide for participants that we called the "Participant Packet" that was distributed in advance of the Institute. The Participant Packet included:

- Instructions for how to communicate with faculty and other participants
- How to ask questions and receive answers during the Institute
- How to use Zoom during out synchronous sessions
- The Institute code of conduct
- Information about social media and Chatham House Rule

The Participant Packet included a detailed day-by-day agenda for the Institute, including assigned meeting groups of various sizes (plenary, small group), free-write activities, and also links to Zoom rooms and shared notes documents for each session.

Importantly, the Participant Packet contained links to readings and pre-recorded short videos (with transcripts and slides) so that participants could be prepared for the next day's topics.

We viewed the Participant Packet as the one-stop-shop for both participants and faculty to be able to reference throughout the week, as it contained nearly all the information we needed to deliver the Institute.

Public communications

We engaged in some public communications around Building LLTDM. We created a website that contained public-facing information about the Institute, including background information and why the Institute was needed, introduction of the project team,

contact information, and information about how to apply (including process, selection criteria, stipend, logistics, and code of conduct). We included a page that discussed how we would be publishing the content and curriculum from the Institute later as an Open Educational Resource (OER). The website was built on WordPress, so it was easy to include a "news" (essentially a blog) section in order to make announcements, provide updates, and discuss outcomes of Building LLTDM. We asked our Library Communications Team to design a simple logo for the Institute, which we used on the website.

We advertised the LLTDM opportunity through our Library's Office of Scholarly Communication Services Twitter account, and urged other faculty to do the same, either through their institutional or personal social media accounts.

In order to provide easy viewing to all the pre-recorded TDM topical videos, we uploaded them all to the Office of Scholarly Communication Services YouTube account. Viewers can also speed up or slow down the video playback, or turn on closed captions; both features are offered automatically by YouTube. We also created playlists under each topical area (copyright, international copyright, licensing, technological protection measures, and privacy & ethics), as well as a comprehensive playlist containing all the videos.

Code of conduct

Building LLTDM participants and faculty were subject to a code of conduct. We drafted our code of conduct based on examples from several other initiatives. The purpose of including a code of conduct was to provide a positive, inclusive, and harassment-free experience for everyone participating in the Institute. The code of conduct outlined the types of behaviors we were aiming to uphold, and described the types of harassing behaviors that would not be tolerated. The code of conduct included information about how to report an incident.

Institute preparation

Participant questionnaire

About a month before the online Institute, we sent a short questionnaire to participants to complete so that the faculty instructors could learn more about their research and professional practices related to text data mining. The responses were compiled into a PDF and shared with the project team prior to the start of the Institute. It allowed faculty to better understand the participants' real-world research experiences and tailor the online sessions and exercises to properly meet participant expectations and needs.

Background reading

We intentionally kept the amount of preparation for the Institute to a minimum, both because we knew the participants were busy individuals with full time jobs and research responsibilities, and also due to the added pressure and stresses of the Covid-19 pandemic. We set the expectation that we hoped the participants would be able to provide as much undivided attention as they could during the actual week of delivery (of course understanding that there might be necessary interruptions due to family or personal responsibilities because of the remote nature of the workshop). We suggested just two pre-reading to set the stage for our week together online. These readings provided an overview of the TDM legal and policy environment.

1. Matthew Sag, "The New Legal Landscape for Text Mining and Machine Learning," SSRN Electronic Journal, 2019, https://doi.org/10.2139/ssrn.3331606

2. Rachael 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, edited by Sara Benson (Chicago: Association of College and Research Libraries, 2019), https://escholarship.org/uc/item/55j0h74g

8. 4-day Institute delivery

RACHAEL SAMBERG AND TIMOTHY VOLLMER

While the previous chapter discussed the development of Building LLTDM, in this part we discuss the detailed day-to-day activities and delivery of the virtual Institute. We also explain our post-Institute reconvening, and describe how we turned the Institute's literacies and pedagogy into an open educational resource for broad dissemination.

Day by Day Institute delivery

This chapter will explain how we delivered the four day online Institute. As we already mentioned, due to the Covid-19 pandemic, we had to move the Institute from a planned in-person event to a fully remote, online experience. We kept the overall length of the Institute the same as we had planned for the in-person event, although the days were somewhat shortened to take into account the fact that participants and faculty were joining from different time zones, thus we wanted to ensure that we completed by end of normal business hours in the latest time zone (Eastern time). We typically ran the online version of the Institute beginning at 8am Pacific time and ending by 2pm Pacific time. We delivered the Institute in a flipped format (with readings and short videos prepared beforehand), so by ending by 2pm Pacific each day, it would provide time for participants to read and view the following days' content, either later the same day for Pacific or Mountain time zone participants, or the following morning before we began again for Central and Eastern time zone participants.

Day 1

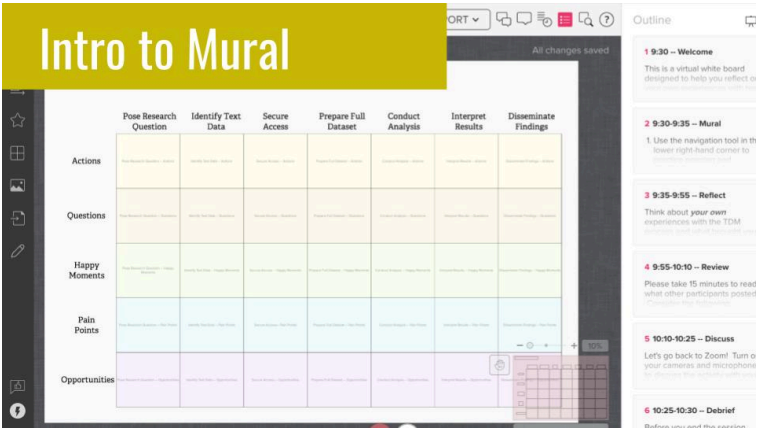
Introductions and setting the stage for the week

The faculty instructors used a primary slide deck throughout the course of the week. Day 1 began with a welcome, logistical information, and framing for the week's activities.

One of the faculty instructors served as a moderator for the Institute. The moderator's role was to observe and synthesize emerging themes from each day. The moderator helped bolster learning outcomes for participants and assist with cross pollination of ideas and themes from across small breakout groups. The moderator observed different groups of discussion sessions and collected individual reflections for sharing at the end of each day.

Empathy building exercise

Participants engaged in a virtual white board exercise designed to help them reflect on their own experiences with text data mining, to build knowledge and understanding among participants, and to surface aspects of divergence and convergence across individual experiences. We used the online "sticky note" software tool called Mural for this journey mapping exercise.



Mural collaborative note-taking.

Free write

Day 1 ended with a free write exercise (the first of three over the course of the week). Free write time wasn't intended for recapturing any notes participants took over the course of the day, but to reflect on the day's sessions and apply them to their personal circumstances: their research interests, institutional culture, team dynamics, etc.

Participants were asked to write for 15 minutes straight without pausing or proofreading. We offered a few prompts to get them writing:

- What did you learn from other participants today about variations in TDM processes and logistical complexities?
- Which pain points highlighted by other participants resonated with you?
- What new questions, concerns, or opportunities emerged during report outs that you didn't capture on the mural board?

At the end of the free write time, participants were asked to email their free write text to our shared faculty email group. Then, a small group of faculty instructors and the moderator reviewed responses each evening and discussed the day's events in preparation for an opening reflection to kick off the next day.

At the end of day 1, faculty and participants were invited to an informal (and optional) "Happy Half-Hour" on Zoom. This time was to socialize, decompress, and answer questions.

Day 2

Report back from moderator on free write themes

At the beginning of day 2, the moderator summarized the themes and learnings that were communicated in the previous day's free writes. This practice reminded participants about the themes discussed in the day before, and tracked progress and accomplishments over the course of the week.

Substantive literacies: Copyright, international copyright, TPMs

On day 2, we began to explore the substantive law and policy literacies for text data mining in the digital humanities. We covered copyright (focusing heavily on U.S. law), copyright in the international/cross-border context, and technological protection measures. As mentioned above, participants were able to watch short pre-recorded videos made by the faculty, as well as view slides and video transcripts.

Participants were asked to read the following articles in advance of day 2:

- Authors Guild v. Google, Inc., 804 F. 3d 202 – Court of Appeals, 2nd Circuit 2015, https://law.justia.com/cases/federal/appellate-courts/ca2/13-4829/13-4829-2015-10-16.html
- Matthew Sag, "The New Legal Landscape for Text Mining and Machine Learning," SSRN Electronic Journal, 2019, https://doi.org/10.2139/ssrn.3331606
- Flynn, Sean and Geiger, Christophe and Quintais, João and Margoni, Thomas and Sag, Matthew and Guibault, L. and Carroll, Michael W., Implementing User Rights for Research in the Field of Artificial Intelligence: A Call for International Action (April 20, 2020). European Intellectual Property Review 2020, Issue 7. Available at SSRN: https://ssrn.com/abstract=3578819

"Putting it together"

After the morning substantive sessions, faculty and participants engaged in a "putting it together" exercise. This activity required individual reading and reflection, as well as small- and medium-sized group discussions, on a pre-prepared TDM scenario.

Free write

Day 2 ended with another free write exercise.

Participants were asked to write for 15 minutes straight without pausing or proofreading. We offered a few prompts to get them writing:

- How do the projects you've worked on, supported, or encountered differ from the scenario you worked on during the Putting it Together session?
- What copyright concerns do you have about accessing data for your own projects? What about publishing it?
- What was your biggest "Ah ha!" moment of the day? What do you still find confusing?

At the end of the free write time, participants were asked to email their free write text to our shared faculty email group. Then, a small group of faculty instructors and the moderator reviewed responses each evening and discussed the day's events in preparation for an opening reflection to kick off the next day.

At the end of day 2, faculty and participants were invited to an informal (and optional) "Happy Half-Hour" on Zoom. This time was to socialize, decompress, and answer questions.

Day 3

Report back from moderator on free-write themes

At the beginning of day 3, the moderator summarized the themes and learnings that were communicated in the previous day's free writes. This practice reminded participants about the themes discussed in the day before, and tracked progress and accomplishments over the course of the week.

Substantive literacies: Licensing, privacy & ethics

On day 3, we explored the substantive law and policy literacies for

text data mining having to do with licensing, privacy, and ethics. Participants were able to watch short pre-recorded videos made by the faculty, as well as view slides and video transcripts.

Participants were asked to read the following articles in advance of day 3:

- California Digital Library 2005 Agreement with Factiva: https://cdlib.org/services-groups/collections/licensed_resources/redacted_licenses/ST_Tier2_Factiva_UCLA_2005_Redacted.pdf
- California Digital Library New Model Agreement: http://ucblib.link/33L
- Nancy Herther, Daniel Dollar, Darby Orcutt, Alicia Wise, and Meg White, "Text and Data Mining Contracts: The Issues and Needs" (2015). Proceedings of the Charleston Library Conference. http://dx.doi.org/10.5703/1288284316233
- Butler, Brandon (2018), "For text- and data-mining, fair use is powerful, but possession is still 9/10 of the law" at http://thetaper.library.virginia.edu/2018/02/28/for-text-and-data-mining-fair-use-is-powerful-but-possession-is-still-9-10-of-the-law-sparc.html
- Suomela, Todd, et al. "Applying an Ethics of Care to Internet Research: Gamergate and Digital Humanities." Digital Studies/Le Champ Numérique, vol. 9, no. 1, Open Library of Humanities, Feb. 2019, p. 4, https://www.digitalstudies.org/articles/10.16995/dscn.302/
- Jules Polonetsky et al., Benefit-Risk Analysis for Big Data Projects, Future of Privacy Forum, (Sep. 2014), https://fpf.org/wp-content/uploads/FPF_DataBenefitAnalysis_FINAL.pdf

"Putting it together"

After the morning substantive sessions, faculty and participants

engaged in a "putting it together" exercise. This activity required individual reading and reflection, as well as small- and medium-sized group discussions, on a pre-prepared TDM scenario.

Free write

Day 3 ended with the final free write exercise.

Participants were asked to write for 15 minutes straight without pausing or proofreading. We offered a few prompts to get them writing:

- What strategies will you use to evaluate the ethical implications of current and future TDM projects?
- What licensing issues surfaced for your own work? Where do you see a path forward and where do you feel stuck?
- What made you feel angry today? What made you feel relieved?

At the end of the free write time, participants were asked to email their free write text to our shared faculty email group. Then, a small group of faculty instructors and the moderator reviewed responses each evening and discussed the day's events in preparation for an opening reflection to kick off the next day.

At the end of day 3, faculty and participants were invited to an informal (and optional) "Happy Half-Hour" on Zoom. This time was to socialize, decompress, and answer questions.

Preparation for implementation mapping discussion

Before Day 4, we asked the participants to read and reflect on

the following questions at three implementation levels: As to their own practice, within their institution, and within their community. "Community" may refer to other digital humanities professionals and researchers with whom you interact, or any relevant broader group of stakeholders.

1. How will you provide guidance to others or integrate the literacies in your own practice? What concrete steps or actions will you take? Are there things that you, your institution, or the broader community should stop doing?
    1. Yourself:
    2. Your institution:
    3. Your community:
2. What challenges might you face as you move forward with implementation of the literacies?
    1. Yourself:
    2. Your institution:
    3. Your community:
3. How would you like to collaborate with other Building LLTDM participants or other DH researchers / professionals to integrate the literacies into DH TDM practice? What would a high level roadmap look like to achieve this vision? What support or funding would you need to make this vision possible?
4. Are there aspects of the current legal landscape that would benefit from community cooperation and advocacy to better address and enable TDM research?

Day 4

Report back from moderator on free-write themes

At the beginning of day 4, the moderator summarized the themes and learnings that were communicated in the previous day's free writes. This practice reminded participants about the themes discussed in the day before, and tracked progress and accomplishments over the course of the week.

Implementation mapping

Faculty and participants reconvened in small groups to discuss the answers to the implementation mapping questions that they'd thought about the night before. These groups worked to identify common themes, next steps, needs, and plans. Later on, we all came together in a final plenary session to share take-aways from the small group discussion. We again used Mural to share virtual "sticky notes" that were viewable by all participants. Finally, we shared stickies of "gratitude" to acknowledge or thank participants, faculty, or recognize a particularly useful or impactful aspect of the Institute.

Participant evaluation

During the next-to-last session, we had participants fill out an evaluation survey. We hoped to get feedback right away while the participants would still have the Institute fresh in their minds, and so they didn't have to respond to an email days or weeks later. We

reminded participants that we would be coming back together in eight months for a meeting to update each other on our progress.

We wrapped the Institute with a short social celebration and goodbye on Zoom.

Institute reconvening & updates

The project team scheduled a check-in meeting with the participants eight months after the completion of the online Institute. The goal of the meeting was to see how participants had been thinking about, performing, or supporting TDM in their home institutions and projects with the LLTDM literacies in mind. In order to get participants thinking about what they wanted to report back to the group based on their experiences in the interim, we asked that they record a 2-minute video and post it on our Slack. In the video, we urged participants (and also faculty, if they had updates) to share their thoughts on the following:

- What have you been thinking about or doing with respect to TDM?
- What's one lasting LLTDM lesson you remember from the Institute?
- What takeaways from the Institute have you been able to implement or share with others?
- What are you still struggling with when it comes to LLTDM?
- What are you proud of with respect to your LLTDM skills?

We asked that each person view all the other posted short videos. This way, they could get up to speed in a relatively short amount of time, and we wouldn't have to spend a lot of time during the meeting itself giving individual updates.

The virtual meeting consisted of a welcome and reflections from the moderator based on the participant and faculty 2-min videos,

small group discussions, and a plenary group exercise to discuss themes that emerged from the smaller discussions.

The last ask of the participants entailed completing a survey that asked about their TDM literacies implementation plans, and also gathered information about the types of resources that would be most useful to include in the open educational resource that would be published.

9. Short instructional sessions

RACHAEL SAMBERG AND TIMOTHY VOLLMER

Not everyone has either the time or the need to host a four-day, intensive extravaganza to teach LLTDM. In this chapter, we present modular opportunities with examples of shorter sessions tailored for different audience needs.

Quick overviews (15-minute sessions)

Why and when?

There are a number of contexts in which a 15-minute overview of LLTDM can be the perfect vehicle by which to introduce core concepts to new audiences. Short overviews work best for attendees who don't need to become experts in the literacies, but wish to (or should!) be aware that the issues exist. As such, we have taught these quick overviews for:

- Students who have been assigned their first TDM projects
- Scholars interested in creating digital research archives
- Librarians who wish to feel more confident advising users about TDM projects

What should you cover?

The main goals of a 15-minute session are simply to help attendees

begin to "issue spot," and learn more about whom to contact if they get stuck or have questions. We typically approach 15-minute sessions as a quick lecture. (Fifteen minutes of an introductory overview of all the literacies just isn't conducive to incorporating an exercise. An exercise is far more feasible if you focus on just one of the literacies.) And note that there is some false advertising here, as these sessions have always run 18-20 minutes despite being billed as 15. We just can't lay enough context for the above takeaway points in only 15 minutes!

We believe the core concepts to establish and instill in a 15-minute session include:

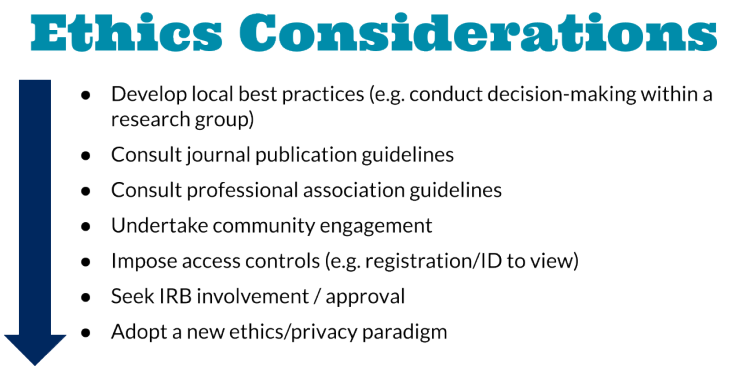
- **Copyright**: It's typically fair use to download or compile copyright-protected content provided you don't break digital locks (DRM), but there are limits on what you can republish or share from what you download or compile. You don't need to worry about fair use and how much you republish at all, though, if you use public domain materials or just facts/ideas.

- **Contracts**: Even if a use is fair under copyright, or if the content is not protected by copyright, there may be a contract that restricts scraping and TDM. Look for fair use savings clauses in applicable license agreements or website terms of use/terms of service. If you're using a library database to download, what matters is the library's license agreement, not the database's generic terms of use online. Finally, consider vendor- or publisher-authorized options like APIs or simply negotiating with the vendor/publisher for what you want.

- **Privacy**: Mining data could violate federal or state privacy laws, but there are important legal exceptions that support TDM research. For instance, state privacy laws (1) often have exceptions for research that is "newsworthy" or of sufficient "public interest," (2) typically don't protect deceased people, and (3) are inapplicable if the subject of the works cannot be

identified. You can consider the applicability of those exceptions or alternatively seek consent from the subjects of the works you're using. Collecting voluntarily-released data from the subject (e.g. a person's public Tweets) does not violate privacy rights, but may present ethical questions.

- **Ethics**: To address ethical concerns, there's a continuum of actions you could consider with increasing degrees of commitment required. Here's a quick example of that spectrum:



Various approaches to addressing ethical considerations in TDM research.

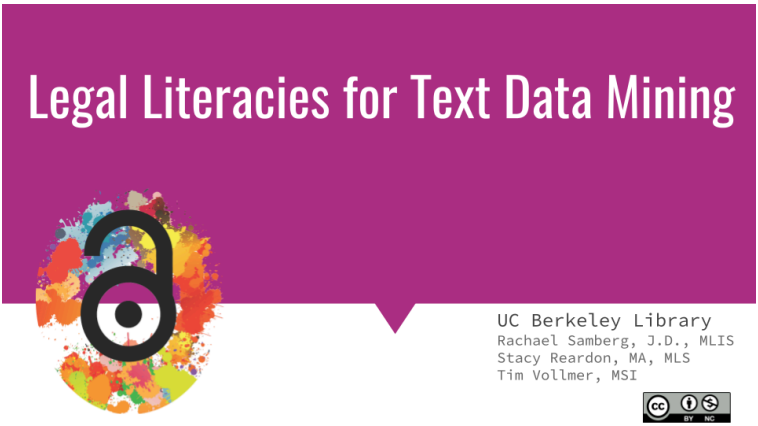
What it looks like

You can check out some of our 15-minute overviews here:

What to know about law and ethics when archiving and mining data...in just 15 minutes! [Video] [Slides + Speaker Notes]

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://berkeley.pressbooks.pub/buildinglltdm/?p=45#oembed-1

Legal literacies for text data mining [Slides + Speaker Notes]



One-shot deep(er) dives (1.5-hour sessions)

If you've got about 1.5 hours for a "one shot" workshop (lecture + exercises), participants can come away with practical, working knowledge of how to implement the literacies for their own projects.

We've successfully run such sessions relying on 45 minutes of lecture plus 15 minutes of questions, followed by a 30-minute exercise—giving participants essential hands-on experience with putting their newly-acquired knowledge to the test.

Why and when?

One-shot workshops are well-suited for graduate students and professional staff who are at the planning stages of or are deeply engaged in supporting TDM research. Catching the interest of scholars before they begin their work can be challenging, but building relationships with digital scholarship centers or labs as well as digital humanities faculty can be essential for bringing attention to the trainings, or even integrating the sessions into required coursework.

What should you cover?

We recommend including all of the takeaways we identified above in the 15-minute sessions. But we also believe it's helpful to provide the following additional context, requiring around 8-10 minutes per topic:

Background

1. Foreground that you're helping people understand what they can do, not telling them that they can't or shouldn't conduct the research
2. Use real-world examples from your practice or scholarly case studies to highlight how many TDM projects intersect with copyright, licensing, privacy, and ethics.
3. In a live session, try to get a sense of the TDM projects with which participants are involved so you can tailor examples as you go to issues arising in participants' own research

Copyright

1. Copyright law grants exclusive rights to original expression for limited periods of time
2. These exclusive rights include reproduction, distribution, display, creation of derivative works, and performance
3. During the protected period of time (currently author's life + 70 years), the author holds these rights exclusively
4. There are exceptions to these exclusive rights that are critical for research and scholarship, and one such exception is fair use
5. Courts have determined that conducting TDM is a fair use, and therefore not a copyright infringement.
6. But that doesn't mean someone can republish the entire copyright-protected corpus they created. While TDM is fair use, republishing the corpus may not be.

Contracts

1. Regardless of whether TDM is fair use, or even if the content you're scraping and analyzing is in the public domain and not protected by copyright at all, there might be other agreements that restrict what you can do with the materials. In other words: Just because TDM is permissible under copyright law doesn't necessarily mean you're free to download, create, and circulate a TDM corpus.
2. This is because there may be a variety of different contracts that supersede what's allowed under copyright law.
3. When you're working with social media or other websites to conduct TDM, you might want to be able to download a large portion of it, or maybe even everything on the site.

It's important to understand that doing so could violate the website's terms. The website's Terms of Use are considered "browse wrap" agreements, meaning you consent to the terms simply by browsing, or viewing, the site.

4. But it's also important to note that these kinds of browse wrap agreements are not always enforceable by a court. Contract issues are questions of an individual state's law, rather than federal law like copyright. Courts in different states may require that users have either actual or constructive notice of the terms of use. This basically means: Should a reasonable person have been aware of the terms based on how the website was presented? Courts that are evaluating whether constructive notice was provided will look to factors like how visible the terms of service were, and whether the users were asked to consent to them. Some courts have simply ruled that browse wrap agreements are indeed enforceable.
5. So what should you know as a general guideline? You should be aware that these terms may exist, and you should make risk calculations accordingly. Often, if you are accessing publicly-available content and downloading it just to scrape—without breaking access barriers to get at the content—then it could potentially be a low risk to violate the terms because it may be hard for the content owner to prove damages.
6. Researchers might also be interested in scraping journal, newspaper, and content databases that are offered by research libraries. When libraries subscribe to these databases, we sign contracts with publishers. If you are accessing material from library databases, then our database agreement applies to you, even if you didn't sign anything yourself.
7. Database licenses can affect researchers' ability to make TDM uses of the material—whether with respect to access

by limiting researchers' right to make downloads, or republishing via restricting circulation of the content.

8. It may be possible to skirt contractual restrictions by using a publisher's application programming interface (API) or negotiating with the publisher to secure the necessary permission.

Privacy

1. There are both federal and state privacy laws that can govern the collection and dissemination of content for TDM research. Often, institutional research boards address federal law applicability since those are more relevant within the context of human subjects research.
2. State privacy laws typically cover what we commonly think of as intrusion and invasion. It's helpful to understand those laws, but perhaps equally helpful to be aware of pertinent exceptions:
    1. The right of privacy is not violated by disclosures of matters of legitimate public interest.
    2. Specifically with respect to public disclosure of private facts, courts also have to balance a person's right to keep information private with your First Amendment right to disseminate information to the public. In achieving this balance, courts sometimes look to whether the facts you're seeking to disclose are of legitimate public concern and/or would be highly offensive to a reasonable person.
    3. When a person dies they lose the common law right of privacy, though not necessarily their commercial right of publicity as to their name or likeness—that depends on state statute. However, you're likely not doing your research for commercial gain anyway, so for all intents

and purposes, if you're mining and disclosing information that would typically be protected by state (as opposed to federal) laws, the state laws usually no longer apply if the subject is deceased.

4. There are no privacy concerns if the people are not identifiable from the information you release.
5. If someone has disclosed the information themselves—such as by posting the content voluntarily on social media sites—or given you permission, they cannot sustain a privacy tort claim.

Ethics

1. There are often questions of ethics that do not fall under privacy, copyright or contract law, but that researchers may still want to consider in their research. This includes information that would be considered "private" under law, but which we (as individuals) may consider to be sensitive in some way.
2. What's unique about ethical concerns in TDM research is that we are bringing together a vast amount of data, in many cases decontextualizing that content from its original source, and making that data available for mining. This can subject individuals to harm, allow for the targeting of disadvantaged communities, or exploit indigineous knowledge, among other risks. In other instances, collecting and mining data may expose cultural heritage sites to looting, or reveal the location of endangered species and subject them to poaching or exploitation. Again, these issues may be present in other types of research; but with TDM, we're looking at exposure at scale.
3. Ethical questions for TDM research can be challenging because there are no legal answers, and TDM researchers

are only beginning to grapple with ethical considerations.

4. So, how do we approach these questions? There's a continuum of actions one could consider with increasing degrees of commitment, which we've excerpted visually above. Researchers may also wish to consider the long-term relationships they hope to build with different communities that they are working with or studying. For now, researchers have to create our own ethical guidelines and seek out guidance from similar projects, professional organizations, publishers, and others.

30-minute exercise

1. In our experience, the real learning in the 1.5-hour workshop comes through the exercise at the end. We recommend dividing participants into groups of 2-4 so that they can talk through the questions together for about 15 minutes before rejoining a plenary discussion. If you're teaching online, having two instructors is helpful so that you can pop in-and-out of breakout rooms.
2. We have found that the groups working on their own can apply basic issue-spotting skills—but when the instructors call everyone back for a plenary discussion of the questions, participants are amazed to discover the many nuances they may have elided. We provide some suggested exercises in our participant packet for the four-day institute. There's no reason these exercises can't be repurposed for 1.5-hour workshops!

What it looks like

Text Data Mining & Publishing [Slides + Speaker Notes] [Exercise]

Text Data Mining & Publishing

UC Berkeley Library
Rachael Samberg, J.D., MLIS
Stacy Reardon, MA, MLS

10. Reflections

RACHAEL SAMBERG AND TIMOTHY VOLLMER

In this chapter, we cover the pedagogical takeaways from the four-day Institute (held June 2020), and reflect on the lasting impact it made eight months later, as evidenced by our observations from the plenary post-Institute check-in.

Design thinking is effective for teaching LLTDM

Participants felt empowered after the Institute to understand the basic contours of the legal literacies for text data mining and applying them to their own work, whether that be developing their own TDM projects, advising DH researchers, or working with TDM issues in libraries and archives. The participants' own words say it best:

- "I can say with confidence that I understand the four literacies better"
- "I really feel that I am coming out with much more both theoretical and practical knowledge than I expected."
- "I will be much more intentional at the outset of any TDM project about working through all of the pertinent literacies in a systematic way…the way the Institute was structured into different literacies provides a repeatable framework to treat potential problems prospectively."
- "I am taking home a lot of new insights from this Institute in combination with a feeling of empowerment that will allow me to reach out to the specialists and directors at my institutions

in order to push for more TDM collaboration and a bolder approach concerning materials and datasets for international cooperation. I know now what the important legal issues are and how to use them to form my arguments and that is more than I could have wished for. Also, the Institute broadened my perspective with regards to issues that I did not have on the radar that much at the beginning and I am looking forward to engaging with these topics in the future, to integrate them into my teaching, and to advocate for them where I can."

The pivot from our initial plan to host an in-person Institute to a virtual one was met with applause. In particular, the participants valued the interactive format fostered by the design thinking model, with different touch points and small group discussions. Again, in their own words:

- "The deliberately thought through breakdown and mix fostered incredibly valuable discussions and I would hope this kind of framework is used as a best practice for future DH institutes of all kinds going forward. Also, thank you for such an amazing virtual experience which I can only imagine took a tremendous amount of work to coordinate and plan with limited time to shift to an entirely different format–I was overjoyed to critically engage with complex subjects and for the chance to get out of my everyday pandemic routines."
- "I found this to be the best example of how to manage hands-on learning in a virtual environment. I think the planning team did a FANTASTIC JOB pivoting to a fully online environment without losing the feel of an in-person intensive."
- "The multi-modal communication (Slack, Mural, Zoom) enabled far more interaction than I anticipated."
- "This is by far the best organized event that I have ever attended. The content was by far the most substantive. The faculty were by far the most engaged. A+ across the board."
- "The flipped learning approach, combined with design learning

> elements, really worked well. The lecture/video materials and reading in particular were well presented and selected, and I really appreciated that we could do that at our own pace. The overall topic of this gathering was well chosen in that it could allow for us to do focused seeking of answers to questions but in a way that had real practical consequences for how we could change the world of TDM research.

We are hopeful that the literacies and methodology developed and shared by the Institute will find a place in broader DH curricula and empower DH researchers to build and analyze their text corpora without fear, thanks to their being more secure in their knowledge of the law.

Lessons for the instructors

The conversations during the Institute and the participant feedback gave us much food for thought. We'd like to expand our commitment to diversity and ensure that the demographics of both faculty and participants reflect those of the broader population, and that the kinds of questions and examples that animate our discussions engage with issues of ethics, equity, and representation.

As we repurpose the Institute training and materials in the future, we will also consider additional ways to emphasize and create discussions around ethics, and perhaps foreground ethics as the first step when thinking through DH projects. We believe questions of ethics loomed large not only because of the critical importance of ethics when addressing data at scale, but also because of the relative absence of guidelines and best practices to help guide us in this area.

We also learned a few specific things that may shape how we approach immersive LLTDM trainings in the future:

Copyright isn't a sticking point (or even that intimidating!)

Questions about using material still under copyright were at the forefront of participants' minds when they entered the Institute, but those concerns evaporated quickly. The copyright portion of the curriculum addressed copyright and the fair use exception extensively. Among others, we discussed the Google Books case, which established that running algorithmic analyses on text was transformative and that using the entirety of books in its corpus was necessary. (One of the authors of a widely-cited amicus brief in the *Authors' Guild v. Google Books* and *HathiTrust* cases was a member of our faculty.) We discussed risk and risk tolerance. Unexpectedly to many, copyright issues turned out to be relatively straightforward, and participants felt empowered to perform analyses on copyrighted materials. One participant said, "I also feel compelled now to do my own research and take advantage of the expansive idea of fair use to examine contemporary, creative works," and another "was mainly relieved that my TDM project was transformative enough to not violate copyright." Rather, the sticking point was how to educate our communities in the possibilities that fair use might allow.

Building a corpus is tough!

Our pre-Institute research and experience indicated that researchers may choose frictionless materials for their corpora, such as materials already in the public domain, or, if they use materials under copyright, they may be unwilling to reveal the process by which they acquired those materials. The former limits the kinds of questions that can be asked, makes certain time periods

easier to study, and may result in bias. The latter makes reproducibility difficult.

The experiences of the participants in the Institute indeed confirmed these challenges. Participants shared their frustrations with finding content and their discomfort with using materials that were under copyright or licensing restrictions. Such challenges limited their work and constituted a major roadblock to their research, one that sometimes exceeded even the technical difficulties of doing the analysis itself. Participants weren't always comfortable sharing how they acquired those materials.

Weave literacies into projects

Another lesson that came up repeatedly was that: We should be building a legal literacies workflow into DH project planning from the very beginning, and refer to it throughout the project lifecycle. Too often, copyright and other legal considerations are unchallenged or brushed aside, to the detriment of our work. This is partly owing to a lack of expertise in these areas or to fear of reprisal. Institute participants suggested ways of addressing these considerations, from trainings, to online documentation, to building legal questions into the project management process for DH work. One participant said, "In our library's center for digital scholarship, we need to develop a better charter/MOU/agreement system for digital projects that will at least touch on data management (DMPs), legal implications (copyright, etc), collaborator expectations, and ethics."

International issues need future institutes

Although we had initially intended to focus mainly on US law, in the

end we realized that international issues are unavoidable given the broad range of humanities research our cohort represented: either scholars are working with materials published under different legal frameworks, or are collaborating with others working in those environments. This obviously complicates the legal picture, so rather than offering clear answers to every question (many of which simply aren't clear), we offered strategies for assessing and mitigating risk. At the same time, we did offer a high-level view of copyright regimes around the world that seemed to be appreciated. Cross-border research collaborations emerged as a clear example of follow-on training that we believe is necessary.

TDM-friendly licenses

Sometimes licenses with publishers, vendors, museums, and other content providers can further restrict uses that would otherwise be allowed under copyright law. While licensing restrictions can be frustrating when terms stand in the way of assembling corpora and running analyses on them, participants learned what a TDM-friendly license might look like, such as one with terms that specifically allow for TDM uses or that contain a fair use clause. The California Digital Library's model license was shared as an example. Licensing was revealed to be an area with the potential for participants to directly intervene in through education, advocacy, and negotiation.

Ethics front and center

Ethics emerged as a major focus of concern for participants in the Institute. Indeed, we quickly realized that although we discussed ethics last, it was difficult to even begin thinking about copyright,

licensing, and other legal issues before ethical considerations were addressed, especially given the Institute's care for questions of social justice. A preferred workflow that emerged for the Institute participants might foreground ethical concerns before moving onto other literacies.

While participants entered the Institute focused on questions of copyright, many reported leaving with their copyright questions solved and their ethical questions awakened. As one participant wrote, the Institute "erased my anxieties in target areas and introduced whole new considerations in areas like ethics. It answered my questions and left me thinking."

Unlike the other literacies, ethics must often be navigated without reliance on the law or clear guidelines. Even IRB guidelines may not always help, particularly as many TDM projects do not have "subjects" in the way that traditional surveys and studies do. Instead, researchers may need to turn to community expectations, other specialists, or disciplinary principles. Sometimes, there may not be any guidance at all, and few solid models for ethics in TDM research are available. In many cases, it will be up to the researchers to determine their own best practices for considering ethics.

One model that resonated with the group was an Ethics of Care approach, which takes into account the relationships between research participants and acknowledges structures of power. Ethics of Care offers an alternative to an individualist consent-based ethical model. In TDM contexts, consent may not always be available or scalable, or the kinds of implied consent (for example, individuals publishing posts to Twitter) may not satisfy the ethical standards of researchers.

Overall, the participants left energized to continue this conversation and contribute to developing ethics models that might guide TDM researchers in the future.

Impact, eight-months on

We analyzed participant update videos and observed not only the lasting impact of the LLTDM literacies, but also a persistent sense of community (or at a minimum, shared experience).

Confidence abounds

One of the themes that arose back in June was the pervasive feeling of imposter syndrome among participants. It seems to permeate this work, perhaps because as one participant so rightly observed, no one person can possibly be a deep expert across an entire landscape of issues in text data mining, from corpus building and computation to legal and ethic issues and all of the many technical, intellectual, and labor issues that underpin the work. But no one mentioned feeling like an imposter in their update videos. Instead we heard about how much more confident they felt integrating the literacies into their work. And this has taken a lot of forms from licensing negotiations to establishing best practices in their labs. The biggest struggle moved from not knowing what to do to finding the time to do it.

Ethics of care

Our closing reflections from the Institute June included strong advocacy for taking an ethics-first approach to teaching the literacies and implementing text data mining projects. It was heartening to see the many ways that participants are living these values by structuring ethics as a key component of their work:

- One scholar added a dedicated ethics section to a paper she submitted that involved the use of YouTube data.
- Another centered ethics in her application of the literacies to a racial reckoning project at her home institution.
- A librarian has adjusted consultations with researchers to take an ethics first approach.
- A faculty member has shifted toward an ethics of care framework in working with students in the classroom and in his research lab.
- Several participants developed workshops and related materials that focus on ethical considerations when doing this work.

They also turned an eye toward institutional gaps where ethics are concerned. One update reflected on the lack of oversight of privacy and ethical issues in TDM research and the need for structures and education that will help with that intervention within our institutions.

Expertise

Across our institutions expertise is both shared and distributed. It would be exceedingly rare to find any one person or even any one office prepared to address the technical, legal, ethical, and logistical nuances of text data mining. Several participants mentioned that it's difficult to build community due in large part to the nature of the work. And living and working through a global pandemic certainly hasn't made that any easier!

Some participants nevertheless made some real gains in community building, and we'd like to celebrate that. One participant described how they initiated conversations across their institution about text data mining to start thinking at an organizational level, and they also noted that they had formed relationships with the

sponsored research office and with the faculty working group on data science. Another participant has taken up the idea of the Data Ombudsperson and is working to introduce it to the scholarly communication group at their library. Yet another participant has established a new research cluster on Critical Practice in Text Data Mining under the auspices of their humanities research center. These kinds of connections hold the potential to make real forward progress within institutions that are notoriously complex.

Institutional risk aversion

One participant described institutional conservatism and risk aversion as their ongoing struggle. And another had hoped to push their institution to be bolder and braver, but it wasn't as easy as they had hoped. Seeding institutional change is long durational work and it begins with small acts of relationship building. It's really important to celebrate these gains while striving for much bigger shifts in practice and perception.

Documentation

One of the most striking things we noticed while watching the update videos was participants' clever use of forms and documentation as tools to help kick start conversations that can ultimately shape practice. One participant described developing an MOU template for use in the digital scholarship lab that includes a section on the legal and ethical implications of the work. The template helps foreground these issues during the negotiation and ensures that they are addressed in the final agreement. In a similar vein, another participant has been developing a rubric for designing new digital projects that incorporates the literacies and is grounded

in the insight that it is best to begin by planning for the end. This presumably helps front load conversations not just about data collection and corpus building but also representation and distribution for publication and long term preservation. To socialize these practices with graduate students, another participant has started requiring a data management plan for student research projects conducted as part of his research lab to ensure everyone in the lab is thinking deeply about ethics in data collection, dehydration, and eventual destruction for social media research. This approach simultaneously generates deep and thoughtful conversations while also making them expected and routine.

Licensing

Several participants have been working to break up their institution's licensing routines with various approaches to address TDM—or not. One participant has been looking at the possibility of regularly including TDM language in institutional licenses, which is in keeping with the approach taken in the California Digital Library's model license agreement. Another participant started working on licensing terms and setting up contracts with vendors at their institution, they ultimately preferred the use of a "Fair Use Escape Clause" rather than outlining specific terms for TDM. They discovered that in an attempt to be explicit, the terms that vendors found acceptable were too confining.

Another piece of the licensing puzzle is making the negotiated terms legible to researchers. One participant has been taking that on with a database evaluation to outline who is eligible to use each resource, how the data may be used, and what content is available. Even when full licenses aren't readily shared with the campus community, this kind of matrix can do a lot of work to help users assess their options when working with content licensed through the libraries.

Workshops

Another way participants have been working with your local communities is by integrating the literacies into your workshops and courses. One participant conducted an hour and a half workshop and has already shared her materials online for those of you who are seeking models for your own efforts on campus. Two other participants collaborated on a workshop foregrounding privacy and ethics in DH projects, which is also available online. And yet another participant has put together a suite of relevant workshops associated with their research cluster.

One challenging thing that came up in an exchange between a participant and a faculty member was the fact that teaching copyright can lead to a lot of fear, uncertainty, and doubt, even when the intention is to empower people to understand their rights. It would be helpful to discuss potential strategies for mitigating that effect as part of our ongoing conversations.

Corpus building

An area where teaching and research appear to intersect is corpus building, and several participants have been applying the lessons from the Institute to your own corpora. One participant has amassed 18,000 YA novels as part of a comparison dataset for use with a digital scholarship project and has also been working to create a standard corpus for each language program in their department so that graduate students have uniform access to a shared dataset right from the beginning of their studies. Another participant has been looking to expand their use of text datasets in their own teaching and has expressed interest in building out a "Law in Literature" text dataset to that end. A third participant has been working on a corpus-building work around that focuses on helping

users run queries that return URLS which can then be downloaded to personal machines. This strategy allows an institution to facilitate TDM while pushing the legal burden to the end user.

Reading list

Required Reading for the Institute

Copyright

Authors Guild v. Google, Inc., 804 F. 3d 202 – Court of Appeals, 2nd Circuit 2015, https://law.justia.com/cases/federal/appellate-courts/ca2/13-4829/13-4829-2015-10-16.html

Rachael 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, edited by Sara Benson (Chicago: Association of College and Research Libraries, 2019), https://escholarship.org/uc/item/55j0h74g.

Matthew Sag, "The New Legal Landscape for Text Mining and Machine Learning," SSRN Electronic Journal, 2019, https://doi.org/10.2139/ssrn.3331606.

Licensing

Nancy Herther, Daniel Dollar, Darby Orcutt, Alicia Wise, and Meg White, "Text and Data Mining Contracts: The Issues and Needs" (2015). Proceedings of the Charleston Library Conference. http://dx.doi.org/10.5703/1288284316233

California Digital Library 2005 Agreement with Factiva: https://cdlib.org/services-groups/collections/

licensed_resources/redacted_licenses/ST_Tier2_Factiva_UCLA_2005_Redacted.pdf

California Digital Library New Model Agreement: http://ucblib.link/33L

Butler, Brandon (2018), "For text- and data-mining, fair use is powerful, but possession is still 9/10 of the law" at http://thetaper.library.virginia.edu/2018/02/28/for-text-and-data-mining-fair-use-is-powerful-but-possession-is-still-9-10-of-the-law-sparc.html

Privacy & Ethics

Suomela, Todd, et al. "Applying an Ethics of Care to Internet Research: Gamergate and Digital Humanities." Digital Studies/Le Champ Numérique, vol. 9, no. 1, Open Library of Humanities, Feb. 2019, p. 4, doi:10.16995/dscn.302.

Jules Polonetsky et al., Benefit-Risk Analysis for Big Data Projects, Future of Privacy Forum, (Sep. 2014), https://fpf.org/wp-content/uploads/FPF_DataBenefitAnalysis_FINAL.pdf

Optional Reading

Flynn, S., Geiger, C., Quintais, J., Margoni, T., Sag, M., Guibault, L., & Carroll, M. (April 20, 2020), "Implementing User Rights for Research in the Field of Artificial Intelligence: A Call for International Action." European Intellectual Property Review 2020, Issue 7. Available at SSRN: https://ssrn.com/abstract=3578819

Van Atteveldt, W., Althaus, S., & Wessler, H. (2020). The trouble with sharing your privates. Pursuing ethical open science and collaborative research across national jurisdictions using sensitive

data. Political Communication https://doi.org/10.1080/10584609.2020.1744780

Additional Sources

62A Am. Jur. 2d Privacy, May 2020 https://drive.google.com/file/d/1Drn41BQyQSJvp_QokXb8IMhgo_R40ssm/view?usp=sharing

Agostinho, Daniela. "Archival Encounters: Rethinking Access and Care in Digital Colonial Archives." Archival Science, vol. 19, no. 2, June 2019, pp. 141–65, doi:10.1007/s10502-019-09312-0.

American Law Institute. Restatement of the Law, Second. Torts. American Law Institute, 1965.

Andrew Garrett, et al. Native American Collections in Archives, Libraries, and Museums at the University of California, Berkeley: Working Group Report. 2019, https://vcresearch.berkeley.edu/research-policies/Native_American_Collections.

Asay, Clark D. and Sloan, Arielle and Sobczak, Dean, Is Transformative Use Eating the World? 61 Boston College Law Review 905 (2020); BYU Law Research Paper No. 19-06. Available at SSRN: https://ssrn.com/abstract=3332682.

Beebe, B. "An Empirical Study of U.S. Copyright Fair Use Opinions, 1978-2005," 156 U. Pa. L. Rev. 549 (2008) https://www.law.upenn.edu/journals/lawreview/articles/volume156/issue3/Beebe156U.Pa.L.Rev.549(2008).pdf

Black, Michael. The World Wide Web as Complex Data Set: Expanding the Digital Humanities into the Twentieth Century and Beyond through Internet Research. International Journal of Humanities and Arts Computing 2016 10:1, 95-109

Buchanan, Elizabeth A., and Michael Zimmer. "Internet Research Ethics." The Stanford Encyclopedia of Philosophy, edited by Edward N. Zalta, Winter 2018, Metaphysics Research Lab, Stanford University, 2018. Stanford Encyclopedia of

Philosophy, https://plato.stanford.edu/archives/win2018/entries/ethics-internet-research/.

California Civil Code § 3345.1. https://codes.findlaw.com/ca/civil-code/civ-sect-3345-1.html.

California Civil Code § 52.5. https://leginfo.legislature.ca.gov/faces/
codes_displaySection.xhtml?lawCode=CIV&sectionNum=52.5.

Caswell, Michelle, and Marika Cifor. "From Human Rights to Feminist Ethics: Radical Empathy in the Archives." Archivaria, vol. 81, no. 1, Association of Canadian Archivists, 2016, pp. 23–43, https://muse.jhu.edu/article/687705/pdf.

Christen, Kimberly. "Relationships, Not Records: Digital Heritage and the Ethics of Sharing Indigenous Knowledge Online." The Routledge Companion to Media Studies and Digital Humanities, Routledge, 2018, pp. 403–12, doi:10.4324/9781315730479-42.

DeCew, Judith. "Privacy." The Stanford Encyclopedia of Philosophy, edited by Edward N. Zalta, Spring 2018, Metaphysics Research Lab, Stanford University, 2018. Stanford Encyclopedia of Philosophy, https://plato.stanford.edu/archives/spr2018/entries/privacy/.

D'Ignazio, Catherine, and Klein, Lauren. Data Feminism. MIT Press, 2020.
http://datafeminism.io/

Dixon, Shane, and Linda Quirke. "What's the Harm? The Coverage of Ethics and Harm Avoidance in Research Methods Textbooks." Teaching Sociology, vol. 46, no. 1, SAGE Publications Inc, Jan. 2018, pp. 12–24, doi:10.1177/0092055X17711230.

Dressler, Virginia, and Cindy Kristof. The Right to Be Forgotten and Implications on Digital Collections: A Survey of ARL Member Institutions on Practice and Policy | Dressler | College & Research Libraries, doi:https://doi.org/10.5860/crl.79.7.972.

Dulong de Rosnay, Mélanie, and Andrés Guadamuz. Memory Hole or Right to Delist? Implications of the Right to Be Forgotten for Web

Archiving. SSRN Scholarly Paper, Social Science Research Network, 1 June 2017, https://papers.ssrn.com/abstract=3107565.

European Privacy Requests Search Removals FAQs – Transparency Report Help Center, https://support.google.com/transparencyreport/answer/7347822?hl=en.

First Archivist Circle. Protocols for Native American Archival Materials. 2007, http://www2.nau.edu/libnap-p/protocols.html.

"General Data Protection Regulation Art. 17 – Right to Erasure ('right to Be Forgotten')." GDPR.Eu, 14 Nov. 2018. gdpr.eu, https://gdpr.eu/article-17-right-to-be-forgotten/.

"General Data Protection Regulation Art. 89 – Safeguards and Derogations Relating to Processing for Archiving Purposes in the Public Interest, Scientific or Historical Research Purposes or Statistical Purposes." GDPR.Eu, 14 Nov. 2018. gdpr.eu, https://gdpr.eu/article-89-processing-for-archiving-purposes-scientific-or-historical-research-purposes-or-statistical-purposes/.

Gilliland, Anne J. and Sue McKemmish. "The Role of Participatory Archives in Furthering Human Rights, Reconciliation and Recovery." (2014). Atlanti: Review for Modern Archival Theory and Practice, vol. 24, Oct. 2014, https://escholarship.org/uc/item/346521tf.

Helen, Kara. Research Ethics in the Real World: Euro-Western and Indigenous Perspectives. Policy Press, 2018. See https://policy.bristoluniversitypress.co.uk/research-ethics-in-the-real-world.

Henttonen, Pekka. "Privacy as an Archival Problem and a Solution." Archival Science, vol. 17, no. 3, Sept. 2017, pp. 285–303, doi:10.1007/s10502-017-9277-0.

Intergovernmental Committee on Intellectual Property and Genetic Resources, Traditional Knowledge and Folklore. Glossary of Key Terms Related to Intellectual Property and Genetic Resources, Traditional Knowledge and Traditional Cultural Expressions. 2018, https://www.wipo.int/edocs/mdocs/tk/en/wipo_grtkf_ic_37/wipo_grtkf_ic_37_inf_7.pdf.

Intergovernmental Committee on Intellectual Property and

Genetic Resources, The Protection of Traditional Knowledge: Draft Articles. 2018, https://www.wipo.int/edocs/mdocs/tk/en/wipo_grtkf_ic_38/wipo_grtkf_ic_38_4.pdf.

Lor, Peter Johan, and J. J. Britz. "An Ethical Perspective on Political-Economic Issues in the Long-Term Preservation of Digital Heritage." Journal of the American Society for Information Science and Technology, vol. 63, no. 11, 2012, pp. 2153–64, doi:10.1002/asi.22725.

Mathiesen, Kay. "A Defense of Native Americans' Rights over Their Traditional Cultural Expressions." The American Archivist, vol. 75, no. 2, Society of American Archivists, Oct. 2012, pp. 456–81, doi:10.17723/aarc.75.2.0073888331414314.

Powell, Timothy B. "The American Philosophical Society Protocols for the Treatment of Indigenous Materials." Proceedings of the American Philosophical Society, vol. 158, no. 4, Dec. 2014, pp. 411–20, https://www.amphilsoc.org/sites/default/files/2017-11/attachments/APS%20Protocols.pdf

Rainie, Stephanie Carroll, et al. Indigenous Data Sovereignty. African Minds and the International Development Research Centre (IDRC), 2019, pp. 300–19, https://researchcommons.waikato.ac.nz/handle/10289/12918.

Research Data Alliance International Indigenous Data Sovereignty Interest Group. CARE Principles for Indigenous Data Governance. The Global Indigenous Data Alliance, Sept. 2019, https://static1.squarespace.com/static/5d3799de845604000199cd24/t/5da9f4479ecab221ce848fb2/1571419335217/CARE+Principles_One+Pagers+FINAL_Oct_17_2019.pdf.

Review of Copyright Exceptions for Research – www.tinyurl.com/researchexceptions

Sag, Matthew. Predicting Fair Use (February 25, 2012). Ohio State Law Journal, Vol. 73:1 47-91 (2012) http://dx.doi.org/10.2139/ssrn.176913

Suomela, Todd, et al. "Applying an Ethics of Care to Internet Research: Gamergate and Digital Humanities." Digital Studies/Le

Champ Numérique, vol. 9, no. 1, Open Library of Humanities, Feb. 2019, p. 4, doi:10.16995/dscn.302.

United Nations Declaration on the Rights of Indigenous Peoples. A/RES/61/295, 13 Sept. 2007, https://www.un.org/development/desa/indigenouspeoples/wp-content/uploads/sites/19/2018/11/UNDRIP_E_web.pdf.

Van Atteveldt, Wouter, Scott Althaus, and Hartmut Wessler. 2020. "The Trouble with Sharing Your Privates: Pursuing Ethical Open Science and Collaborative Research across National Jurisdictions Using Sensitive Data." Political Communication: 1-7. https://doi.org/10.1080/10584609.2020.1744780

Vitak, J., Shilton, K., & Ashktorab, Z. 2016. Beyond the Belmont Principles: Ethical Challenges, Practices, and Beliefs in the Online Data Research Community. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (CSCW '16). Association for Computing Machinery, New York, NY, USA, 941–953. DOI:https://doi.org/10.1145/2818048.2820078

Waldman, Ari W. Privacy Law's False Promise, 97 WASH. U. L. REV. 0773 (2020). Available at: https://openscholarship.wustl.edu/law_lawreview/vol97/iss3/7

Wright, David, and Renée Saucier. "Madness in the Archives: Anonymity, Ethics, and Mental Health History Research." Journal of the Canadian Historical Association / Revue de La Société Historique Du Canada, vol. 23, no. 2, The Canadian Historical Association / La Société historique du Canada, 2012, pp. 65–90, doi:https://doi.org/10.7202/1015789ar.

Zwitter, Andrej. "Big Data Ethics." Big Data & Society, vol. 1, no. 2, SAGE Publications Ltd, July 2014, pp. 1-6. doi:10.1177/2053951714559253.

Videos, Slides, Transcripts

Videos, Slides, Transcripts

Below are links to instructional videos, slides, and transcripts used in the delivery of Building LLTDM.

Copyright

- View all Copyright videos in playlist
- Copyright Video 1: https://youtu.be/3QSj3P7vL0o (0:38)
- Copyright Video 2: https://youtu.be/K2b7olTIaao (5:22)
- Copyright Video 3: https://youtu.be/fCnWPtjstuA (7:57)
- Copyright Video 4: https://youtu.be/OkJluvCZMXU (7:08)
- Copyright Video 5: https://youtu.be/F1dsqXEkFEM (4:54)
- Copyright Video 6: https://youtu.be/TguLxmLd_l4 (6:54)
- Copyright Video 7: https://youtu.be/2K9A0Vtgtms (8:49)
- Copyright Video 8: https://youtu.be/CzH_wIantaM (4:38)
- Copyright Video 9: https://youtu.be/ovwfHTgW1yw (6:33)
- Copyright Video 10: https://youtu.be/jBgqe1ljGhI (5:24)

Total viewing time: 1 hour

If you don't learn best through videos, we've got you covered:

- Copyright Videos Transcript
- Copyright Videos Slides

International Copyright

- View all International Copyright videos in playlist
- Sean Flynn Introduction: https://youtu.be/tcAM3VyXs9I (3:05)
- International Copyright Video 1: https://youtu.be/TBHHI7ZtqFk (3:21)
- International Copyright Video 2: https://youtu.be/5atNHDz9MjY (7:43)
- International Copyright Video 3: https://youtu.be/-9ZEWnXsQ_0 (8:55)
- International Copyright Video 4: https://youtu.be/KNp3ffbHiec (8:52)
- International Copyright Video 5: https://youtu.be/dtc0brOVzTk (10:04)
- International Copyright Video 6: https://youtu.be/Pz5KOBXWgFw (18:21)
- International Copyright Video 7: https://youtu.be/c_SuEN0qk6w (5:04)
- International Copyright Video 8: https://youtu.be/sCG5X0iwSx8 (8:11)
- International Copyright Video 9: https://youtu.be/Tr3pr78iD6Y (5:29)
- International Copyright Video 10: https://youtu.be/tod3U42AASc (3:35)
- International Copyright Video 11: https://youtu.be/bCSq-F5PLuU (7:11)

Total viewing time: 1 hour, 30 minutes

If you don't learn best through videos, we've got you covered:

- International Copyright Videos Transcript
- International Copyright Videos Slides

Technological Protection Measures

- View all TPM Videos in playlist
- Matt Sag Introduction: https://youtu.be/EC0kXNomii0 (1:08)
- Technological Protection Measures Video 1: https://youtu.be/mjTsctueZRU (8:38)
- Technological Protection Measures Video 2: https://youtu.be/TYZixxLFdv0 (5:34)
- Technological Protection Measures Video 3: https://youtu.be/iBcuwC_HhcY (7:53)

Total viewing time: 23 minutes

If you don't learn best through videos, we've got you covered:

- Technological Protection Measures Videos Transcript
- Technological Protection Measures Videos Slides

Licensing

- View all Licensing videos in playlist
- Licensing Video 1: https://youtu.be/woIV5SOaeVQ (0:49)
- Licensing Video 2: https://youtu.be/StX4Gk6Y-dc (3:28)
- Licensing Video 3: https://youtu.be/R6EYyo2lqyU (5:14)
- Licensing Video 4: https://youtu.be/brg8IlqJtB8 (8:03)
- Licensing Video 5: https://youtu.be/lybowwUhZQQ (10:17)
- Licensing Video 6: https://youtu.be/HtxGVkImVsk (10:11)
- Licensing Video 7: https://youtu.be/pXLLWjv3QCw (6:11)
- Licensing Video 8: https://youtu.be/lVItEFW2aOE (9:46)
- Licensing Video 9: https://youtu.be/p0zBcLVFYNk (8:00)

Total viewing time: 1 hour, 5 minutes

If you don't learn best through videos, we've got you covered:

- Licensing Videos Transcript
- Licensing Videos Slides

Privacy & Ethics

- View all Privacy & Ethics videos in playlist
- Privacy & Ethics Video 1: https://youtu.be/ClFG2DAFMzM (7:43)
- Privacy & Ethics Video 2: https://youtu.be/77E-IsmmlZM (10:48)
- Privacy & Ethics Video 3: https://youtu.be/xBskFg8g2Lk (4:33)
- Privacy & Ethics Video 4: https://youtu.be/nRyKaWhkMU8 (8:22)
- Privacy & Ethics Video 5: https://youtu.be/IFt8ifmlm2U (9:08)
- Privacy & Ethics Video 6: https://youtu.be/bvZVNUBV0k0 (6:56)
- Privacy & Ethics Video 7: https://youtu.be/nf7Gy753zto (7:47)
- Privacy & Ethics Video 8: https://youtu.be/WUgA5KHKKM0 (7:30)
- Privacy & Ethics Video 9: https://youtu.be/t2s7ULoT23E (2:55)

Total viewing time: 1 hour, 6 minutes

If you don't learn best through videos, we've got you covered:

- Privacy & Ethics Videos Transcript
- Privacy & Ethics Videos Slides

Privacy & Ethics has an optional bonus set of videos exploring privacy in more depth:

- Privacy Closer Look Video 1: https://youtu.be/g2z2fyu7q20
- Privacy Closer Look Video 2: https://youtu.be/ZKBBlI-hLwI
- Privacy Closer Look Video 3: https://youtu.be/nN80mdh4apI
- Privacy Closer Look Video 4: https://youtu.be/IN_Zi4eoOI4

- Privacy Closer Look Video 5: https://youtu.be/WEU0Gtc-ilg

Building Legal Literacies for Text Data Mining: Institute White Paper

Also available as a Google doc.

Project Summary

Until now, digital humanities (DH) researchers conducting text data mining (TDM) in the U.S. have had to maneuver through a thicket of legal issues without much guidance or assistance. Uncertainty about the breadth and contours of TDM rights and obligations has impeded the scope of DH research questions, or unnecessarily exposed scholars to risk. We designed Building Legal Literacies for Text Data Mining (Building LLTDM) to address these questions and barriers to facilitate DH TDM research. Funded as an NEH Institute for Advanced Topics in the Digital Humanities, and hosted by UC Berkeley from June 23-26, 2020, Building LLTDM provided 32 DH TDM researchers, librarians, and professionals with foundational skills to:

1. confidently navigate law, policy, ethics, and risk within DH TDM projects;
2. integrate workflows at their home organizations to provide law and policy support for DH TDM projects;
3. practice sharing these new skills and workflows through authentic consultation exercises;
4. prototype plans for broadly disseminating their knowledge; and

5. develop communities of practice to promote cross-institutional outreach about the DH TDM legal landscape.

While we originally planned Building LLTDM to be held on the UC Berkeley campus, the COVID-19 pandemic necessitated a transition to online teaching. Our faculty of legal experts, librarians, and researchers from across the U.S. provided interactive remote instruction. We presented the substantive content through pre-recorded videos and held live group discussions in a flipped classroom model. We also provided the video transcripts and slides to participants to promote accessibility and accommodate multiple learning styles.

To maximize the reach and impact of Building LLTDM, we compiled the legal literacies covered during the institute into an Open Educational Resource (OER) with a public domain (CC0) dedication. The OER covers copyright (both U.S. and international law), technological protection measures, privacy, and ethical considerations. It also helps other DH professionals and researchers run their own similar institutes by describing in detail how we developed and delivered programming (including our pedagogical reflections and take-aways), and includes ideas for hosting shorter literacy teaching sessions.

Project Origins & Goals

Growth of Text Data Mining in Digital Humanities

If one were to crack open popular English-language novels written in the 1850s—say, ones from Brontë, Hawthorne, Dickens, and Melville—one would find they describe men and women in very

different terms. While a male character might be said to "get" something, a female character is more likely to have "felt" it. Whereas the word "mind" might be used when describing a man, the word "heart" is more likely to be used about a woman. As the 19th Century became the 20th, these descriptive differences between genders diminish within these novels. And we know all this because researchers have used automated techniques to extract information from the novels, and analyzed word usage trends at scale.[1] They crafted algorithms to turn the language of those novels into data about the novels.

In fields of inquiry like the digital humanities, the application of such automated techniques and methods for identifying, extracting, and analyzing patterns, trends, and relationships across large volumes of unstructured or thinly-structured digital content is called "text data mining" or "TDM". (One may also see it referred to as "text and data mining" or "computational text analysis"). TDM is an increasingly important and prevalent research methodology leveraging algorithms to sift, organize, and analyze vast amounts of thinly-structured textual content.[2] For instance, these methods make it possible to: discern racial disparity by evaluating language

1. Underwood, T., Bamman, D., & Lee, S. (2018). The transformation of gender in English-language fiction. Journal of Cultural Analytics. Available at https://doi.org/10.22148/16.019
2. Hearst, M. (2003, October 17). What is Text Mining? Available at http://people.ischool.berkeley.edu/~hearst/text-mining.html.

from police body camera footage;[3] assess visual culture;[4] and examine conversation patterns on Twitter regarding social justice issues such as violence against women.[5] TDM methodologies and tools continue to expand, posing great opportunities for advancements across education, literature, society, politics, and beyond.[6]

Law and Policy Hurdles

Until Building LLTDM, DH researchers conducting TDM faced confusing legal considerations, and a marked absence of community

3. Voigt, R., et al., (2017). Language from police body camera footage shows racial disparities in officer respect. Proceedings of the National Academy of Sciences, 114(25), 6521. Available at https://doi.org/10.1073/pnas.1702413114.
4. Arnold, T., & Tilton, L. (2019). Distant viewing: Analyzing large visual corpora. Digital Scholarship in the Humanities. Available at https://doi.org/10.1093/digitalsh/fqz013.
5. Xue, J., et al., (2019). Harnessing big data for social justice: An exploration of violence against women-related conversations on Twitter. Human Behavior and Emerging Technologies, 1(3), 269–279. Available at https://doi.org/10.1002/hbe2.160.
6. Hassani, H., et al., (2020). Text Mining in Big Data Analytics. Big Data and Cognitive Computing, 4(1). Available at https://doi.org/10.3390/bdcc4010001.

guidance for navigating them. For instance, imagine that researchers wish to digitally crawl and download content about Egyptian tombs and artifacts from online websites, in order to conduct an automated computational analysis on the web-scraped materials. Then imagine the researchers also want to share these content-rich datasets to encourage research reproducibility or enable other scholars to query the datasets with new questions. This kind of work can raise issues of:

- **Copyright** (e.g. Are the images protected by copyright? Does an exception like fair use apply?)
- **Contracts** (e.g. Are there database license agreements or website terms of use that govern what researchers are permitted to scrape or download? Do these agreements override copyright exceptions?)
- **Privacy** (e.g. Do the images reveal information that could infringe upon the privacy rights of the subjects under federal and state laws? Does downloading images that should not have been made public constitute a further privacy violation?)
- **Ethics** (e.g. Are there social and religious customs, or other circumstances like indigenous knowledge that could impact downloading and use of the materials?)

If researchers are not comfortable navigating these issues or feel that, in doing so, they or their institutions would take on too much risk, they may abandon their projects. Indeed, a study of humanities scholars' text analysis needs found that access to and use of copyright-protected texts was a "frequent obstacle" in participants' ability to select appropriate texts for TDM.[7]

7. Green, H., et al., (2016). Scholarly Needs for Text Analysis Resources: A User Assessment Study for the HathiTrust Research Center. Proceedings of the Charleston Library

Potential legal hurdles do not just deter TDM research; they also bias it toward particular topics and sources of data. In response to confusion over copyright, website terms of use, and other perceived legal roadblocks, some digital humanities researchers have gravitated to low-friction research questions and texts (e.g. materials exclusively in the public-domain or datasets already compiled) to avoid decisions about rights-protected data. Restricting research to such sources can skew inquiries, leave important questions unanswered, and render resulting findings less broadly applicable. A growing body of research also demonstrates how race, gender, and other biases found in openly available texts have contributed to and exacerbated bias in developing artificial intelligence tools.[8]

Sound guidance from information professionals can help researchers traverse these concerns. Yet, scholars have reported hesitation to seek help from institutional staff whom they fear will question the legality of their TDM methods, or advocate for a more risk-averse approach than the law warrants. Those worries may be validated when libraries sign or enforce license agreements to datasets with unclear or, in some cases, hostile TDM provisions. If equipped with legal and ethical literacies, institutional staff as well as researchers would be better positioned to understand what the law already permits, and negotiate for better usage rights overall.

Conference. Available at http://dx.doi.org/10.5703/1288284316464.

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

Past Work Demonstrated Need for Training

For all of these reasons, our project team wanted to help DH scholars and research professionals better navigate the law and policy landscape of TDM—using a pedagogical approach[9] that enables researchers to fully and fairly utilize rights-protected works, and disseminate their resulting TDM scholarship broadly. Our intended framework would also need to support TDM researchers in understanding and navigating ethical issues like corpus bias and subject consent.

We began designing our institute by canvassing existing educational programs—which cemented the need for our training. In reviewing a broad sample of digital humanities, humanities, and information science curricula, professional development training programs, and library guides, we found scant trainings or resources that integrate TDM legal literacies into outreach and instruction. While there were a growing number of DH training opportunities on TDM methods and tools, they almost universally omitted copyright and other law or policy concerns. Moreover, our own experiences suggested that DH scholars and professionals face many of the questions that arise around legal issues and TDM at the time of crisis (e.g., when university access to a database is suspended due to systematic downloading). This places undue stress on DH scholars'

9. An early formulation of that approach was articulated in project team members' 2019 paper: Samberg, R. G., & Hennesy, C. (2019). Law and literacy in non-consumptive text mining: Guiding researchers through the landscape of computational text analysis. Copyright Conversations: Rights Literacy in a Digital World (pp. 289–315). ACRL. Available at https://escholarship.org/uc/item/55j0h74g.

ability to conduct DH TDM research and may lead institutions to unduly restrict such research via institutional policy.

We understood that addressing this educational need would require cross-organizational training aimed at both (1) the *scholars conducting* TDM, and (2) the *professional staff who assist and collaborate* with them. Digital humanities professionals are people like librarians, consultants, and other institutional staff who conduct digital humanities text data mining or aid researchers in their text data mining research. The DH professional stakeholder group is essential to maximizing the efficacy of legal literacy education for a number of reasons. DH professionals who teach or consult on TDM are well-positioned to incorporate legal literacies into existing trainings. Further, academic libraries, labs, and departments license many of the databases and datasets DH researchers seek to use. Staff are then called upon to provide input on or information about database terms and conditions, and they may be positioned to secure better licensing terms from the start. Many libraries also employ legal experts within scholarly communications or copyright units—some of whom have established TDM training programs and service models that could be adjusted to incorporate law and policy workflows.

There was ample reason to believe that an institute devoted to the development of these legal literacies for DH researchers and professionals would be highly productive. For example, copyright training sessions for librarians have already been found to be effective in building understanding and confidence around copyright research consultations. Educating DH researchers and professionals through a focused institute also offers the benefit of creating shared understanding across the scholarly landscape. This in turn would offer the potential for downstream impact as all participants would be poised to return to their home institutions or professional communities and share what they have learned.

Goals

The law and policy impediments to DH TDM research, coupled with the need for training to help researchers and professionals navigate them, prompted us to provide DH professionals and researchers with foundational skills to:

1. confidently navigate law, policy, ethics, and risk within DH TDM projects;
2. integrate workflows at their home organizations to provide law and policy support for DH TDM projects;
3. practice sharing these new skills and workflows through authentic consultation exercises;
4. prototype plans for broadly disseminating their knowledge; and
5. develop communities of practice to promote cross-institutional outreach about the DH TDM legal landscape.

Project Overview

Our aim is to facilitate the replicability of Building LLTDM institute by others. Accordingly, in this section we detail project design and administration chronologically for easier implementation:

1. Faculty & participant recruitment
2. Provision of financial support
3. Pre-institute preparation
4. Institute schedule & activities
5. Post-institute "catch up"
6. Creation of OER

People

Faculty

Building LLTDM was led by the Office of Scholarly Communication Services at the University of California, Berkeley Library. Rachael Samberg, Scholarly Communication Officer & Program Director, served as Project Director. She oversaw curricular design and execution, as well as the administrative and operational aspects of the institute. Timothy Vollmer, Scholarly Communication & Copyright Librarian, served as Project Manager, and was responsible for coordinating the design and execution of the institute, and streamlining administrative and operational aspects. Both the Project Director and Project Manager also served as faculty instructors for the institute, helping to create and deliver educational materials and training.

The remaining institute faculty hailed from more than a dozen North American universities and institutions, and were each responsible for contributing to institute curricular design and delivery. Faculty were recruited through professional connections and networks, and were composed of legal experts ("LE"), librarians ("L"), and humanities researchers ("HR"). Their real-world roles straddled these boundaries (e.g. some legal experts are also librarians); yet, the nominal divisions ensured that institute sessions were led by a set of experts who collectively offer a full range of relevant DH expertise. We also had an additional legal expert on call via e-mail during the institute to field any questions that instructors were unable to answer in real time.

Project team members included:

- Scott Althaus, Professor of Political Science & Communication, and Director of the Cline Center for Advanced Social Research at University of Illinois

- David Bamman, Assistant Professor at UC Berkeley's School of Information
- Brandon Butler, Director of Information Policy at the University of Virginia (UVA) Library
- Beth Cate, Associate Professor at Indiana University Bloomington's School of Public and Environmental Affairs (SPEA)
- Kyle K. Courtney, Copyright Advisor for Harvard University, within the Office for Scholarly Communication
- Sean Flynn, Associate Director of the Program on Information Justice and Intellectual Property (PIJIP) and Professorial Lecturer in Residence
- Maria Gould, Research Data Specialist/Product Manager, California Digital Library
- Cody Hennesy, Journalism and Digital Media Librarian at University of Minnesota
- Eleanor Dickson Koehl, HathiTrust Digital Scholarship Librarian at the University of Michigan Libraries, and Associate Director for Outreach and Education, HTRC
- Thomas Padilla, Visiting Digital Research Services Librarian at University of Nevada Las Vegas
- Stacy Reardon, Literatures and Digital Humanities Librarian at UC Berkeley
- Matthew Sag, Professor of Law at Loyola University Chicago School of Law
- Brianna L. Schofield, Executive Director of Authors Alliance
- Glen Worthey, Associate Director for Research Support Services, HathiTrust Research Center
- Megan Senseney, Head of the Office of Digital Innovation and Stewardship at University of Arizona Libraries
- Sara Benson, Copyright Librarian at University of Illinois

Participant Recruitment

We designed the institute to support 32 participants, resulting in what we believed would be a suitable instructor-to-attendee ratio to accommodate the highly immersive and discursive aspects of a design thinking framework (discussed further below).

We sought participation from both DH researchers and professionals. We anticipated that both groups would have mutually beneficial insights and experiences to share. For instance, DH researchers would benefit from LLTDM training that can be applied to their own research projects and publications, and integrated into their teaching and advising—thereby broadening downstream community impact. Conversely, DH professionals are often the first contact point for DH researchers with law-related TDM questions; handle licensing and negotiate access to datasets and digital collections for TDM; and provide training and documentation for DH researchers on workflows and tools. Educating DH professionals would enable ongoing institute impact as these professionals can bring the skills they have gained back to their own campuses and professional communities. We also aimed for approximately equal numbers of DH researchers and DH professionals to maximize impact—recognizing that these two groups are variously situated in their organizations and thus can provide future advocacy and support in different ways. For similar reasons, we encouraged participation from institutional pairs of participants (e.g. one digital humanities researcher and one professional affiliated with that same organization or project) with the hope that greater representation from a given institution could result in broader literacy implementation at that institution following participant training.

With institute scope and intended reach determined, we developed a project website to host information about the institute application process, timeline, and criteria. We advertised the application process on the Building LLTDM blog, digital humanities

and library-related email lists, and social media. The submission window was open for two months, and the application process required individuals to submit two documents: (1) a current CV, and (2) a 2-page (maximum) letter of interest. In their letters of interest, we asked applicants to account for experience with or interest in: the intersection of TDM in DH research and the law; their goals for applying knowledge and skills to be acquired at the institute to their own activities; their goals for sharing knowledge and skills with others at their home institutions/affiliations; and, how they might support the institute's commitment to diversity and equity.

We posted selection criteria prominently on the Building LLTDM website, and gave particular influence to diversity, equity, and inclusion. In particular, the faculty believed that the institute would work best if it reflected the race and gender demographics of the broader population, and not just those of higher education—and we strived to achieve equity by reflecting these more representative demographics. Additionally, we worked to develop a participant group that was representative of different institution types, research advising and support experience, professional roles, levels of experience with DH TDM research, career stages, and disciplinary perspectives.

The selection process took place over two main rounds. First, a subset of the faculty conducted an initial assessment of all applications based on the selection criteria. Our subgroup then met in successive sessions to discuss and normalize rankings, and reached consensus on recommended candidates. We presented our recommendations to the project team for discussion in a full group meeting. The suggested group was composed of 15 DH researchers and 17 DH professionals hailing from 15 different states. We are also pleased to report that all of our selected participants accepted our offer to be institute participants, and included:

- Ilya Akdemir, University of California, Berkeley
- Tara Baillargeon, Marquette University
- Trevor Burrows, Purdue University

- Matthew Cannon, University of California, Berkeley
- Nathan Carpenter, Illinois State University
- Ashleigh Cassemere-Stanfield, University of Chicago
- James Clawson, Grambling State University
- Mark Clemente, Case Western Reserve University
- Quinn Dombrowski, Stanford University
- Alyssa Fahringer, George Mason University
- Heather Froehlich, Penn State University
- Nicole Garlic, Temple University
- Casey Hampsey, New York University
- Devin Higgins, Michigan State University
- Christian Howard, Bucknell University
- Daniel Johnson, Notre Dame University
- Spencer Keralis, University of Illinois
- Sarah Ketchley, University of Washington
- Melanie Kowalski, Emory University
- Barbara Levergood, Bowdoin College
- Jes Lopez, Michigan State University
- Rochelle Lundy, Seattle University
- Jon Marshall, UC Berkeley
- Jens Pohlmann, Stanford University
- Caitlin Pollock, University of Michigan
- Sarah Potvin, Texas A & M University
- Andrea Roberts, Texas A & M University
- Daniel Royles, Florida International University
- Hadassah St. Hubert, Florida International University
- Todd Suomela, Bucknell University
- Nicholas Wolf, New York University
- Madiha Zahrah Choksi, Columbia University

Financial support

On our project website, we made clear to potential applicants that

participant stipends would be distributed in advance of the institute. This was designed to promote equity by helping participants avoid having to expend personal funds or await reimbursement. Had the institute been in person, the paid-in-advance stipends would have been sufficient to cover travel, lodging, and related expenses with the aim of eliminating out-of-pocket expenses. As we found ourselves having to rapidly transition the institute online while participant stipends were concurrently being distributed by the university business office, we conferred with our NEH program officer about how to proceed. With NEH guidance, we maintained stipend distribution as awarded in the grant—with stipends being repurposed to compensate for participant time and incentivize participation.

We also offered instructor honoraria to faculty. The honoraria were originally intended to both (1) cover faculty travel costs to the institute, and (2) recognize the substantial contributions project team members were making for developing and teaching curriculum and creating the post-institute OER. (No faculty member time was being charged to the grant, and instead all efforts were contributed from people's personal time.) As COVID-19 unfolded, and as with participant stipends, we consulted with our NEH program officer and were advised that honoraria should similarly continue, with a focus shifting to rewarding faculty contributions.

Pre-institute Preparation

After participants were chosen, the pre-institute timeline was filled with both substantive and logistical planning:

- **Four months pre-institute**: While simultaneously developing instructional content, we also began regular communications with participants, which increased in frequency as the

pandemic spread. We began communications through individual and group e-mails. As the start date for the institute approached, we transitioned to Slack for announcements and community information sharing, and to help build familiarity and collegiality. We created a Slack sub-channel for faculty and participant introductions. In addition, faculty and participants created sub-channels to discuss specific TDM research areas, such as social media and oral histories.

- **One month pre-institute**: We sent participants a short questionnaire so that the faculty instructors could learn more about participants' research or professional practices related to TDM. This allowed faculty to better understand the participants' real-world experiences and struggles, and tailor the upcoming sessions and exercises to properly meet participant expectations and needs. We also developed a Faculty Facilitation Guide (referred to as the "Faculty Packet") for instructors to help faculty prepare for administering the institute. This Google doc contained faculty and participant contact information, information about how to use Zoom effectively, and guidance about how to support participant contributions and positive interactions during the online institute.

- **One week pre-institute**: We distributed pre-reading to participants that provided an overview of the TDM legal and policy environment. However, we kept the amount of required preparation to a minimum—both because we knew the participants were busy individuals with full time jobs and research responsibilities, and also due to the added pressure and stresses of the COVID-19 pandemic. We set the expectation that we hoped the participants would be able to provide as much undivided attention as they could during the actual week of delivery (of course understanding that there might be necessary interruptions due to family or personal

responsibilities because of the remote nature of the institute).

We also distributed a comprehensive guide for participants that we called the Participant Packet—essentially a one-stop-shop to guide participants through the week ahead. The Participant Packet included:

- Information about how to communicate with faculty and other participants
- Instructions for how to use Zoom during institute sessions
- Institute code of conduct (to which the cohort had consented upon acceptance of the offer to participate)
- Information about social media usage and the applicability of the Chatham House Rule to protect participant communications
- Day-by-day agenda for the institute, including assigned meeting groups of various sizes (plenary, small group), free-write activities, and also links to Zoom rooms and shared notes documents for each session
- Links to readings and pre-recorded short videos (with transcripts and slides) so that participants could be prepared for the next day's topics[10]

10. In order to provide easy public viewing for all the pre-recorded TDM topical videos, we uploaded them to the UC Berkeley Library's Office of Scholarly Communication Services YouTube account. Viewers can also speed up or slow down the video playback, or turn on closed captions; both features are offered automatically by YouTube. We also created playlists under each topical area (copyright, international copyright, licensing, technological protection measures,

Institute Schedule & Activities

Design Thinking Approach

We believed that design thinking offered an apt instructional framework to convey the literacies while sufficiently engaging participants. Design thinking relies upon experiential meeting methodologies that foster hands-on learning and allow participants to experiment with developing their own solutions for their TDM hurdles. The institute tracked the five stages of design thinking as follows:

- **Empathy (Institute Day 1)**: Building trust and common understanding through experience sharing can foster robust discussion and collaborative inquiry. We thus began the first day of the institute by developing our collective understanding of participants' experiences with TDM through storyboarding sessions. These empathy-supporting activities served as an opportunity for participants to get to know each other and to start learning about each other's hurdles and successes with the LLTDM literacies. The exchanges helped participants discover that they are not alone in their struggles but rather are part of a burgeoning community.

- **Define & Ideate (Institute Days 2 & 3)**: For days two and three, we cycled iteratively through the "define" and "ideate" phases of design thinking. Defining and ideation are foundational for developing a shared language to discuss the contours of TDM challenges, and to lay the groundwork for participants to

and privacy & ethics), as well as a comprehensive playlist containing all the videos.

strategize about customized solutions. For these stages, faculty worked with participants to articulate and contextualize TDM issues and literacies through: (1) asynchronous videos conveying the substantive literacies, and (2) synchronous small group time to discuss case studies and undertake "putting it together" exercises (more below under "Daily Agenda") to simulate real-world problems.

- **Prototype & Test (Institute Day 4; Post Institute)**: Prototyping involves developing a personalized approach to implementing takeaways and solutions. To model this stage, on the final day of the institute, participants crafted implementation plans regarding how they will integrate the literacies into their work and at their home institutions. Testing would then occur in the months following the institute, as participants put their plans into place. To follow up on testing, we stayed connected through Slack and reconvened the cohort eight months after the institute to learn from each other's outcomes (more below under "Post-Institute Meeting").

Daily Agenda

General Schedule

We adjusted the institute's prime content delivery mechanism to asynchronous (pre-recorded) instructional videos so that we could utilize synchronous sessions for small group discussions and experiential exercises. This minimized sedentary time in front of a computer and allowed participants the opportunity to pace themselves according to their personal schedules and learning styles. We also spaced sessions with intermittent breaks to help participants focus and avoid Zoom burnout. Because participants

and faculty were joining from different time zones, we began sessions at 8 a.m. Pacific Time and concluded by 2 p.m. Pacific Time[11], to wrap by the end of normal business hours on the East Coast. This allowed participants sufficient time to prepare for each subsequent day's content and activities irrespective of time zones.

Day 1

1. *Introductions and stage setting*: Faculty instructors used a master slide deck throughout the week. Day 1 began with a welcome, logistical information, explanation of the code of conduct and Chatham House Rule, and a framing for the week's activities. One of the faculty instructors also served as an institute moderator. The moderator's key roles were to: (1) observe and synthesize emerging themes from each day to bolster learning outcomes, and (2) assist with cross pollination of ideas and themes from across small breakout groups. The moderator tuned in to small group discussion sessions and collected individual reflections for sharing at the end of each day.
2. *Empathy building exercise*: Following the moderator's introduction, participants engaged in a virtual white board exercise designed to help them storyboard their own experiences with TDM; build knowledge and understanding among participants; and surface aspects of divergence and convergence across individual experiences. We used the online "sticky note" software tool called Mural for this journey mapping exercise.

11. At the end of each day, we offered optional and informal "Happy Half-Hours" on Zoom. This time was to socialize, decompress, and answer participant questions.

3. *Free Write*: Day 1 ended with a free write exercise (the first of three such exercises over the course of the week). Freewriting was intended as an opportunity to reflect on the day's sessions and apply them to one's personal circumstances, research interests, institutional culture, team dynamics, etc. Participants were asked to write without pausing or proofreading and in response to the following prompts:

    - What did you learn from other participants today about variations in TDM processes and logistical complexities?
    - Which pain points highlighted by other participants resonated with you?
    - What new questions, concerns, or opportunities emerged during report outs that you didn't capture on the mural board?

Participants e-mailed their text to our shared faculty email group. The institute moderator and several instructors reviewed the submitted responses each evening in preparation for an opening reflection to kick off the next day.

Day 2

1. *Report back from moderator on free write themes*: At the beginning of day 2, the moderator summarized the motifs and lessons evidenced in the previous day's free writes. This practice reminded participants about the themes discussed the day before, and helped them track progress and accomplishments throughout the week.
2. *Substantive literacies—Copyright, international copyright, TPMs*: On day 2, we began to explore the substantive law and policy literacies for text data mining in the digital humanities. We covered copyright (focusing heavily on U.S. law), copyright in the international/cross-border context, and technological

protection measures. As mentioned above, participants were asked to watch short pre-recorded videos made by the faculty, as well as view slides and video transcripts.

3. "*Putting it together*" *exercise*: After the morning substantive sessions, faculty and participants engaged in a real-world simulated exercise. This activity required individual reading and reflection, as well as small- and medium-sized group discussions, on a pre-prepared TDM scenario.
4. *Free Write*: Day 2 ended with another 15-minute free write exercise, with prompts tied to the day's learnings:

    - What copyright concerns do you have about accessing data for your own projects? What about publishing it?
    - How do the projects you've worked on, supported, or encountered differ from the scenario you worked on during the Putting it Together session?
    - What was your biggest "Ah ha!" moment of the day? What do you still find confusing?

Day 3

1. *Report back from moderator on free-write themes*: At the beginning of day 3, the moderator again summarized topics and progress communicated in the previous day's free writes.
2. *Substantive literacies Licensing, Privacy & Ethics*: On day 3, we explored the substantive law and policy literacies for text data mining having to do with licensing, privacy, and ethics. Participants had watched pre-recorded videos, and synchronous sessions were used for small group discussions.
3. "*Putting it together*" *exercise*: After the morning substantive sessions, faculty and participants engaged in another "putting it together" exercise. This time, however, the exercise was comprehensive of all literacies—requiring participants to apply not just the day's learnings but also tap into their copyright

knowledge from the day before.

4. *Free Write*: Synchronous sessions on day 3 ended with the final 15-minute free write exercise, in which participants reflected on the following prompts:

    - What strategies will you use to evaluate the ethical implications of current and future TDM projects?
    - What licensing issues surfaced for your own work? Where do you see a path forward and where do you feel stuck?
    - What made you feel angry today? What made you feel relieved?

5. *Preparation for Implementation Mapping discussion*: At the conclusion of day 3, we also asked the participants to prepare for day 4 by considering the following questions:

    - How will you provide guidance to others or integrate the literacies in your own practice? What concrete steps or actions will you take? Are there things that you, your institution, or the broader community should stop doing?
    - What challenges might you face with implementation of the literacies?
    - How would you like to collaborate with other Building LLTDM participants or other DH researchers / professionals to integrate the literacies into DH TDM practice? What would a high level roadmap look like to achieve this vision? What support or funding would you need to make this vision possible?
    - Are there aspects of the current legal landscape that would benefit from community cooperation and advocacy to better address and enable TDM research?

Day 4

1. *Report back from moderator on free-write themes*: The moderator summarized the free write motifs and lessons.
2. *Implementation mapping*: Faculty and participants convened in small groups to discuss their prepared thoughts on implementation mapping questions. Each group worked to identify next steps, needs, and plans for bringing the literacies to life in their work and at their institutions. We reconvened in a final plenary session to share plans and take-aways from the small group discussion, using the Mural tool to exchange virtual "sticky notes" viewable by all participants. Participants also had an opportunity to post "gratitude" messages to acknowledge or thank other participants, faculty, or recognize a particularly useful or impactful aspect of the institute.
3. *Participant Evaluation*: With impressions and lessons still fresh in their minds, participants completed an evaluation survey prior to attending a final optional "happy half hour."

Post-Institute Meeting

To model the "testing" phase of design thinking, we organized a 1.5-hour check-in meeting eight months after the institute. Our goal was to help the cohort reflect upon their implementation experiences so they could evaluate whether their strategies had been successful.

Approximately two months before the check-in meeting, we re-oriented the cohort to the literacies through a post-institute survey that inquired about their implementation plans and desires for follow-up programmatic resources.

One month before the check-in meeting, we asked participants to share brief (2-minute) videos documenting how they had been

supporting TDM legal literacies in their home institutions and projects. We offered the following prompts:

- What have you been thinking about or doing with respect to TDM?
- What's one lasting LLTDM lesson you remember from the Institute?
- What takeaways from the Institute have you been able to implement or share with others?
- What are you still struggling with when it comes to LLTDM?
- What are you proud of with respect to your LLTDM skills?

When we convened for the plenary meeting in February 2021, we began again with the moderator's reflections on themes evidenced in participants' videos. We then transitioned to small group discussions focused on successes, frustrations, or opportunities that the cohort had experienced in implementing the literacies. We concluded with a plenary group exercise to share individual and collective next steps brainstormed during the smaller discussions.

Open Educational Resource

In order to broadly share the materials developed to deliver the institute, we published an openly licensed ebook (open educational resource, or "OER") under the Creative Commons Public Domain Dedication (CC0). This means that the OER can be accessed, reused, and repurposed without restriction.

The OER serves two key purposes:

- **Substantive Literacies**: The first part of the OER covers all the legal literacies covered during the virtual institute, including copyright (both U.S. and international law), technological protection measures, privacy, and ethical considerations. We

hope this content will enable any member of the public to gain similar skills and insights as institute participants.
- **Pedagogy**: In the second part, we focus on pedagogy to help anyone who might want to teach the Building LLTDM literacies to others. It describes in detail how we developed and delivered the 4-day institute, and provides ideas and exemplars for hosting shorter instructional sessions. We also include our reflections on both substance and administration to facilitate effective teaching of Building LLTDM literacies by others.

The OER is published on Pressbooks, a web-based platform used to create and share ebooks and other OERs. The ebook is available in a variety of formats, including a web version and downloadable formats such as PDF and EPUB. We are publicizing it through our project website (www.buildinglltdm.org), the UC Berkeley Library blog, via email lists, and through faculty and participants' professional networks.

Impact, Reflections, & Next Steps

Impact

We analyzed participant evaluations and post-institute update videos and survey responses. We observed not only the lasting impact of the LLTDM literacies, but also a persistent sense of shared experience and community.

Confidence now abounds

One theme that arose early during the institute was the pervasive feeling of imposter syndrome among participants. It seemed to permeate this work, perhaps because as one participant so rightly observed, no one person can be a deep expert across an entire landscape of issues in text data mining, from corpus building and computation to legal and ethic issues and all of the many technical, intellectual, and labor issues that underpin the work. Yet in post-institute surveys, videos, and discussions, imposter syndrome was absent. Instead, participants commented about how much more confident they felt integrating the literacies into their work. This integration has taken a lot of forms, from licensing negotiations to establishing best practices in their labs. The key struggle transitioned from being unsure of one's skills to finding the time to apply them all.

Successful incorporation of ethics into TDM practices

Participants' closing reflections from the institute in June 2020 included a strong desire for taking an ethics-first approach to teaching the literacies and implementing text data mining projects. It has been heartening to see the many ways that participants are living these values by structuring ethics as a key component of their work. For instance:

- One scholar added a dedicated ethics section to a submitted paper involving the use of YouTube data.
- Another centered ethics in application of the literacies to a racial reckoning project at her home institution.
- A librarian has adjusted consultations with researchers to take

an ethics-first approach.
- A faculty member has shifted toward an ethics of care framework in working with students in the classroom and in his research lab.
- Several participants developed workshops and related materials that focus on ethical considerations when doing this work.

Participants also turned an eye toward institutional gaps where ethics are concerned. One video update reflected on the lack of oversight of privacy and ethical issues in TDM research, and the need for structures and education that will help with that intervention within our institutions.

Overall, the participants left energized to continue the conversation around ethics and contribute to developing ethics models that might guide TDM researchers in the future.

Community education

Across academic institutions, TDM expertise is both shared and distributed. It would be exceedingly rare to find any one person or even any one office prepared to address all of the technical, legal, ethical, and logistical nuances of text data mining. Several participants mentioned that it is difficult to build community due in large part to the dispersed nature of the work. Living and working through a global pandemic has not made that any easier.

Some participants nevertheless made some real gains in community building, and we can celebrate that. One participant described how they initiated conversations across their institution about text data mining to start thinking at an organizational level, and they also noted that they had formed relationships with the sponsored research office and with the faculty working group on data science. Another participant has taken up the idea of the "Data

Ombudsperson" and is working to introduce it to the scholarly communication group at their library. Yet another participant has established a new research cluster on "Critical Practice in Text Data Mining" under the auspices of their humanities research center. These kinds of connections hold the potential to make real progress within institutions that are notoriously complex.

Struggles with institutional risk aversion

One participant described institutional conservatism and risk aversion as their ongoing struggle. And another had hoped to push their institution to be bolder and braver, but it was not as easy as they had hoped. Seeding institutional change is long durational work and it begins with small acts of relationship building. We reinforced the need to celebrate these gains while striving for much bigger shifts in practice and perception.

Efforts to improve institutional licensing

Several participants have been working to break up their institution's licensing routines with various approaches to address TDM. One participant has been looking at the possibility of regularly including TDM language in institutional licenses, which is in keeping with the approach taken in the California Digital Library's model license agreement. Another participant started working on licensing terms and setting up contracts with vendors at their institution, and they ultimately preferred the use of a "Fair Use Escape Clause" rather than outlining specific terms for TDM. They discovered that in an attempt to be explicit, the terms that vendors found acceptable were too confining.

Participants also recognized the need to make the negotiated

terms visible to researchers. One participant has been taking that on with a database evaluation to outline who is eligible to use each resource, how the data may be used, and what content is available. Even when full licenses are not readily shared with the campus community, this kind of matrix can help users assess their options when working with content licensed through the libraries.

Development of workshops

Another way participants have been working with local communities is by integrating the literacies into their workshops and courses. One participant conducted an hour-and-a-half workshop and shared materials online. Two other participants collaborated on a workshop foregrounding privacy and ethics in DH projects, which is also available online. And yet another participant has put together a suite of relevant workshops associated with their research cluster.

One participant observed that the mere mention of copyright to students can lead to a lot of fear, uncertainty, and doubt, even when the intention is to empower people to understand their rights. It would be helpful to discuss potential strategies for mitigating that effect as part of our ongoing conversations with participants and the research community.

Pedagogical Reflections

The conversations during the institute and the participant feedback gave us much food for thought. We would like to expand our commitment to diversity by ensuring that the demographics of future faculty are as representative as those of the participants, and

that the questions and examples that animate discussion sessions themselves engage with issues of ethics, equity, and representation.

We also learned a few specific things that may shape how we approach immersive LLTDM trainings in the future:

Design Thinking is effective for teaching LLTDM

The institute empowered participants to understand the basic contours of the legal literacies for text data mining and apply them to their own work, whether that be developing their own TDM projects, advising DH researchers, or working with TDM issues in libraries and archives. The participants' own words from institute evaluations affirm the pedagogical efficacy:

- "I can say with confidence that I understand the four literacies better"
- "I really feel that I am coming out with much more both theoretical and practical knowledge than I expected."
- "I will be much more intentional at the outset of any TDM project about working through all of the pertinent literacies in a systematic way…the way the institute was structured into different literacies provides a repeatable framework to treat potential problems prospectively."
- "I am taking home a lot of new insights from this institute in combination with a feeling of empowerment that will allow me to reach out to the specialists and directors at my institutions in order to push for more TDM collaboration and a bolder approach concerning materials and datasets for international cooperation. I know now what the important legal issues are and how to use them to form my arguments and that is more than I could have wished for. Also, the institute broadened my perspective with regards to issues that I did not have on the radar that much at the beginning and I am looking forward to

engaging with these topics in the future, to integrate them into my teaching, and to advocate for them where I can."

Design thinking can also work in virtual instructive environments, as the pivot from an in-person institute to a virtual one was met with applause. In particular, the participants valued the interactive format with different touch points and small group discussions. Again, in their own words:

- "The deliberately thought through breakdown and mix fostered incredibly valuable discussions and I would hope this kind of framework is used as a best practice for future DH institutes of all kinds going forward. Also, thank you for such an amazing virtual experience which I can only imagine took a tremendous amount of work to coordinate and plan with limited time to shift to an entirely different format–I was overjoyed to critically engage with complex subjects and for the chance to get out of my everyday pandemic routines."
- "I found this to be the best example of how to manage hands-on learning in a virtual environment. I think the planning team did a fantastic job pivoting to a fully online environment without losing the feel of an in-person intensive."
- "The multi-modal communication (Slack, Mural, Zoom) enabled far more interaction than I anticipated."
- "This is by far the best organized event that I have ever attended. The content was by far the most substantive. The faculty were by far the most engaged. A+ across the board."
- "The flipped learning approach, combined with design learning elements, really worked well. The lecture/video materials and reading in particular were well presented and selected, and I really appreciated that we could do that at our own pace. The overall topic of this gathering was well chosen in that it could allow for us to do focused seeking of answers to questions but in a way that had real practical consequences for how we could change the world of TDM research.

Copyright is a straightforward literacy to teach

Questions about using material under copyright were at the forefront of participants' minds when they entered the institute, but those concerns evaporated quickly. The copyright portion of the curriculum addressed copyright and the fair use exception extensively, and its applicability to TDM work was solidified. Unexpectedly to many, copyright risk issues turned out to be relatively straightforward and largely confined only to corpus republishing. As a result, participants felt empowered to perform analyses on copyrighted materials. One participant said, "I also feel compelled now to do my own research and take advantage of the expansive idea of fair use to examine contemporary, creative works," and another "was mainly relieved that my TDM project was transformative enough to not violate copyright." The greater challenge the cohort recognized was finding ways to educate our communities about the full scope of what fair use allows for TDM.

Literacies should be woven into research project plans

As scholars and educators, we should be building a legal literacies workflow into DH project planning from the very beginning, and refer to it throughout the project lifecycle. Too often, copyright and other legal considerations are unexamined or brushed aside to the detriment of DH research, partly due to lack of confidence in these areas or fear of institutional or rights-holder reprisal. Institute participants suggested ways of instantiating a lifecycle approach to literacy integration into DH project planning—including intermittent trainings, online guidance about process and sample documentation templates, and building legal questions into the project management process for DH support work. One participant

said, "In our library's center for digital scholarship, we need to develop a better charter/MOU/agreement system for digital projects that will at least touch on data management (DMPs), legal implications (copyright, etc), collaborator expectations, and ethics."

Institutions need support for adopting TDM-friendly licenses

Licenses with publishers, vendors, museums, and other content providers can further restrict uses that would otherwise be allowed under copyright law. While licensing restrictions can be frustrating when their terms impede the assembly of corpora or application of automated corpora analysis, participants learned what a TDM-friendly license might look like, such as one with terms that specifically allow for TDM uses or that contain a fair use clause. Participants were interested in shaping their institutional licenses—but desired additional instructional materials focused specifically on advocacy and negotiation support.

Ethics should be front-and-center

While participants entered the institute focused on questions of copyright, many reported leaving with their copyright questions solved and their ethical questions awakened. As one participant wrote, the institute "erased my anxieties in target areas and introduced whole new considerations in areas like ethics. It answered my questions and left me thinking." We believe questions of ethics loomed large not only because of the critical importance of ethics when addressing data at scale, but also because of the relative absence of guidelines and best practices to help guide us in this area.

We quickly realized that although we discussed ethics as the final substantive literacy during the institute, it was difficult for participants to even begin thinking about copyright, licensing, and other legal issues before ethical considerations were addressed, especially given the institute's care for questions of social justice. As we repurposed the institute training and materials into the OER, we considered additional ways to emphasize and create discussions around ethics, and perhaps foreground ethics as the first step when thinking through DH projects, and in teaching Building LLTDM.

Next Steps

Overall, we are encouraged that the literacies and methodology developed and shared by the institute has empowered DH researchers to build and analyze their text corpora without fear, thanks to their being more secure in their knowledge of the law and ethics. We hope these literacies become rooted more broadly in DH curricula.

In the meantime, we have been considering two specific future courses of action: (1) development of cross-border training, and (2) creation of documentation templates.

Cross-Border Issues Need Future Institutes

Cross-border research collaborations emerged as a clear example of follow-on training that we believe is necessary. Although we had initially intended to focus mainly on U.S. law for most literacies, cross-border and foreign law issues pervaded given the broad range of humanities research in which our cohort engaged: Scholars are working with materials published under different legal frameworks, or are collaborating with others working in those environments.

This obviously complicates the legal landscape. Rather than offering clear answers to every question participants raised in the context of cross-border inquiries, we offered strategies for assessing and mitigating risk. Yet, the need for expanding or extending Building LLTDM to international and cross-border contexts is clear.

Need for Documentation Templates

While watching participants' update videos, we also observed their clever use of forms and documentation as tools to help kick start conversations that can ultimately shape practice. One participant described developing an MOU template for use in the digital scholarship lab that includes a section on the legal and ethical implications of the work. The template helps foreground these issues during the negotiation and ensures that they are addressed in the final agreement.

In a similar vein, another participant has been developing a rubric for designing new digital projects that incorporates the literacies and is grounded in the insight that it is best to begin by planning for the end. This presumably helps front-load conversations not just about data collection and corpus building but also representation and distribution for publication and long term preservation. To socialize these practices with graduate students, another participant has started requiring a data management plan for student research projects conducted as part of his research lab to ensure everyone in the lab is thinking deeply about ethics in data collection, dehydration, and eventual destruction for social media research. This approach simultaneously generates deep and thoughtful conversations while also making them expected and routine.

A comprehensive guide or set of customizable templates to document project development choices relative to the literacies is a sound direction for follow-on work.

Appendices

1. Building LLTDM Open Educational Resource
2. Participant Packet
3. Institute Videos, Slides, Transcripts
4. Reading List