RACHAEL

The purpose of this video is to provide you with rapid tips for how to navigate law and policy issues when you're scraping, archiving, or text mining third party content, like social media posts, website text or images, or articles from databases.

RACHAEL

We expect you're watching this video because you want to scrape and archive digital content, and use that content for text mining. Or perhaps because you wish to scan and extract information from print or digital books. These research processes are referred to as text data mining, or TDM.

As you know, TDM research techniques mean you can collect materials and examine them *en masse* through probing algorithms that reveal otherwise hidden trends and stories. TDM has been used to detect everything from the presence of race-based disparities revealed through police body camera footage, to comparing the vocabulary spectrum across rap songs.

But we also expect that you are watching this because you have concerns about conducting TDM as a research methodology.

- Perhaps you're worried that archiving or downloading materials violates the copyright of whoever owns the content.
- Perhaps you're worried about a database license agreement or the website's terms of use, which may expressly restrict downloading, scraping and text mining.
- Or maybe you have have questions about the privacy of the individuals in the materials you're collecting, or any ethics associated with working with them.

Today we're going to briefly explain why and how you should consider each of these law and policy areas in archiving and text mining projects, and in the course of doing so, we will give you tips about navigating each issue.

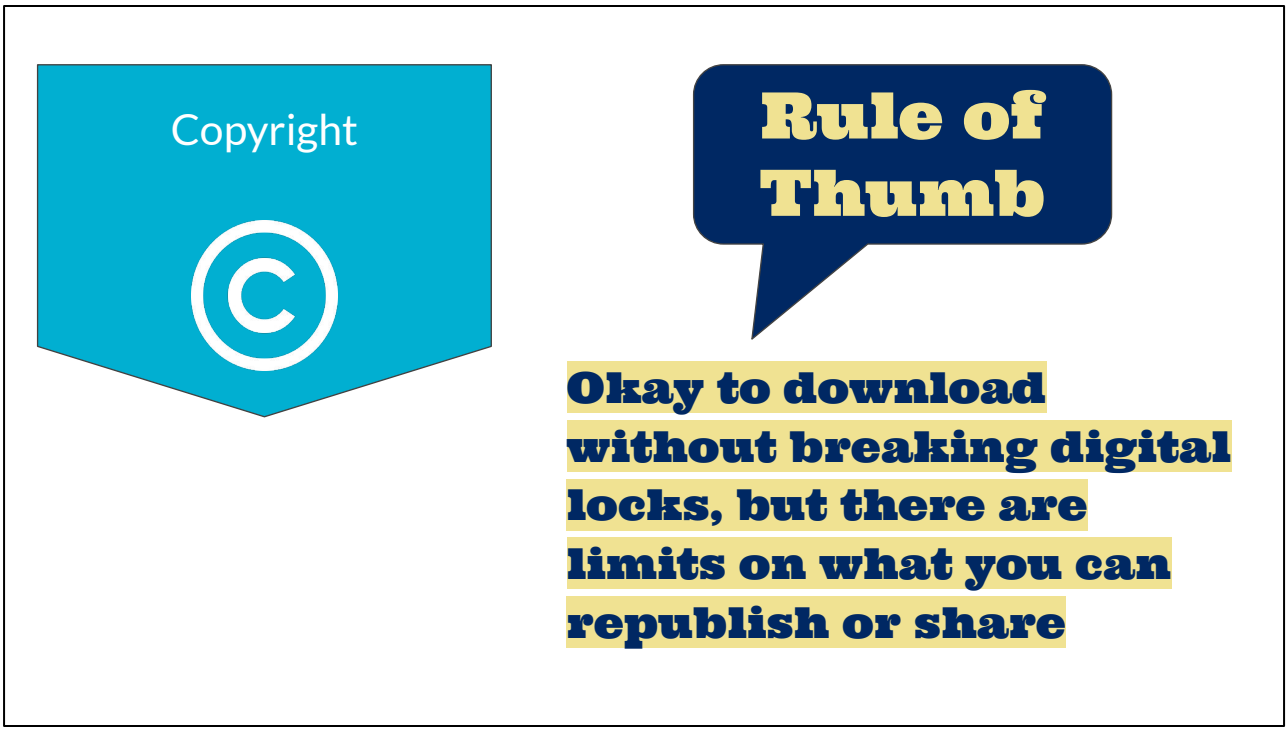
Quick disclaimer: We're not providing legal advice, but seeking to help you make your

own decisions about using and sharing materials based on a deeper understanding of the law, and in accordance with your own risk tolerance.

RACHAEL

Of course, in a 15 minute video, we can't provide you with detailed information on each of these law and policy matters, so we'd like to direct you to a great resource, which is a whole set of short videos we've made about each of these issues. We've conveniently compiled playlists of the videos that cover each of these topics in a lot more depth, and we recommend checking them out if you really need help and support in designing or executing your archiving or TDM project.

You can find all of these videos on our YouTube site, under playlists, at the URL shown here.



RACHAEL

Okay, so first up we'll talk about copyright and what you should know about it when doing text data mining.

Throughout this video, we're going to start with the rules of thumb first, and then explain how we got there. So the quick answer about copyright is:

> If you are downloading content from the Internet or a library database, or if you are scanning materials and using optical character recognition on them to perform your automated extractions for non-profit educational research, all of this has been considered fair use— meaning it's A-OK to do it as far as copyright issues go — provided you don't break digital locks called technical protection measures when doing so.

However, it's critical to stress that there's a difference between downloading and collecting your content to use for your automated extractions, and then *republishing* or *circulating or sharing* all of that same content with other researchers. The rule of thumb there is that:

> When it comes to actually sharing the downloaded content you've collected with others, that's when you might hit the limit of what is considered fair use. So you need to carefully think about how much of what you're collecting you can republish or share with others.

Now we’ll explain what all o this means, beginning with demystifying copyright.

RACHAEL

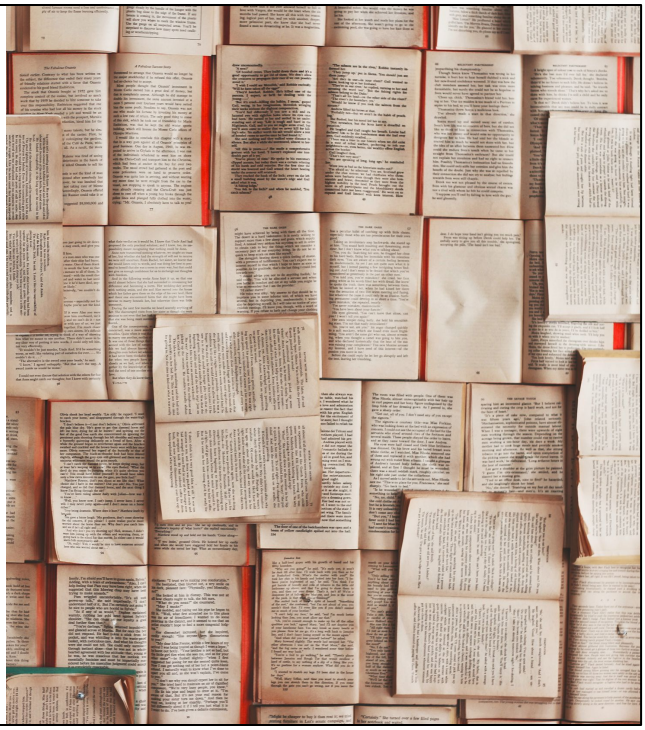
To understand why it's okay to download or compile materials to conduct TDM, but why it is not necessarily okay to republish or reshare everything you've collected, we have to first understand what copyright is. Copyright is a collection of rights given both to content creators but also the public. These rights and exceptions or limitations are all set forth in what's called the U.S. Code and, in the case of copyright, all of the laws are in Title 17 of the U.S. Code.

The very reason copyright laws exist is because the Constitution of the United States authorized the U.S. Congress to enact laws that would encourage people to create and write things. Specifically, the Constitution wanted Congress to promote progress in arts & science by rewarding authors with exclusive rights to their writings or art for a limited period of time. But the Constitution also told Congress to limit the length of time that these exclusive rights for authors last because, if they lasted indefinitely, this would cut against the competing goal of having the public able to make use of the writings in order to continue advancing science and art.

So, at its core, copyright is really simple and can be thought of just as the exclusive rights that authors hold to their original works of expression, with the understanding that these exclusive rights don't last forever to ensure that you can make use of the works, too.

Exclusive Rights

- Reproduction
- Distribution
- Derivative works
- Public performance
- Public display



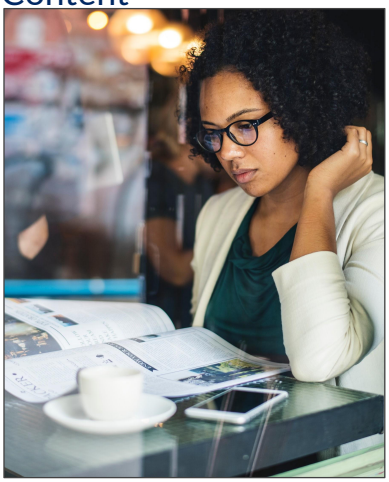
RACHAEL

So, what are these exclusive rights that an author gets when they write or create something? Well they are often referred to as a bundle of the following rights:

- Reproduction: That means in the context of writing a book, if I'm the author, I'm the only one who can make copies of what I'm writing.
- Distribution: I'm also the only one who can distribute copies of my book.
- Derivative works: I'm the only one who can write a second edition, or an adaptation of the book, or turn it into a movie (all of those are considered "derivative works")
- Public performance: With respect to public performance, I'm the only one who can read the book aloud to a public paying audience
- Public display: And finally, I'm the only one who can take each page of the book and display it publicly for everyone to see.

TDM researchers can use copyrighted content!

Content



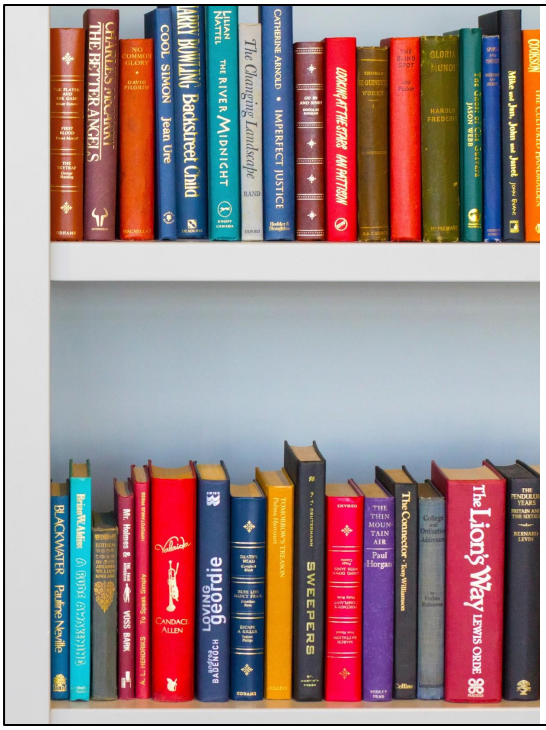
Data about the content

RACHAEL

So if authors have these exclusive rights under copyright law, why did we say that text data mining is okay for you to do? Why is it okay for TDM researchers scrape, download, reproduce, or publish work that's subject to copyright?

The answer is that creating reproductions to conduct TDM falls into an exception to these exclusive rights, an exception called fair use. And the main reason why TDM fits within that fair use exception is because digitizing or reproducing content in order to extract data about that content is very transformative, which is one important consideration under the fair use balancing test.

So let's dig in to what fair use is, and why U.S. courts have determined that TDM is both transformative and a fair use overall.



Fair Use

17 U.S.C. § 107

"The fair use of a copyrighted work…for purposes **such as criticism, comment, news reporting, teaching…, scholarship, or research**, is not an infringement of copyright."

RACHAEL

Fair use is an exception built into the copyright statutes to encourage uses of copyrighted work in order to engage in criticism, comment, news reporting, teaching, scholarship, or research.

So by its nature, it applies to the very kind of work TDM scholars are doing in exposing relationships or information within or between content.

Four-Factor Balancing Test

1. **Purpose & character of use**
   "Transformativeness" often dominates
2. **Nature of copyrighted work**
   Whether factual/scholarly work
3. **Amount and substantiality**
   Size & importance of portion
4. **Effect on potential market**
   Whether it supplants market

RACHAEL

Section 107 of U.S. Code Title 17 goes on to suggest four factors to be balanced when determining whether a use is fair.

1. **Factor one looks at why you're using the copyrighted content.** Are you doing so in an educational context, and are you adding new insights or understandings to the copyrighted content? If so, then your use is more likely to be fair. In the context of TDM, if you take a bunch of novels and digitize them or download social media posts to run algorithms on them that show how certain language is used, then you are indeed adding new insights or understandings and transforming that text or content. So factor 1 leans in favor of TDM being a fair use.

2. **Factor two looks at whether the copyrighted work is factual and scholarly or more "creative"/expressive.** The more factual the content is, the more fair it is to use it. Depending on what type of material you're using to conduct TDM, this factor could weigh in your favor or not. Text mining scholarly articles technically might be considered more fair than text mining images. But overall, this factor winds up not being very determinative for a court, though it remains something to consider.

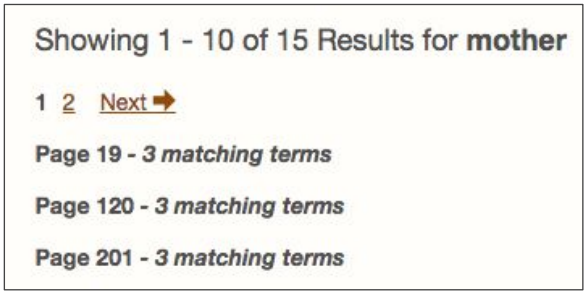
3. **Factor three looks at how much of the copyrighted work you are using.** The less you use, the more fair your use is. Well, with TDM, often you're

1. downloading or reproducing the entire work. So, technically this factor would weigh against fair use. But it's important to understand that fair use is a balancing test, and we are very strong under factor 1 with TDM. And we're also strong under factor 4, as I'll explain

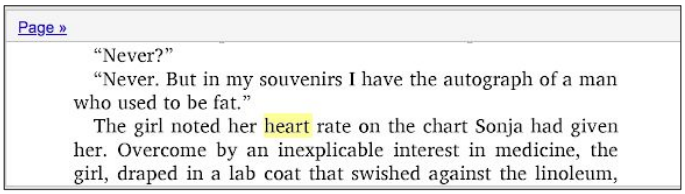
2. **Factor four considers whether your use supplants the sales or licensing market of the work.** And if what you're doing is extracting information from books or content already lawfully purchased or downloading social media content made freely available online, you're not supplanting the market for the author or rightsholder to sell more books or content. So factor 4 also weighs in favor of TDM being a fair use.

Therefore, on balance, text data mining is very strong with factors 1 and 4, and as such courts have determined that digitizing or downloading material to conduct TDM is considered a fair use overall.

Authors Guild v. HathiTrust
755 F.3d 87 (2d Cir. 2014)
Textual analysis that digital library enabled was transformative under factor one, and overall fair



Authors Guild v. Google
804 F.3d 202 (2d Cir. 2015)
Creation of full-text searchable database with "snippet view" and "ngram viewer" [search strings] were fair uses



RACHAEL

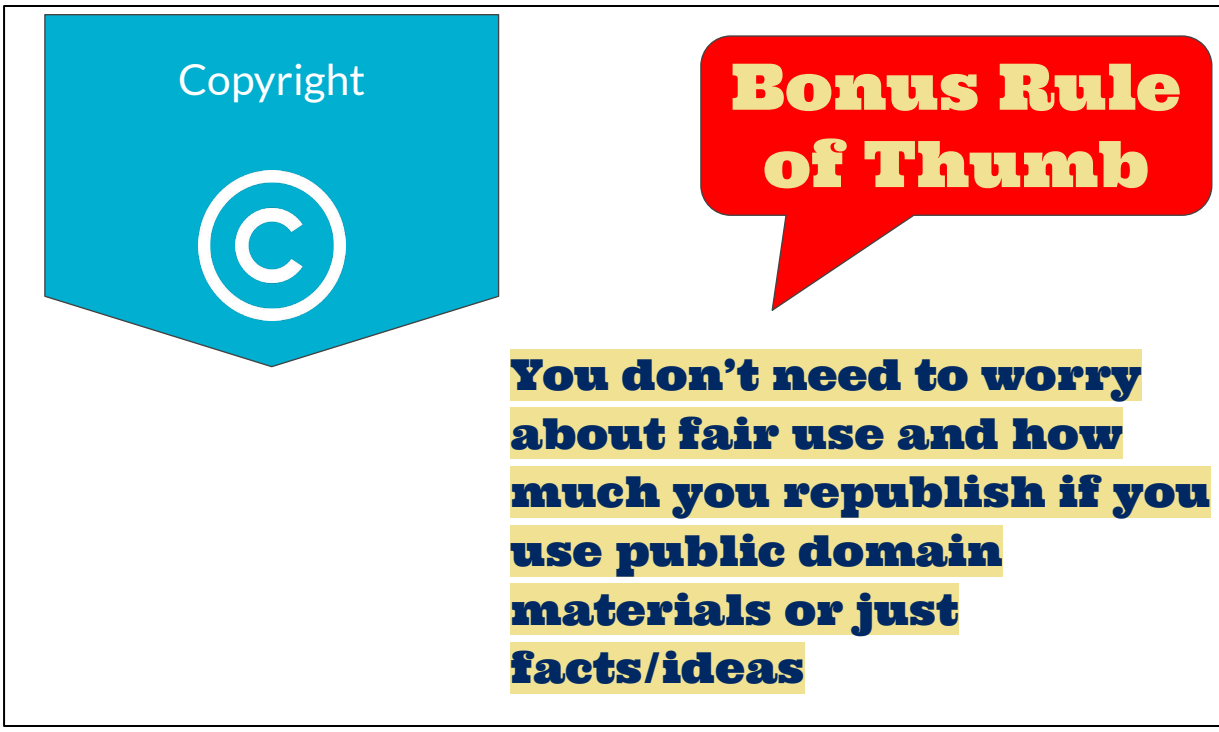
If you're interested in learning more about the court cases that have found TDM to be both transformative under factor 1, and a fair use overall, we encourage you to check them out.

One of the leading cases relates to the HathiTrust digital library, which scanned books and allows you to search within them for the number of occurrences and locations of various words. What you get to see is the result list of where your search terms occur, but you don't get to see the underlying text of the books. That's because: It's fair use to digitize and do the searching, but may not be fair to then display all of the protected content you compiled.

A similar result arose in a case involving the Google Books project, which digitized books and allows you to search for strings of words. Google Books then displays a very short snippet of where the search string occurs in the book. Once again: It was fair use for Google to digitize and allow algorithmic searching, but Google had to limit how much it could display or republish of the underlying books.

So to recap the rule of thumb for copyright and text data mining: If you are downloading content from the Internet or a database, or scanning materials and using optical character recognition on them to perform your automated extractions, this has been considered fair use in the context of research—provided you don't break digital locks called technical protection measures when doing so. But the court cases make

clear that the amount you are allowed to redistribute, display, or republish from the original copyrighted content has limits—meaning you may not be able to share the content you've downloaded; doing so could exceed fair use.



RACHAEL

Two last copyright tips to keep in mind.

1. Not everything is protected by copyright, so in some cases you don't even need to worry about whether republication is fair use. Remember we said that copyright lasts only a certain period of time. Well, after that time expires, the material is in the public domain—meaning not protected by copyright —and you don't have to worry about copyright anymore. In addition, federal government works are automatically in the public domain.

2. Copyright protects creative expression, not facts. So, for instance, researchers just seeking to download and mine information about statistics or citations—and mapping information about who gets cited and where/how—are not using any content with expressive elements. That is, the content they're using isn't protected by copyright to begin with, because citations are just facts. So, these researchers also don't have to worry about fair use or other copyright constraints when it comes to republishing that content.

TIM

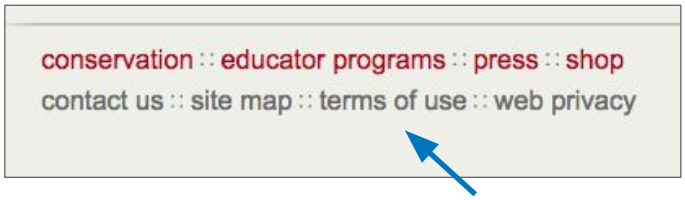
Now we'll turn to the next topic area: Contracts: Just as we did with copyright law, we're going to start with the rule of thumb for contracts and then explain how we got there.

And the takeaway we want you to have for contracts related to archiving and mining data is:

> Regardless of what's okay under copyright law—it doesn't necessarily mean you're free to download, create, and circulate your TDM corpus. Why? Because there may be a variety of different contracts that supercede what's allowed under copyright law. So, we need to pay attention to contracts, including license agreements and terms of service, because 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.

So, we'll now explain what those contracts are, and what you should be aware of when reviewing them.

Review Website Terms



"If you intend to quote extensive amounts of text, use other original content, or **reproduce images** from this site, please contact us for **permission**."

TIM

One of the most common types of contracts that can affect what you are allowed to do with respect to TDM are a website's terms of use, sometimes also call its terms of service.

For example if you want to harvest user posts from Facebook for a project, Facebook's terms of service govern both how you're allowed to search Facebook, and what you're allowed to download and do with the downloaded content. 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.

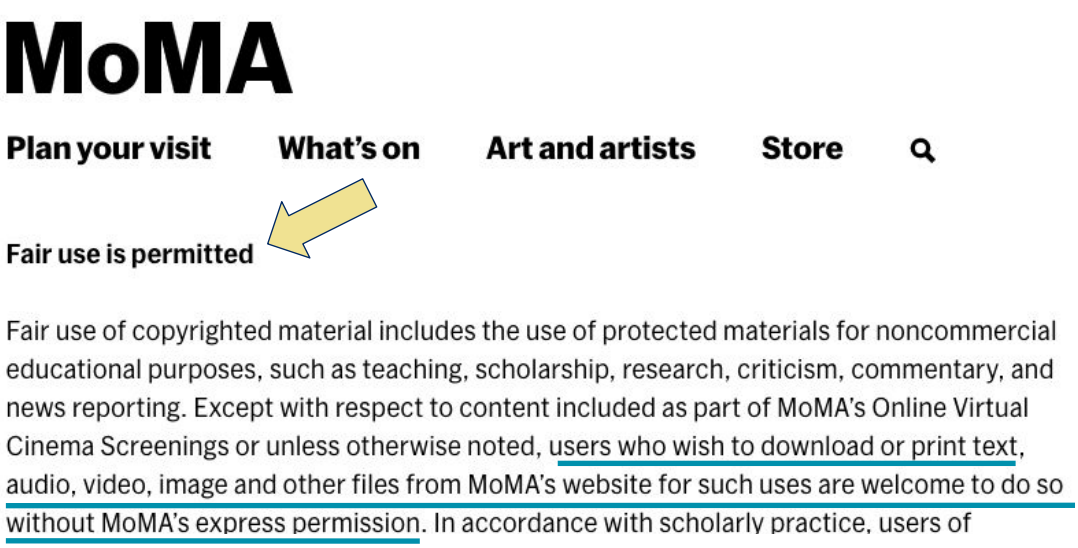
Take a look at the sample website terms on this slide. Here, the site owner wants you to get permission for reproducing any images you download, even if doing so would be fair use.

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—where copyright is governed. 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 things 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.

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. So, the relevant question is: what did the website's owner suffer if its publicly available content was used for textual analysis? Maybe nothing, but that doesn't mean they won't sue to try.

Look for Fair Use "Savings Clauses"

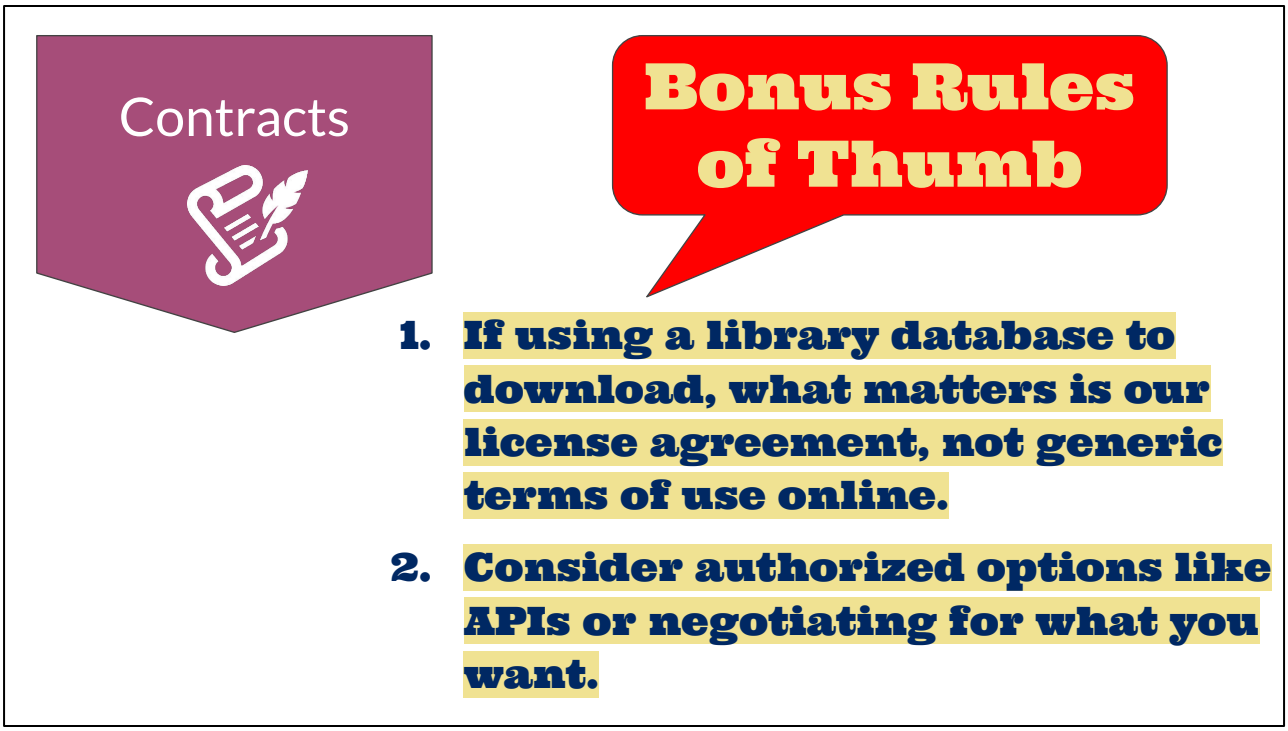


TIM

One good tip when reviewing website terms of service is to look for what's called a fair use "savings clause."

On the previous slide, the website wanted you to get permission from them to do any kind of reproduction of their content. Well, when we look at the terms of service on a different site, in this case here the Museum of Modern Art, the terms provide that you can do whatever is considered fair use. And remember, as we learned in the earlier copyright segment of this video, text data mining is indeed a fair use. So, by inference the MOMA website is authorizing TDM research, provided you stay within the bounds of fair use when doing it. In other words, the MOMA website terms **save or preserve your fair use rights.**

That's why these are sometimes referred to as fair use savings clauses.

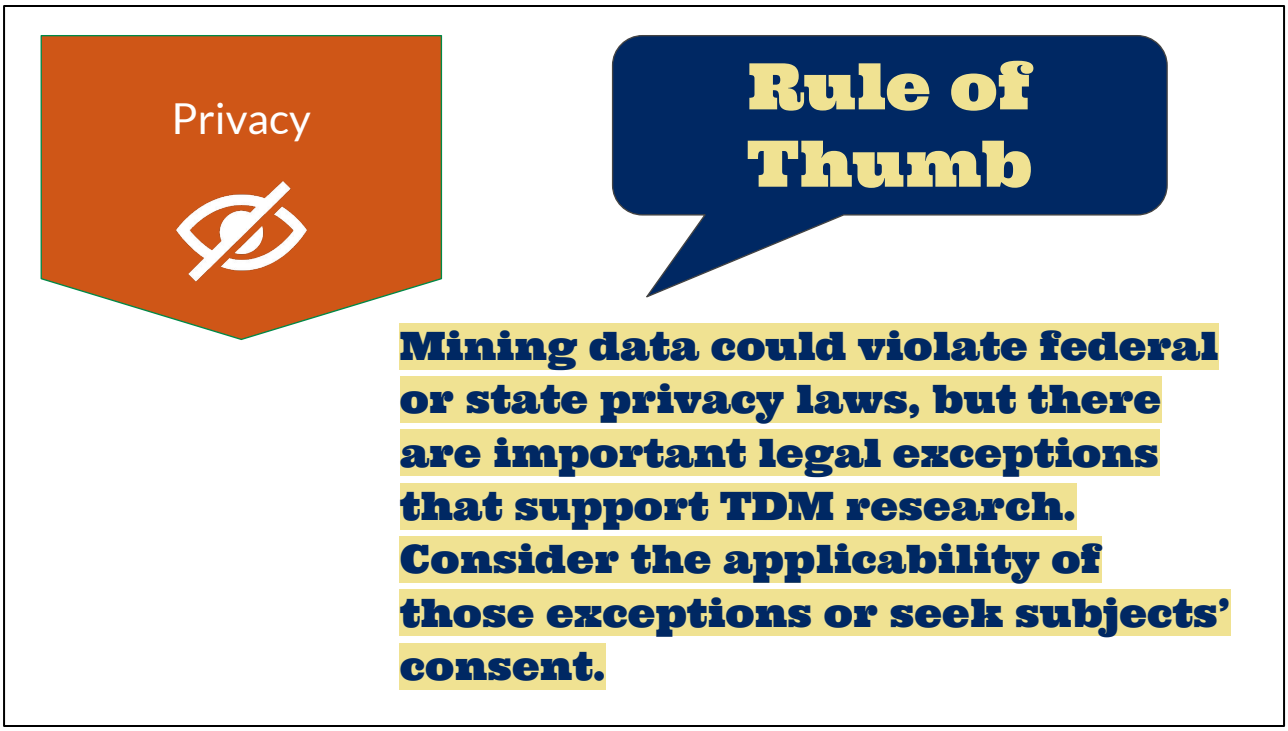


TIM

Two more rules of thumb when considering contracts in your TDM project.

1. First, if you're downloading or mining content from a library-licensed database, what matters is not the generic Terms of Use you find online for that database, but the license agreement that the library actually signed with the database provider. You should contact the library to find out what that license agreement allows with respect to TDM. Often, libraries will try to build fair use savings clauses into the agreements they sign with vendors.

2. Second, even if the terms of use for the website or database restrict or prohibit text mining, the provider may offer an application programming interface, or API, with its own set of terms that allows scraping and TDM. If you're still out of luck, you could also try contacting the provider and requesting permission for the research you want to do.



RACHAEL

Okay we're moving on to considering privacy issues in TDM.

Privacy law is pretty complex and is better treated in that series of videos we mentioned. But the most important takeaways for TDM are the following:

Assuming you're not working on behalf of the government, the two main sources of privacy laws for you to consider when text data mining are federal statutes and state tort laws. There are many federal privacy statutes that control what data you're allowed to access and what you can do with it. For instance, there's the Fair Credit Reporting Act, the Family Educational Rights and Privacy Act (FERPA), and the Health Insurance Portability and Accountability Act (HIPAA). You'll of course need to comply with all of these.

But when you're dealing with mining data that is **not** covered by federal privacy statutes, you still need to think about state privacy statutes or other state privacy case law. State privacy laws are meant to protect against things like unlawful intrusion upon people's seclusion, or public disclosure of embarrassing personal facts. But these state privacy laws have important exceptions that may permit you to conduct your research.

Exceptions & limitations critical for TDM research

1. **Newsworthiness / public interest**
    i. Also, for public disclosure tort: First Amendment
2. **Death**
3. **Person not identifiable**
4. **Permission / voluntary disclosure / waiver**

RACHAEL

So, what are the typical and applicable exceptions to state privacy laws?

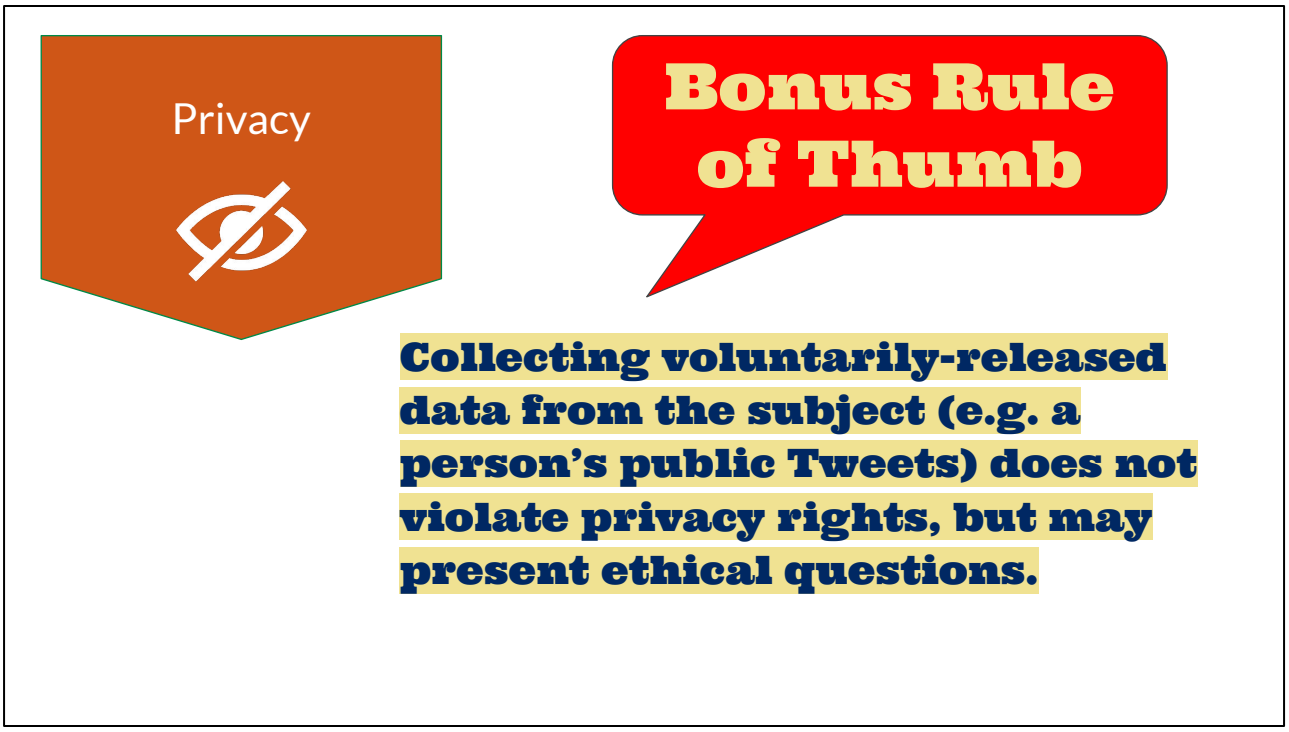
1. First, the right of privacy is not violated by disclosures of matters of legitimate public interest. Additionally, specifically with respect to public disclosure of private facts, courts also have to balance a person's right to keep information private with your First Amendment right to disseminate information to the public. In achieving this balance, courts sometimes look to whether the facts you're seeking to disclose disclose are of legitimate public concern and/or would be highly offensive to a reasonable person.

2. Second, when a person dies they lose 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 about people who are deceased that would typically be protected by state (as opposed to federal) laws, the state laws usually no longer apply.

3. Third, there are no privacy concerns if the people are not identifiable from the information you release.

4. And fourth and finally, if someone has released the information themselves—such as by posting the content voluntarily on social media

1. sites—or given you permission, they cannot sustain a privacy tort claim.



RACHAEL

Now, before we move on to the final topic area, we want to point out that when you're collecting content that fits within a privacy exception we just discussed, or if it's content that someone voluntarily disclosed, then you may **not** be violating someone's legal rights to privacy—but your actions might still have ethical implications. Ethical norms are not a matter of law in the United States, but may still be important for you to consider. So we'll talk about them next.

TIM

Imagine in your social media archiving and text mining project, you are collecting social media posts from victims of domestic violence. There are no privacy concerns if the posters have voluntarily divulged information about themselves. But, by collecting, analyzing, and resharing the content, you might be amplifying the posts related to domestic violence and making the victims, or perhaps the perpetrators, easier to find and potentially subjecting them to harm.

Ethical rules that focus only on the characteristics of the data itself ignore the ethics of what we do with that data.

Consider, for example, the published use case of an algorithm created to detect someone's sexual orientation from a photograph. The training data for this algorithm came from profile photographs collected from online dating sites, and in this case the images were already labeled with sexual orientations.

However, as acknowledged in the research paper for this example, a different use for this kind of algorithm could be the identification and persecution of people in a country where homosexuality is illegal. So the question is, what responsibility should you take as a data collector and extractor in trying to anticipate and evaluate harm.

As we mentioned, there are no legal answers to these ethical problems. All we can do is look for various types of guidance.

Ethics Considerations

- Develop local best practices (e.g. conduct decision-making within a research group)
- Consult journal publication guidelines
- Consult professional association guidelines
- Undertake community engagement
- Impose access controls (e.g. registration/ID to view)
- Seek IRB involvement / approval
- Adopt a new ethics/privacy paradigm

TIM

As to that guidance, researchers often consider a wide range of sources and options, including:

- Developing agreed-upon practices within a research group
- Consulting their research advisors
- Consulting journal publication guidelines
- Consulting professional association guidelines and best practices
- Undertaking engagement with the affected communities, particularly if the researchers wish to build long-term relationships with them
- Seeking institutional review board involvement or 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 (not to mention overwhelm some IRB offices), so some fundamental structural changes at your institution might be needed if want to go down that route
- Finally, you could even think about adopting a specific ethics/privacy paradigm. Unless you adopt a strict "do no harm" approach, you may need to

- develop a balancing test that you like.

Next up, we'll take a look at one example of a balancing test.

Balancing Principle

Does the **value** to cultural communities, researchers, or the public outweigh the **potential for harm or exploitation** of people, resources, or knowledge?

TIM

If you are interested in considering specific ethics paradigms, we'd recommend our ethics video playlist for an overview of some of the leading theories. But briefly, we just want to highlight an offshoot of the framework called the "ethics of care" because it might be of interest to you.

An ethics of care model would consider impact on content creators or the subjects of the content you're using, infusing empathy and responsibility, relative to any embedded power structures that affect the creators or subjects. Essentially, what we chose to do as information collectors or analyzers will affect other people, particularly when people have less structural power, and according to the ethics of care, we should care about that.

Through its focus on relationships, an ethics of care approach also enables a progression from accounting for the rights and obligations of individuals, to the rights and obligations of groups. This mean means we can talk about not just an individual in a photograph, but about where and how, for example, Japanese American

internment materials should be archived or algorithmically analyzed.

The UC Berkeley Library has adopted a form of ethics of care in our approach to making decisions about what collection materials we'll digitize and put online. Our version of ethics of care is framed as a balancing principle:

That is, we look to

> ***whether the value*** to researchers, the public, or cultural communities in digitizing and sharing the content ***outweighs the potential for harm or exploitation*** of people, resources, or knowledge.

The balancing framework is just a suggestion, and for your individual projects you could consider whatever values you want to uphold, as well as any relevant community or disciplinary standards you wish to follow as you're developing and publishing digital projects.

And it's a good idea to make these choices clear to your viewers and contributors. Let them know about the ethical considerations in how they are using and repurposing the data, especially considering that it could be combined and displayed in ways not originally intended.

Law & Ethics in Archiving and Mining Data

TDM law & policy questions
schol-comm@berkeley.edu

Video playlists
https://www.youtube.com/channel/UCNUMwTyK0raTNNZVjhgB7KA/playlists

TIM

We hope this overview was helpful. If you're a UC Berkeley community member, we invite you to contact us with questions at the e-mail address here.

And our video playlists going into all of these issues in more detail are available to everyone on our YouTube channel linked from this slide.