Text and Data Mining: Negotiating Computational Access to Library Resources

Citation: "Text and Data Mining: Negotiating Computational Access to Library Resources," Katie Zimmerman (pgs. 197-212), in Copyright: Best Practices for Academic Libraries, Donna L. Ferullo and Dwayne K. Buttler, eds., 2023.

As Published: https://rowman.com/ISBN/9781538168219/Copyright-Best-Practices-for-Academic-Libraries

Publisher: Rowman & Littlefield Publishers

Persistent URL: https://hdl.handle.net/1721.1/153820

Version: Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

Copyright: Best Practices for Academic Libraries

Link to published version: https://rowman.com/ISBN/9781538168219/Copyright-Best-Practices-for-Academic-Libraries

Chapter Title: Text and Data Mining: Negotiating Computational Access to Library Resources

INTRO

Text and data mining (TDM), also sometimes called computational access, refers to computer-mediated analysis of massive amounts of text. Unlike traditional library use, where the end user reads content directly, TDM instead uses the corpus of library content as data and applies machine learning techniques to draw conclusions from vastly higher quantities of text than a single human could process. Text and data mining can be used in a wide range of analyses, from analysis of depictions of gender in popular literature,[1] race and prejudice in news coverage,[2] or predicting new chemical reactions based on the existing chemistry literature.[3]

As machine learning becomes increasingly common as a research tool, it is increasingly necessary for academic libraries to be able to provide computational access to library resources. Making library resources accessible for TDM, however, implies copyright and licensing considerations, both for researchers and libraries themselves.

COPYRIGHT CONSIDERATIONS

In many ways, text and data mining should not implicate copyright concerns any more than use of any other tool. If a reader has the right to read the text, they should also have the right to include it in a scientific analysis, even if that analysis is mediated by technology.[4] Practical application of text mining tools usually requires making copies of the texts, however, and the volume of copying required for TDM can make the rightsholders of those texts very nervous, which can be a recipe for litigation. A good understanding of the copyright implications of TDM is therefore essential for librarians providing support to patrons making computational use of library resources.

Fortunately for researchers and the librarians seeking to support them, text and data mining is well-established in the statutory exceptions and limitations to copyright. Internationally, a recent survey of national copyright laws found that approximately one quarter of countries have copyright provisions

---

[1] "Computational Reading of Gender in Novels 1770–1922," *The Gender Novels Project*, accessed December 15, 2022, http://gendernovels.digitalhumanitiesmit.org/info/gender_novels_overview.

[2] Rochelle Terman, "Islamophobia and Media Portrayals of Muslim Women: A Computational Text Analysis of US News Coverage," *International Studies Quarterly* 61, no. 3 (2017): 489–502, accessed December 15, 2022, https://doi.org/10.1093/isq/sqx051.

[3] Jiang Guo et al., "Automated Chemical Reaction Extraction from Scientific Literature," *Journal of Chemical Information and Modeling* 62, no. 9 (2022): 2035–2045, accessed December 15, 2022, https://doi.org/10.1021/acs.jcim.1c00284.

[4] See , e.g., Heather Joseph, "The Right to Read is the Right to Mine…," *SPARC*, November 19, 2015, https://sparcopen.org/news/2015/the-right-to-read-is-the-right-to-mine/; https://github.com/ContentMine.

that allow for most computational research applications.[5] An increasing subset of countries have recognized explicit exceptions for text and data mining, including the United Kingdom,[6] Germany,[7] Japan,[8] and Estonia.[9] An explicit exception provides welcome clarity to researchers in that jurisdiction that computational research activities do not run afoul of copyright, as long as the conditions giving rise to the exception continue to apply. A researcher in Germany, for example, would enjoy the legal right under German law to assemble and analyze corpora of copyrighted works, and share corpora with research collaborators and for data validation. Explicit TDM exceptions also come with limitations, however: in the German example the TDM exception only applies while the research is non-commercial; if a research project made the jump from academic to commercial application, they would need to reassess the legal basis of the work.

FAIR USE

In the United States, text and data mining is well established under the general "fair use" copyright exception.[10] Unlike countries with an explicit statute about TDM, US copyright law does not mention text mining specifically. Instead, it is covered by fair use, which by its design is flexible enough to cover a wide range of applications and has been repeatedly interpreted by the courts to cover computational access to copyrighted works.

There are two main cases on this topic, *Authors Guild v. Google*[11] and *Authors Guild v. HathiTrust*,[12] both involving similar facts. Both cases stemmed from mass digitization done by Google in order to create machine-readable scans of books. Starting in the early 2000s several research libraries partnered with Google in the Google Books Library Project;[13] libraries had the books, and Google had the scanning capacity. Copies of the scans coming out of the project went into Google Books[14] (for Google) and the HathiTrust Digital Library[15] (for the libraries). In Google Books, Google makes the books available in several ways: if the book is in the public domain, the full text is available to read; if the book is within

---

[5] Sean Flynn, Michael Palmedo, and Andrés Izquierdo, "Research Exceptions in Comparative Copyright Law," *PIJIP/TLS Research Paper Series*, no. 72 (2021), accessed December 15, 2022, https://digitalcommons.wcl.american.edu/research/72.

[6] Copyright, Designs and Patents Act 1988, c. 48(29A) (incorporating amendments up to the Digital Economy Act 2017) (UK.) https://wipolex.wipo.int/en/legislation/details/18023.

[7] Act on Copyright and Related Rights, 1965, Section 60d (Copyright Act, as amended up to Act of September 1, 2017) (Ger.), https://wipolex.wipo.int/en/text/474263.

[8] Copyright Act, 1970 (Act No. 48 of May 6, 1970, as amended up to Act No. 72 of July 13, 2018) Art. 30-4(ii) (Japan), https://wipolex.wipo.int/en/legislation/details/18696.

[9] Copyright Act, 2017 (consolidated text of February 1, 2017), Section 19(3), (Est.), https://wipolex.wipo.int/en/text/510476.

[10] Copyright Act, 17 U.S.C. § 107 (2018), https://www.law.cornell.edu/uscode/text/17/107.

[11] *Authors Guild v. Google, Inc.*, 804 F.3d 202, 116 U.S.P.Q.2d (BNA) 1423 (2d Cir. 2015), accessed December 15, 2022, https://casetext.com/case/guild-v-google-inc-1.

[12] *Authors Guild, Inc. v. Hathitrust*, 755 F.3d 87 (2d Cir. 2014), accessed December 15, 2022, https://casetext.com/case/authors-guild-inc-v-hathitrust-1.

[13] "Google Books Library Project – An enhanced card catalog of the world's books," *Google Books*, accessed December 15, 2022, https://books.google.com/intl/en-GB/googlebooks/library.html.

[14] "About Google Books," *Google Books,* accessed December 15, 2022, https://books.google.com/intl/en/googlebooks/about/index.html.

[15] "Welcome to HathiTrust!" *HathiTrust Digital Library*, accessed December 15, 2022, https://www.hathitrust.org/about.

copyright, the full text is indexed so that users could search it, and Google will display page numbers and a "snippet view" consisting of several lines of text surrounding a search term. Additional safeguards sought to ensure that the snippet view function didn't provide access to the full text: only a certain percent of any given page is shown, and some pages are never displayed. HathiTrust provides similar functionality, returning only a list of page numbers on which a search term appears for books currently under copyright.[16] Most relevant for TDM research, Google Books also created an "ngram" tool[17] which enables some text and data mining of the digitized content. The ngram viewer indexes every word found in the millions of digitized volumes alongside basic metadata about where it was found, including date of publication and part of speech. This allows researchers to easily view and analyze changes in word frequency over time. Authors of books digitized and made available in Google Books and HathiTrust, led by the Authors Guild organization, objected to this use, and sued for copyright infringement.

In both cases, all of the uses described above were affirmed as falling under fair use. The court recognizes in these cases that the ultimate goal of copyright is to benefit the public, and that a robust fair use doctrine is essential to that purpose. A close look at the assessment of these cases under the four-factor fair use test is a helpful point of comparison for librarians and researchers assessing their own use cases:

The first fair use factor examines the purpose and character of the use. According to the Second Circuit court, the indexing and search functions are "quintessentially transformative use,"[18] decisingly favoring fair use under this factor. "Transformativeness" supports cases that use content for a new purpose, using the content as a building block for new creativity, rather than "merely repackaging or republishing the original[s]."[19] Allowing users to search a full corpus of one thousand books is a very different use case than actually reading those one thousand books. Rather than reading the direct content of the text, users are gathering frequency and location information. The ngram viewer, in particular (and by extension, other similar TDM applications), is a very different user experience than reading the books that constitute the corpus. Even the snippet view, however, which most resembles traditional reading, is satisfactorily transformative under the first factor, because it "is designed to show the searcher just enough context surrounding the searched term to help her evaluate whether the book falls within the scope of her interest (without revealing so much as to threaten the author's copyright interests)."[20]

The second fair use factor covers the nature of the work. This factor looks at whether the works being copied are closer to the creative core of copyright, or, alternately, whether we should give a greater latitude to fair use due to the informative nature of the content. In these cases the scanned books spanned many genres, and the court found this factor unpersuasive in either direction. Because the use doesn't substitute for reading the books, the specific content of the books essentially doesn't matter for purposes of the fair use analysis.

---

[16] HathiTrust copies were also used in full to provide access to print-disabled patrons.
[17] "Google Books Ngram Viewer," *Google Books*, accessed December 15, 2022, https://books.google.com/ngrams/.
[18] *Authors Guild, Inc. v. Hathitrust*, 755 F.3d 87, 97 (2d Cir. 2014).
[19] Pierre N. Leval, "Toward a Fair Use Standard," *Harvard Law Review* 103, no. 5 (1990): 1111 cited in *Authors Guild, Inc. v. Hathitrust*, 755 F.3d 87, 97 (2d Cir. 2014).
[20] *Authors Guild v. Google, Inc.*, 804 F.3d 202, 218 (2d Cir. 2015).

The third fair use factor looks at the quantity of work used. In order to create the Google Books and HathiTrust platforms, Google and the libraries had to copy the entirety of vast numbers of books. Importantly, however, they were not sharing the full text content with their users. In order to be able to search the full text of a book you need to copy every word in the book, but users of both platforms only ever had access to limited search results and aggregate data. Because the amount used was necessary for the transformative purpose under the first factor, and because public availability was limited in order to avoid inappropriate interference with the market for the original works, under the fourth factor (discussed below), copying even this vast amount of content was a reasonable fair use.

The fourth fair use factor analyzes the market effect of the use. The question under this factor is whether Google Books and the HathiTrust Digital Library will substitute for sales of the books that constitute the underlying corpus. Because access to the full text is limited, the courts conclude that they do not. Even the Google Books snippet view, which the court acknowledges may replace use of the underlying works if the user is looking up specific information, survives this factor. In the case of quick reference questions answered through snippet views, the court notes that factual information gleaned from the text is not covered by copyright anyway, and the chance of snippet view substituting for experiencing protected authorial expression is minimal.

Weighing all of the factors together, the court easily concludes that these are fair uses, and these cases provide a useful benchmark for other TDM projects. The exposure of the underlying corpus must be limited, but the TDM function itself, reasonable (snippet view-esque) public disclosure of the corpus as data, and the non-public mass copying required to facilitate the use, are all established fair uses.

IS IT COPYRIGHTABLE? IMPLICATIONS FOR RELEASING DATASETS

Another aspect to take away from the Google Books and HathiTrust cases relates to the nature of datasets that are created for TDM purposes. While the underlying corpus for a text mining project may consist of copyrighted material, the output of the project, and even the dataset that results after the data is cleaned and prepared for analysis, may contain none of the copyrightable expression from that material. To see this, let's consider the data required to power the Google ngram viewer.

The ngram viewer is designed to give the user word frequency indexed against a designated publication year. To do that, the creators of the dataset certainly had to start with the full text of the corpus, but the actual data needed for that result doesn't have to look anything like the original texts: the words don't have to be in order, and can be aggregated into the total count per word per publication year. Factual information extracted from or about the texts isn't copyrightable,[21] and the actual data for analysis could be an alphabetized list of word, date, and count, which contains none of the copyrightable expression from the source material.[22] Consistent with this analysis, Google makes the full ngram dataset (i.e. the full text of millions of books, organized alphabetically by word) freely available for download.[23]

---

[21] See *Feist Publications, Inc. v. Rural Tel. Service Co.*, 499 U.S. 340, 111 S. Ct. 1282 (1991).

[22] Even if the dataset can't be broken down quite that far, "snippet view" is a useful benchmark for providing subsets of text as data under fair use.

[23] "Google Books Ngram Viewer Exports," *Google Books*, accessed December 15, 2022, https://storage.googleapis.com/books/ngrams/books/datasetsv3.html.

This has implications for researchers constructing similar datasets and the librarians that are providing consultative services or helping source the research corpus. Usually it's not enough to be able to analyze the corpus – an academic researcher also needs to publish their results, and increasingly frequently, the dataset as well.[24] This can be problematic for researchers starting from a proprietary corpus, but keeping copyrightability principles in mind can help get a researcher to a releasable dataset. FAIR data principles[25] and good practice means that the fulltext original dataset should still be available to validate the cleaning and data preparation process (e.g., did you miscount?), but that access can be more controlled while still allowing for maximal public disclosure of the final dataset.

CONTRACT CONSIDERATIONS

Frequently, contract concerns are more problematic than copyright for text and data mining uses. Most electronic resources in libraries are subject to license agreements, which may constrain use beyond the restrictions inherent in copyright. When that happens, regardless of whether a use is permissible under copyright law, computational uses could be a breach of that contract. It is therefore very important for librarians licensing eresources to negotiate terms that do not (either explicitly or accidentally) prohibit text and data mining.

See chapter x of this volume for information on licensing in general. For text and data mining applications, specifically, certain clauses will be particularly relevant.

TEXT AND DATA MINING CLAUSES

Most obviously, if the contract addresses computational use or text and data mining directly, you should read those terms carefully. Academic publishers, who may be focused on human readability and other non-computational use cases, may prohibit text and data mining directly, or include specific provisions around conducting it. Here's an example that effectively prohibits TDM:

> "Systematic or programmatic downloading, printing, transmitting, or copying of the Licensed Material is prohibited. "Systematic or Programmatic" means downloading, printing, transmitting, or copying activity of which the intent or the effect is to capture, reproduce, or transfer the entire output of a journal volume, a journal issue, or a journal topical section, or sequential or cumulative search results, or collections of abstracts, articles, or tables of contents. Other such systematic or programmatic use of the Licensed Materials that interferes with the access of Authorized Users or that may affect performance of the Publishers' systems, for example, the use of 'robots' to index content, or downloading or attempting to download large amounts of material in a short period of time, is prohibited. example, the use of 'robots' to

[24] See, e.g., "Funder Requirements," *MIT Libraries*, accessed December 15, 2022, https://libraries.mit.edu/data-management/plan/funder-requirements/; Alondra Nelson, "Ensuring Free, Immediate, and Equitable Access to Federally Funded Research" (Memorandum for the Heads of Executive Departments and Agencies, Office of Science and Technology Policy), August 25, 2022, https://www.whitehouse.gov/wp-content/uploads/2022/08/08-2022-OSTP-Public-Access-Memo.pdf.

[25] Mark Wilkinson et al., "The FAIR Guiding Principles for scientific data management and stewardship," *Scientific Data* 3 (2016): 160018, https://doi.org/10.1038/sdata.2016.18.

index content, or downloading or attempting to download large amounts of material in a short period of time, is prohibited."[26]

This example doesn't mention "text and data mining" explicitly, and suggests that the overall concern is with potential service disruptions resulting from high-volume downloading, rather than objection to TDM itself. This starting point, is, perhaps, more promising than it first appears: if you can work out an access solution that addresses the technical concerns, that may be a viable way forward.

In the author's experience as a licensing librarian, explicit language to the effect of "text and data mining of this content is prohibited" is relatively rare, but you may still see it in several particular contexts. First, for database-type products where the product being sold is more similar to what a TDM researcher might produce themselves: the vendor in that case may be concerned that TDM could make an ongoing subscription to the content unnecessary after initial access, or provoke concerns about competition with the original product. Interestingly, the fair use basis for creating a TDM corpus from such content would also be reduced, due to a greater likelihood of market harm under the fourth factor, but the copyrightability basis for allowing it is likely greater: this situation is most likely when the database content is not, or is only minimally, copyrightable in the first place (e.g. databases of factual content), and the vendor is relying primarily on gatekeeping access via contracts, rather than underlying copyrights, to keep the content proprietary. This makes for difficult negotiations.

Secondly, you may see explicit prohibitions on TDM where the vendor offers a separate product specifically for TDM of the content. If they allow TDM of the content natively, they may reason, they're unlikely to get you to pay for their TDM product. In general, this should be resisted as rent-seeking behavior: if users have a right to conduct TDM under fair use, you shouldn't have to pay twice for the privilege of doing it. Some TDM products legitimately offer additional functionality – they may save researchers work preparing the data, or make TDM available to researchers with less technical expertise, for example – and those may provide value-added services that are worth the extra cost, but that should be separate and additional to TDM rights derived from general access.

Finally, you may also see terms like the prohibition against "systematic or programmatic downloading" above when the vendor has no objection to TDM, but has a different access method for it than for human-readable access. This can be a reasonable and mutually beneficial way to mediate TDM access, if handled appropriately, as we will see next.

When the vendor is on board with text and data mining of their content, it is a good idea to include affirmative language to that effect in the license. A good place to start this conversation is around the way that researchers will access the licensed content for TDM. If the vendor agrees that users may

---

[26] All example clauses in this chapter are drawn from actual eresource contracts. Citations provided are to publicly available versions. This example clause is common across several society publishers, e.g.: "Terms of Use," *ASCE Library*, American Society of Civil Engineers, accessed December 15, 2022, https://ascelibrary.org/page/termsofuse; "American Institute of Physics Multi-Site License Agreement," *California Digital Library*, clause 5.b, August 15, 2009, https://cdlib.org/services-groups/collections/licensed_resources/redacted_licenses/STAmericanInstituteofPhysicsLicenseAgreement_Redacted.pdf; "Electrochemical Society (ECS) Digital Library, Terms and Conditions 2018," *Northeastern University Library*, PROHIBITIONS ON CERTAIN USE (b), accessed December 15, 2022, https://library.northeastern.edu/research/a-to-z-databases/database-policies-and-terms/electrochemical-society-ecs-digital-library.

conduct TDM directly through the online journal interface, then you should make sure you're in agreement about what that looks like: should users limit downloading to a certain rate (and what is that rate)? If the vendor has an API for TDM access: do users need an API key, and how do they get one? If the vendor will provide bulk data upon request: how are those requests made, and what information needs to be provided? If a vendor has a well established TDM process, they will likely have pre-existing terms to describe it, which you should evaluate carefully to ensure they meet your community's needs. In general, you probably want access to be as seamless as possible, rate limits and similar restrictions should be reasonably tied to functional concerns, and information about the TDM project shared with the vendor should be minimized.[27]

Somewhat frequently, a negotiation around TDM rights will result in philosophical agreement that it is permitted, but the specifics won't be there yet. That's ok, too. In that case, adding language to the agreement to commit to meeting this need might be appropriate, for example:

> Licensor will cooperate with Licensee and Authorized Users as reasonably necessary in making the Publisher Content available in a manner and form most useful to the Authorized Users [for Text and Data Mining].[28]

FAIR USE SAVINGS CLAUSE

> Licensee and Authorized Users may make all use of the Licensed Materials as is consistent with the United States Copyright Act of 1976, as amended (17 U.S.C. §101, et seq.) including all limitations on and exceptions to the exclusive rights as provided therein.[29]

A fair use savings clause, such as the above example, is a clause that explicitly preserves the balance of uses envisioned under copyright law, by acknowledging that limitations and exceptions to copyright, including fair use and the library exceptions under section 108, apply to the content under this contract. The addition of this clause helps to ensure that you are not "contracting away" rights the user would otherwise have, and provides a backstop on use parameters which is consistent with the underlying law. This preserves and unifies the contract with the larger body of copyright law, and avoids having to reinvent the wheel in every case.

For text and data mining applications, a fair use savings clause is extremely helpful because, as we have seen, text and data mining is a well-established fair use under US law. If you have a correctly constructed fair use savings clause, TDM is included in your contract. This is usually not the end of the story – you will run into complications quickly if other parts of the contract can be interpreted as conflicting with this language, or if they impose undue impediments to actually using the preserved rights – but it is a very good place to start.

---

[27] If information about the project does need to be shared with the vendor, or if you can't negotiate it out, then the decision to share information about any particular project should be made by that researcher, rather than the library.

[28] See, e.g.: "Liblicense Model License Agreement with Commentary," *Center for Research Libraries*, clause 3.2(j), revised November 2014, http://liblicense.crl.edu/wp-content/uploads/2015/05/modellicensenew2014revmay2015.pdf.

[29] "NorthEast Research Libraries Consortium Generic License Agreement for Electronic Resources," *NERL*, "Authorized Uses" clause, revised April 23, 2012, https://nerl.org/wp-content/uploads/2019/06/NERLModelLicense-061019.pdf.

PROHIBITED USES CLAUSES

Other clauses that may affect the ability of library users to effectively obtain computational access to the content are frequently found in the "shall nots" section: specific actions prohibited by the contract. Consider again the clause quoted above on "systematic or programmatic downloading." This is a very common prohibited use clause, and is usually designed to avoid overloading the vendor's technical systems, and to prohibit mass redistribution of the content. Both of those goals are legitimate concerns, but the contract language should be written more specifically to address them, while not limiting legitimate computational uses. The risk here is that it could be completely permissible for users to conduct TDM (under fair use, and incorporated into the contract through a fair use savings clause), but any reasonable method of getting the content to the user in the volume required for TDM would still be a breach of contract. To avoid this, you could negotiate for removal of this language, or to specify that it doesn't apply to uses that are otherwise permitted under the contract (assuming, that is, you also have a fair use savings clause, or language specifically addressing TDM). If taking the latter approach, you may also want to specify what users should do to differentiate their permitted use from an unpermitted one: specifying times for permitted crawling, rate limits that will not trigger an automatic blocking process, (more cumbersomely) a notification process to whitelist specific uses, or an alternate access method altogether, are among the options here.

Another common clause that can be unintentionally problematic are clauses that say users may not "remove, obscure or modify any copyright notices"[30] that appear in the licensed content. Presumably, the intent behind this clause is to keep the rights metadata associated with the content, and to generally prevent misuse. What legitimate reason, a rightsholder might ask, would you have for removing the copyright notice? An answer to that is TDM. TDM usually requires data preparation that involves extracting information from the text or separating out sections of the text, which may, incidentally, remove copyright notices.[31] Instead, a requirement that, in the case of TDM outputs, the components of the corpus will be appropriately cited in the dataset (for example, by including source DOIs) may be a better approach which can also serve the researchers' needs.

DOWNSTREAM USES, RELEASE OF DATA, DELETION REQUIREMENTS

Occasionally you may find a contract that tries to regulate the downstream use of TDM output. For example:

> The Subscriber will provide [vendor] with drafts for review of any reports, papers or presentations or other written communications generated regarding [a TDM project] prior to submission for publication or public distribution.[32]

---

[30] E.g. "Science Online Journals Institutional License Agreement," *Science*, Annex A clause 2(g), accessed December 15, 2022, https://www.science.org/content/page/institutional-license-agreement; "Business Research Guide: Acceptable Uses," *University of Illinois at Chicago University Library,* last updated October 26, 2022, https://researchguides.uic.edu/c.php?g=252391&p=1683521; "Elsevier: Compendex, INSPEC & GeoRef Terms of Use," *Northeastern University Library*, July 2016, https://library.northeastern.edu/research/a-to-z-databases/database-policies-and-terms/elsevier-compendex-inspec-georef-terms-of-use.

[31] Removing the copyright notice could also be entirely appropriate, if the extracted sections are not copyrightable content.

[32] Stanford Libraries, "Warning: Elsevier Reaxys API License Violates Academic Freedom," April 3, 2018, accessed December 15, 2022, https://cpi.tamu.edu/meetings/CPI_Newsletter_April_2018_FINAL_with_Attachments.pdf.

This is extremely problematic for academic freedom, and opens the door to censorship if not properly limited. Agreements for limitations on disclosure of data are not inconceivable in academic spaces, but are usually in the form of data use agreements[33] which are regulated through an institution's office of sponsored research, and subject to oversight processes which are usually separate from the library. If an eresources license starts looking like a data use agreement, something is wrong. Approval requirements prior to publication of research results, audit requirements for security compliance, or preferential sharing of research results with the data vendor should raise red flags in any contemplated TDM agreement.

With that said, some licenses that include TDM might contain reasonable restrictions on what a released dataset should look like, and sometimes those can provide some reassuring clarity for the end user. A license that specifies that only reasonable portions of the licensed content be released, in no event to create a substitute for subscription access to the content, for example, could be reasonable. Similarly, reasonable citation requirements are common, and if they are truly reasonable, that could be acceptable. All downstream requirements should be caveated, however, to limit the library's responsibility to "reasonable efforts to inform" researchers of any requirements: generally librarians are not going to be directly in contact with all researchers doing TDM, and you want to avoid direct liability for compliance.

On the topic of data sharing, researchers frequently have collaborators at other institutions, and may need to share TDM data with those collaborators. This can be a licensing issue because collaborators at other institutions typically would not be authorized users under your library's contract, and standard article sharing clauses won't cover the volume necessary for a TDM collaboration. And while the publicly released dataset might be more streamlined, typically the researcher will need to share more than that with their direct collaborators, as well as with peer reviewers and others for appropriate validation of the research methodology and dataset. This is easily addressed by adding license language to cover collaborator access and for replication and validation,[34] and is a potential complication you can avoid by specifically including in your contract.

A final complication that sometimes arises in negotiations around TDM is requirements for the deletion of data. This issue is particularly likely to arise if the vendor usually interacts with the corporate or industry market, where such requirements are more common. This can be extremely problematic in the academic context, however, where researchers may build off of datasets for years, and should, if they're following good research practices,[35] keep copies for validation and replication. Such needs are unlikely to coincide with the duration of the contract term. A requirement that data received under the

---

[33] E.g. "Research Data Use Agreement," *Massachusetts Institute of Technology Office of Sponsored Programs*, last revised May 13, 2014, https://nda.mit.edu/images/MODEL-Research-DUA-2014-05-13.pdf; "Data Use Agreements (DUAs)," *University of Pittsburgh Office of Sponsored Programs,* accessed December 15, 2022, https://www.osp.pitt.edu/osp-teams/clinical-corporate-contract-services/negotiations/data-use-agreements-duas; "Data Use Agreements," *University of Massachusetts Amherst Research and Administration Compliance*, last modified October 5, 2022, https://www.umass.edu/research/guidance/data-use-agreements.

[34] See, e.g.: Association for Computing Machinery and California Digital Library, "2020-2022 ACM Digital Library Tiered-Band Open Access Model Pilot Agreement Terms and Conditions," *SPARC,* January 15, 2020, https://sparcopen.org/wp-content/uploads/2019/05/ACM-CDL-OA-Agreement-2020-2022.pdf.

[35] E.g. "MIT Research Data Principles," *MIT Libraries*, September 2019, https://libraries.mit.edu/data-management/mit-research-data-principles/.

agreement be deleted upon termination of the contract could disrupt research mid-project, and it's questionable if the library could even reasonably comply: imagine trying to contact every library patron who's used a given product and telling them they need to delete everything they've downloaded – most libraries would not have the user data needed to do this[36] and also would not have the authority within their institution to enforce compliance, even if they could. For these reasons, deletion requirements should be removed, if possible. If deletion must be included, carve-outs for validation and compliance with research policies, and limitation of the library's responsibility to "reasonable efforts to inform" end users, are essential.

HANDLING TDM DISPUTES

Another consideration when licensing content for computational access is what the responsibilities and processes are when something goes awry. Text and data mining uses, even when clearly laid out in the license, frequently need to be defended and it's best to prepare for that ahead of time. Consider this scenario: you have an electronic subscription for journal content, with a fair use savings clause. A researcher at your institution (without talking to a librarian, or even knowing that they should) starts a data mining project, and writes a script to automatically download any papers at the publisher's URL with certain keywords in the abstract. The publisher has an automated system to identify downloading that doesn't match usual human browsing patterns – usually this consists of rapid downloading of many papers by the same IP address – and the publisher flags this as potential misuse, blocks access for the user, and sends the library an angry email. Even with TDM written into the license, this can be a frequent occurrence: the automated system usually can't tell whether this is misuse or not. Having a process in place for resolving the access disruption and also meeting the users needs is essential. At MIT, the license language we've developed for this is:

> Cure Activities. In the event of any unauthorized use of the Licensed Materials by an Authorized User, Licensee shall respond to Licensor in relation to unauthorized use of the Licensed Materials of which it is made aware and shall use reasonable efforts to remedy such unauthorized use and reduce the likelihood of its recurrence. In the case of unauthorized use which is causing serious and immediate material harm to the Licensor, Licensor may temporarily suspend an individual Authorized User's access to the Licensed Materials (e.g. by blocking an individual user's IP address), provided that Licensor immediately notifies the Licensee of any such suspension, including the reason for the block and any supporting details. Such temporary suspensions will be of the shortest duration possible sufficient to terminate the alleged unauthorized activity and prevent its resumption.[37]

In practice, when a vendor notifies us of a block, we escalate the technical investigation to our campus IT security team, who attempts to identify the user and contact them. Once we know what was going on, we communicate back to the publisher that the issue has been resolved. Separately, then, we will or direct the user to a pre-defined access solution that won't trigger a block, or otherwise reach out to coordinate TDM access for the user. It is important to separate the incident response from the TDM

[36] In general, libraries should not retain these usage logs long term in order to preserve patron privacy. See, e.g., "MIT Libraries Patron Data Privacy Policy," *MIT Libraries*, last updated November 2020, https://libraries.mit.edu/about/policies/privacy-policy.

[37] MIT Libraries, internal documentation, September 22, 2016.

process in order to preserve the privacy of the patron, and to facilitate the TDM interaction, since starting off with an allegation of misuse is usually not a great start.

WEBSCRAPING

What about electronic resources that don't have a negotiated license? Webscraping – collecting information through automated means from the publicly available internet – is also a topic that a librarian advising TDM researchers should have at least a passing familiarity with. The fair use principles discussed above still apply in this context – if anything it is a stronger fair use because the market harm is presumably reduced for freely available content – but additional non-copyright complications can arise.

Many websites have terms of use (ToU) that at least proport to govern use of the content on that website. Because it is the public internet, neither the end user nor the library has negotiated these terms, and they may prohibit TDM or actions – such as scraping – required for TDM. Frequently, website ToUs will not be enforceable as a contract between the user the and site. A contract requires a manifestation of assent from both parties and passively hosted ToU that the user never actually agrees to, or likely even sees, do not satisfy that requirement.[38] In those cases, violation of the ToU will not be actionable as a matter of contract law. Where a ToU becomes enforceable is when the user has actively agreed to the terms in some verifiable way, most usually by clicking "I agree" during a registration process. Whether that agreement is sufficient and binding has been the subject of voluminous litigation,[39] the details of which will be highly relevant if a particular site your researcher is interested in requires an account, but will not usually be an issue for content available on the open web.

As an additional complication, however, there has long been a question of whether violation of ToU, separate from being a potential breach of contract, could be a violation of the Computer Fraud and Abuse Act (CFAA). The CFAA criminalizes accessing a computer system "without authorization or exceed[ing] authorized access."[40] Does violating the posted ToU run afoul of this? Fortunately for TDM researchers, recent cases have narrowed the interpretation of this law. In 2019, the 9th Circuit held that a company scraping data from the LinkedIn website in violation of the ToU was not liable under the CFAA.[41] This narrower interpretation of the CFAA was subsequently endorsed by the Supreme Court[42] which adopted a "gates up or down" approach, which the 9th Circuit subsequently reaffirmed means that "the concept of 'without authorization' does not apply to public websites."[43] Circumventing passwords or authentication would still be a problem under the CFAA, but scraping of open websites no

---

[38] See, e.g., *Meyer v. Uber Techs., Inc.,* 868 F.3d 66, 77–80 (2d Cir. 2017).
[39] For more information see: Eric Goldman, "Online Contracts," in *Internet Law: Cases & Materials*, *Santa Clara Univ. Legal Studies Research Paper* (2022), https://dx.doi.org/10.2139/ssrn.3201352.
[40] 18 U.S.C. § 1030(a)(2) (2018), accessed December 15, 2022, https://www.law.cornell.edu/uscode/text/18/1030.
[41] *hiQ Labs, Inc. v. LinkedIn Corp.*, 938 F.3d 985 (9th Cir. 2019), accessed December 15, 2022, https://casetext.com/case/hiq-labs-inc-v-linkedin-corp-2. See also *Sandvig v. Barr*, 451 F. Supp. 3d 73 (D.D.C. 2020), accessed December 15, 2022, https://casetext.com/case/sandvig-v-barr.
[42] *Van Buren v. United States*, 141 S. Ct. 1648, 210 L. Ed. 2d 26 (2021), accessed December 15, 2022, https://casetext.com/case/van-buren-v-united-states-5/.
[43] *HiQ Labs, Inc. v. LinkedIn Corp.*, 31 F.4th 1180, 1199 (9th Cir. 2022), accessed December 15, 2022, https://casetext.com/case/hiq-labs-inc-v-linkedin-corp-5.

longer faces that issue. Webscraping remains, however, a complicated and evolving legal issue, which researchers and the librarians that support them should keep an eye on.[44]

PRACTICAL APPLICATION AND BEST PRACTICES

For practical purposes, it's helpful to consider a series of questions when advising researchers on TDM projects:

WHAT IS THE SOURCE OF THE CORPUS?

- If the corpus is sourced through digitization of print resources, then fair use is going to be the guiding principle and you should look to the fair use section of this chapter to help design a project plan that stays within a fair use assessment informed by the researcher's and institution's risk tolerance.
- If the corpus includes licensed eresources, then you should check the terms of the relevant licenses, with an eye to anything that might limit effective TDM access, and potentially negotiate TDM-specific terms if they are not already present.
- If advising on data obtained from miscellaneous sources, you should additionally keep in mind the webscraping issues discussed above.

WHAT DOES THE ANALYSIS REQUIRE?

Generally, this part is where the researcher spends most of their time, and it is actually the most straightforward, legally. Fair use and the standard use provisions of most academic licenses will almost always cover the actual research activities, once the corpus has been legitimately acquired.

PUBLISHING AND SHARING THE DATASET

Once the analysis is done, the researcher also needs to share the result. It's the librarian's role to minimize barriers to the subsequent publication of TDM datasets and output resulting from use of the corpus. You can do this by advising researchers on construction of datasets to minimize copyright concerns, and comply with licensing requirements. You can also anticipate and head off this need by preemptively addressing likely issues in eresource licenses at the point of negotiation.

A final bit of advice, if you are navigating text and data mining rights for a library resource, is to work closely with a TDM researcher, if you have the opportunity. Researchers using these tools know what they need, and can help determine whether particular terms are reasonable. Our TDM support at MIT has been greatly improved through direct collaboration with the researchers we serve.

---

[44] For example, as of the date of this chapter, a petition for certiorari is pending with the U.S. Supreme Court for a case on the enforceability of state-law contract remedies in a webscraping context. See "ML Genius Holdings LLC v. Google LLC," *SCOTUSblog,* accessed December 15, 2022, https://www.scotusblog.com/case-files/cases/ml-genius-holdings-llc-v-google-llc/. See also, Keiran McCarthy, "Hello, You've Been Referred Here Because You're Wrong About Web Scraping Laws (Guest Blog Post, Part 2 of 2)," *Technology and Marketing Law Blog,* December 9, 2022, https://blog.ericgoldman.org/archives/2022/12/hello-youve-been-referred-here-because-youre-wrong-about-web-scraping-laws-guest-blog-post-part-2-of-2.htm.

ADDITIONAL RESOURCES

Building LLTDM (Legal Literacies for Text and Data Mining): https://buildinglltdm.org/; https://berkeley.pressbooks.pub/buildinglltdm/.

FAIR for Data and Texts not in the Open: https://osf.io/tbkj7/

Resources and Tools for Computational Research: https://libguides.mit.edu/comptools