Kyle K. Courtney, Rachael Samberg, and Timothy Vollmer

Big data gets big help

Law and policy literacies for text data mining

A wealth of digital texts and the proliferation of automated research methodologies enable researchers to analyze large sets of data at a speed that would be impossible to achieve through manual review. When researchers use these automated techniques and methods for identifying, extracting, and analyzing patterns, trends, and relationships across large volumes of un- or thinly structured digital content, they are applying a methodology called *text data mining* or TDM.[1] TDM is also referred to, with slightly different emphases, as "computational text analysis" or "content mining."

The "distant reading" that TDM makes possible supports the discovery of scientific and social insights, such as how gender is depicted in fiction over time or evidence of racial disparity in police camera footage.[2] Libraries are eager to provide and expand institutional access to data sets so that scholars can continue exploring unknown connections, yet both scholars and professional staff who support TDM research often run into roadblocks.[3] Law and policy questions are paramount and shape not only how TDM scholarship is disseminated, but also the very questions being asked. If researchers are limited to corpora unencumbered by legal restrictions, they risk perpetuating bias in the scholarly record.[4]

With a basic set of law and policy literacies in hand, libraries can help scholars navigate these issues so that they can confidently use, create, and share a far wider set of corpora and research results.[5]

Copyright

Imagine that a researcher in the United States wants to analyze a corpus of 20th-century literature to examine the stylistic commonalities among literary prize winners. The researcher digitizes or downloads dozens of prize-winning texts and runs their computational software on the materials. They discover some interesting details and decide to publish their findings along with parts of their corpus for data validation. But the researcher begins to have doubts about the activity. Since the literature is protected by copyright, was it lawful to digitize or download and run the analyses on the corpus in the first place? Can the researcher share parts of the literary works for purposes of reproducibility, or so that other scholars can query the corpus for other research questions? The answers lie in what TDM researchers and librarians need to understand about copyright law.

By providing rewards for authorship in the form of exclusive rights, copyright law incentivizes the creation and dissemination of knowledge. But copyright law is also in-

Kyle K. Courtney is copyright advisor and program manager at Harvard Library, email: kyle_courtney@harvard.edu, Rachael Samberg is scholarly communication officer and program director at the University of California-Berkeley Library, email: schol-comm@berkeley.edu, and Timothy Vollmer is scholarly communication and copyright librarian at the University of California-Berkeley Library, email: schol-comm@berkeley.edu

tended to benefit the public, and if authors held exclusive rights to their works indefinitely, public access to knowledge would be impeded.

Congress has actively limited these rewards in important ways. Much of the public benefit of copyright is incorporated by the *expiration* of rights. When copyright ends, works enter the "public domain."[6] The copyright term has been lengthened significantly,[7] but even for works still protected by copyright, Congress built critical exceptions into the Copyright Act to promote the progress of science and art. One of the strongest such exceptions is the right of fair use, codified in 17 U.S.C. § 107, which states that "notwithstanding" the bundle of rights granted to the copyright owner, the fair use of a copyrighted work . . . is not an infringement."[8]

Courts consider four factors in making a fair use determination: 1) the purpose and character of the use (nonprofit uses and uses that "transform" a work by adding new insights or understanding are more likely to be fair), 2) the nature of the copyrighted work (use of factual works is more likely to be fair than works coming closer to the "core of creative expression"), 3) the amount and substantiality of the portion used (amounts appropriate to the new transformative purpose are more likely to be fair), and 4) the effect of the use upon the potential market for or value of the copyrighted work (uses that do not usurp the market for the original are more likely to be fair). Evaluating whether a given use of copyrighted material is "fair" overall requires balancing these four factors on a case-by-case basis.

Courts that have considered computational research have found TDM to be a fair use. For instance, in *Authors Guild v. HathiTrust,* 755 F.3d 87 (2d Cir. 2014), scanning and creating a database of digitized materials so that users could conduct full-text searching within content, rather than read that content, was highly transformative under factor one and a fair use overall.

In that case, a collection of authors and author associations had sued HathiTrust and certain of its member universities for copyright infringement. The basis of their claims was the fact that, pursuant to a relationship with Google, HathiTrust received digital copies of nearly ten million books—the majority of which were still in copyright. HathiTrust then made these books available for full-text searching, without the researcher being able to read the book.

The court found this arrangement to be fair use, notably because the textual analysis that the HathiTrust Digital Library enabled was transformative under the first fair use factor: "[T]he result of a word search is different in purpose, character, expression, meaning, and message from the page (and the book) from which it is drawn." In *Authors Guild v. Google, Inc.,* 804 F.3d 202 (2d Cir. 2015), the same court found that Google Books' creation of a full-text searchable database and "Ngram Viewer" were fair—as was allowing users to view three-line snippets of the underlying works to provide context for where desired phrases appear.[9]

There is less clarity around how, or if, a researcher may share the underlying corpus in order to enable verification of research results or offer new querying opportunities. For instance in *Fox News Network, LLC v. TVEyes, Inc.,* No. 15-3885 (2nd Cir. Feb. 27, 2018), media aggregator TVEyes was recording commercial news and radio audiovisual content, importing it into a database, and permitting its clients to search for, view, download, and share that content in ten-minute clips.

While keyword-enabled searching would be both transformative and a fair use overall, permitting redistribution was not because it made "available to TVEyes's clients virtually all of Fox's copyrighted content that the clients wish[ed] to see and hear, and because it deprive[d] Fox of revenue." Therefore, the key copyright issue scholars will face is typically *how much* of the corpus they used or created can be shared or republished.

Contracts

Researchers and librarians also need to understand circumstances in which the contracts they have signed or to which they have assented can control—and even supersede—TDM uses that would otherwise have been permitted under copyright law. Even if the act of downloading and sharing copyright-protected materials when conducting TDM may have constituted fair use, some license agreements expressly forbid it.

Shrewd TDM researchers may try their luck compiling a dataset from the "open web" instead, but often encounter confusing hurdles with application programming interface or website terms of service governing how researchers may access, use, and share the content. Some courts find that these website "terms of service" or "terms of use" can constitute an enforceable "browsewrap" agreement—one to which a party assents simply by using the website, others dismiss browsewraps entirely.[10] Some courts require browsewrap terms to have certain visual characteristics and cues to be enforceable, or seek proof that a user was actually aware of them.[11] This complex landscape makes it confusing for researchers to understand how to proceed, but ignoring the browsewrap is not advisable either—particularly as doing so may also violate a university or library's Internet policies.

Libraries and researchers can negotiate to retain both fair use rights and the right to conduct TDM, expressly.[12] In some cases, vendors may charge a hefty (and prohibitive) fee in their licenses to preserve these benefits for researchers. The ability to negotiate favorable license agreements also varies from publisher to publisher, leaving the prospective TDM researcher with a patchwork of differing rules they must follow for each content source they wish to include in their corpus. At other times, vendors might require researchers to ask permission to conduct TDM on a case-by-case basis, which may involve additional obstacles.

Privacy and ethics

Researchers engaging in digital scholarship that incorporates materials stewarded by libraries and archives typically are no strangers to issues of privacy and ethics. Library special collections of personal writings and correspondence, photographs, and audio-visual recordings often contain information protectable under federal statutes (such as financial, medical, or student record data) or state privacy laws (which prohibit actions like disclosure of facts that would not otherwise be made public, or intrusion in places where people have a reasonable expectation of privacy). Researchers working with such materials face these questions regardless of their research methodologies, but TDM transforms these challenges into ones of greater scale and impact. TDM enables the potential review and disclosure of much greater volumes of data, exacerbating the risk that scholars may run afoul of privacy protections and increasing the need for careful data management practices.

Sometimes questions that seem like ones of privacy are ethical issues. For instance imagine a TDM scholar wanted to explore "Gamergate"—the harassment of women who spoke out on Twitter and other sites against misogyny within video game development culture.[13]

As Todd Suomela et al. note, women who shared their views received rape and death threats, but often there was nothing "private" (from a legal perspective) in these messages of hate.[14] Yet the mere act of compiling and publishing a corpus containing instances of harassment could amplify the messages or make the published information more readily discoverable, thus exposing the women who had spoken out to additional threats. The researchers needed to consider what ethical standards should be applied to mining and publishing data in these contexts.

Building legal literacies

Researchers may also face specialized questions of cross-boundary collaborations complicated by the inconsistent framework of international copyright and privacy laws. Or perhaps the researchers need to "break" digital rights management protections to access the content they want to mine. How can scholars and librarians acquire an understanding of all relevant concerns?

In a review of digital humanities and information science curricula, professional development training programs, and library guides, we observed few training opportunities or resources that integrate legal literacies into TDM outreach and instruction, particularly in the context of digital humanities. We viewed this as an opportunity for our team of librarians, legal experts, and scholars to build and offer a robust curriculum at a four-day institute at the University of California-Berkeley in June 2020.

"Building Legal Literacies for Text Data Mining," supported by the National Endowment for the Humanities, will bring together digital humanities researchers and professionals to share and learn together.[15] The project team will publish the curriculum as an open educational resource to foster a broader community of practice. The goal is for TDM researchers and professionals to confidently build, mine, and publish corpora with a solid understanding of legal, ethics, and risk choices they will make along the way.

Notes

1. Marti Hearst, "What Is Text Mining?" SIMS, UC-Berkeley, October 17, 2003, http://people.ischool.berkeley.edu/~hearst/text-mining.html.

2. Ted Underwood, David Bamman, and Sabrina Lee, "The Transformation of Gender in English-Language Fiction," *Journal of Cultural Analytics,* 2018, https://doi.org/10.22148/16.019; R. Voigt, et al., "Language from Police Body Camera Footage Shows Racial Disparities in Officer Respect," Proceedings of the National Academy of Sciences 114, no. 25 (May 2017): 6521–26, https://doi.org/10.1073/pnas.1702413114.

3. Matthew Sag, "The New Legal Landscape for Text Mining and Machine Learning," *SSRN Electronic Journal,* 2019, https://doi.org/10.2139/ssrn.3331606.

4. Megan Senseney, Eleanor Dickson, Beth Namachchivaya, and Bertram Ludäscher, "Data Mining Research with In-Copyright and Use-Limited Text Datasets: Preliminary Findings from a Systematic Literature Review and Stakeholder Interviews," *International Journal of Digital Curation* 13, no. 1 (2018): 183–94, https://doi.org/10.2218/ijdc.v13i1.620; Amanda Levendowski, "How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem," *UW Law Digital Commons,* accessed February 11, 2020, https://digitalcommons.law.uw.edu/wlr/vol93/iss2/2/.

5. 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.

6. *Sony Corp. of Am. v. Universal City Studios, Inc.,* 464 U.S. 417, 429 (1984).

7. *Eldred v. Ashcroft,* 537 U.S. 186 (2003).

8. 7 U.S.C. § 107 (2020).

9. See also A.V. ex rel *Vanderhye v. iParadigms,* 562 F.3d 630 (4th Cir. 2009).

10. See *Zaltz v. JDATE,* 952 F.Supp.2d 439 (E.D.N.Y.2013).

11. See *Specht v. Netscape Commc'ns Corp.,* 306 F.3d 17 (2d Cir. 2002).

12. See, e.g., the California Digital Library Model License, available at https://cdlib.org/cdlinfo/2017/01/25/cdl-model-license-revised/.

13. Todd Suomela, Florence Chee, Bettina Berendt, and Geoffrey Rockwell, "Applying an Ethics of Care to Internet Research: Gamergate and Digital Humanities," *Digital Studies/Le Champ Numérique* 9, no. 1 (2019). https://doi.org/10.16995/dscn.302.

14. Ibid.

15. "The Institute," Building LLTDM, accessed February 11, 2020, https://buildinglltdm.org/about/institute-basics/.