Finding the Haystack: Literacies for Accessing and Using Text as Data

Cody Hennesy (University of Minnesota, Twin Cities), Stacy Reardon (University of California, Berkeley), and Rachael Samberg (University of California, Berkeley)

Legal Literacies for Text Data Mining

Copyright

Can I use copyright-protected materials in my TDM project?

- **What:** Copyright law gives exclusive rights to authors of expressive works, but these rights are subject to important limitations such as fair use exceptions.
- **Implications:** TDM in which a computer statistically analyzes text (as opposed to a human simply reading the text) can be considered a transformative use of text, one of the key factors in determining fair use exemptions.
- **Takeaways:** You don't need to limit your TDM project to public domain materials (materials for which copyright has expired). Using copyright-protected materials for TDM has been found a be a fair use.
- **Corpus sharing:** Publishing or sharing your corpus after analysis is less likely to be considered fair use. Consider sharing derived data instead of your full corpus.

Contracts

Can I scrape this database or website?

- **What:** Contract law can supersede copyright, meaning that agreements you may enter into when you use materials from an archive, library database, website, or other provider can restrict otherwise fair uses.
- **Implications:** Even when TDM may be a fair use, programmatically downloading content may violate a provider's Terms of Use or license agreement.
- **Takeaways:** Before scraping a website or database, review the Terms of Use. If there is an API, use that. If the Terms prohibit your intended uses you may wish to ask permission, negotiate with the content provider, find an alternative source, or perform a risk assessment.
- **Corpus sharing**: Licenses may include separate provisions regarding how and how much of the copyright-protected data you can share.

Privacy

What should I do if my corpus contains private data?

- **What:** Privacy rights protect people who are the subjects of a work. Most privacy rights expire at death. Newsworthiness of the content or getting permission are also defenses for publishing certain content otherwise considered private.
- **Implications:** If your corpus contains private data this may affect how you store and share it.
- **Takeaways:** Protect sensitive data by password and encryption. Store cloud-based data during the phase of active research only on secured servers provided by your campus or organization. When working with private data, work in an environment without access to web browsing, Skype, and other potentially vulnerable tools.
- **Corpus sharing:** If you decide to share your corpus strip out private information that is protected by law.

Ethics & Policy

What ethical issues should I consider in compiling/sharing text data?

- **What:** Questions related to indigenous knowledge, at-risk communities, cultural heritage sites, and endangered species may be subject to de-identifying regulations or otherwise call for ethical research choices.
- **Implications:** The scale of data amassed and sometimes exposed for TDM research can have broader impacts on the safety or exploitation of subjects being studied.
- **Takeaways:** Consider your own ethical principles, guidelines in your discipline, best practices statements, and the perspectives of the individuals and communities under study.

Other Statutes/ Use Cases

What other risks are there in downloading and sharing corpus data?

- **What:** The *Computer Fraud and Abuse Act*, the *Digital Millennium Copyright Act (DMCA)*, and international law may all be applicable to your project.
- **Implications:** Sidestepping measures that protect materials online, such as breaking digital rights management (DRM) or hacking into a library's authentication system, can be a crime, not just a violation of contract. US copyright law will generally apply if you are doing research in the US, but you may also acquire content subject to foreign license agreements.
- **Takeaways:** Don't hack into systems or break DRM to access materials. When materials are from international providers, closely examine agreements, and consider whether other ethics or policy issues might relate to your corpus.

TDM Opportunities

Books & primary sources
Gale Digital Scholar Lab
HathiTrust Research Center (HTRC)

News
Chronicling America (LoC, 1836-1923)
MediaCloud
NewsAPI.org

Scholarship & citations
Elsevier APIs
JSTOR Data for Research

Social media
Social Media Macroscope
Stanford Large Network Datasets (SNAP)

Public & government data
ProPublica Data Store
Court Listener APIs and Bulk Data

More resources are available from:
guides.lib.berkeley.edu/text-mining
libguides.umn.edu/text-mining

New Modes of Access

Ngrams

What: Ngrams are lists of word counts from texts, out of order, that enable certain text analysis methods (e.g., topic models, classifiers...) while not revealing the underlying human-readable text.
Takeaways: Share your own text-as-data as ngrams to avoid violating copyright or contracts.
Examples: HTRC Extracted Features, JSTOR Data for Research, Google Ngrams

Secured computing environments

What: A secured computing environment is a virtual space where you can access and work with materials, instead of downloading content that is under copyright.
Takeaways: You may not be able to export your corpus, or otherwise share data outside of the secured environment.
Examples: HTRC Data Capsules, Gale Digital Scholar Lab

APIs

What: Application Programming Interfaces provide formal access hooks to query and download data from different publishers.
Takeaways: Before web scraping, check first for an API which provides a legal and transparent way to download data following limits set by the publisher.
Examples: Twitter APIs, Elsevier APIs, ProPublica Congress API, NewsAPI.org

References

Sag, Matthew. The New Legal Landscape for Text Mining and Machine Learning, *Journal of the Copyright Society of the USA*, Vol 66 (2019). http://dx.doi.org/10.2139/ssrn.3331606

Samberg, Rachael, & Hennesy, Cody. (2019, forthcoming). Law & Literacy for 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, IL: Association of College & Research Libraries.

Astro-chart illustrations are from Alesha Sivartha's *The Book of Life* (1898), via www.oldbookillustrations.com.

DASH Digital Arts, Sciences, & Humanities University of Minnesota

Libraries University of Minnesota

Berkeley Library University of California