

# **Text Data Mining, Copyright, & Licensing/Contractual Restrictions**



**Rachael Samberg**  
**Leslie Barnes**  
**Stephanie Orfano**

Colloquium on Text & Data Mining in Libraries  
2023

# **Copyright Primer**

## **US & Canadian Law**



# What is COPYRIGHT?

Exclusive rights

to make certain uses

of original expression

for limited period of time

RQ: <i>wh</i> in Braun et al. (2019)	ISQ: polar in Braun et al. (2019)
<i>Deine Freundin serviert bei einem Fest Garnelen als Vorspeise. Doch es ist offensichtlich, dass sich all eure Freunde vor dem gummiartigen Zeug eckeln. Du fragst deine Freundin:</i>	<i>Auf einer Dinnerparty servierst du Garnelen. Du möchtest wissen, ob deine Freunde das essen und davon möchten oder nicht. Du fragst deine Freunde:</i>
<i>wer mögen Garnelen</i>	<i>mögen Garnelen</i>
At a party, your friend is serving shrimp as an appetizer. But it is obvious that all of your friends are disgusted by the rubbery stuff. You ask your friend:	At a dinner party you serve shrimps. You would like to know which of your friends like this and whether they want some of it or not.
who like shrimps	like shrimps

## **U.S. Exclusive Rights**

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

## **Canadian Exclusive Rights**

- Creation & recreation
- Publish (if unpublished)
- Public performance
- Translation
- Derivative works
- Public exhibition
- Computer program and musical work rentals

# Limited Period

- Varies, but at least author's life + 70 years (canadian change from 50 years in 2022)
- Within “protected” period, you need author's permission to reproduce, display, perform, etc.

RQ: <i>wh</i> in Braun et al. (2019)	ISQ: polar in Braun et al. (2019)
<i>Deine Freundin serviert bei einem Fest Garnelen als Vorspeise. Doch es ist offensichtlich, dass sich all eure Freunde vor dem gummiartigen Zeug ekeln. Du fragst deine Freundin:</i>	<i>Auf einer Dinnerparty servierst du Garnelen. Du möchtest wissen, ob deine Freunde das essen und davon möchten oder nicht. Du fragst deine Freunde:</i>
<i>wer mögen Garnelen</i>	<i>mögen Garnelen</i>
At a party, your friend is serving shrimp as an appetizer. But it is obvious that all of your friends are disgusted by the rubbery stuff. You ask your friend:	At a dinner party you serve shrimps. You would like to know which of your friends like this and whether they want some of it or not.
who like shrimps	like shrimps

---

# Get permission?

## Yes, unless...

- Facts/ideas
- Public domain
- **Exception like:**
  - U.S. fair use**
  - Canada fair dealing**
  - Canada temp reproductions for tech process**

---

# Do the exceptions apply to TDM?

- U.S. fair use: **Yes, to conduct TDM. May hit limits on republishing**
- Canadian fair dealing or temporary reproduction for tech processes: **Uncertain**

---

# Breaking DRM to do TDM?

- U.S.: **Yes with restrictions**
- Canada: **No**




# Contracts



**Even if a use is fair / fair dealing or if the content is not protected by copyright, there may be a contract that restricts scraping, TDM, and/or breaking DRM to do TDM.**

# Website Terms

conservation :: educator programs :: press :: shop  
contact us :: site map :: terms of use :: web privacy



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

# LinkedIn User Agreement

## You agree that you will not:

1. Develop, support or use software, devices, scripts, robots or any other means or processes (including crawlers, browser plugins and add-ons or any other technology) to **scrape the Services or otherwise copy profiles and other data** from the Services;
2. **Override any security feature or bypass or circumvent any access controls** or use limits of the Service (such as caps on keyword searches or profile views);
3. **Copy, use, disclose or distribute any information obtained from the Services**, whether directly or through third parties (such as search engines), without the consent of LinkedIn;

# “What are we doing to stop scraping?”

On LinkedIn, our members trust us with their information, which is why **we prohibit unauthorized scraping on our platform.**

Our teams at LinkedIn [also] **create, deploy, and maintain models and rules that detect and prevent abuse, including preventing unauthorized scraping.**

- To detect public profile scraping, our models look for signs of automated viewing....
- We...defend against logged-in scraping...we look for signals of bot-like activity
- In addition to rate limits, we also employ a funnel of additional defenses that detect and take down fake accounts engaged in scraping at multiple stages.

# Fair Use Expressly Preserved

## Fair use is permitted

Fair use of copyrighted material includes the use of protected materials for noncommercial educational purposes, such as teaching, scholarship, research, criticism, commentary, and news reporting. Except with respect to content included as part of MoMA's Online Virtual Cinema Screenings or unless otherwise noted, users who wish to download or print text, audio, video, image and other files from MoMA's website for such uses are welcome to do so without MoMA's express permission. In accordance with scholarly practice, users of

# TDM Expressly Permitted

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

# Contract prohibiting DRM circumvention

## Amazon Kindle:

Limitations. You may not remove or modify any proprietary notices or labels on the Kindle Content. In addition, you may not **attempt to bypass, modify, defeat, or otherwise circumvent any digital rights management system or other content protection or features used as part of the Service.**

# Procedural / Tech Restrictions

Members and Authorized Users should notify the Licensors via [ejournals@rsc.org](mailto:ejournals@rsc.org) at least two (2) weeks before they wish to carry out the Text and Data Mining and give the Licensors the following information:

- Date to start:
- Completion date:
- Institution:
- Crawler IP address:
- Crawler user agent: TDMCrawler (please set user agent to this) Types of content (HTML / PDF)
- Institution contact email: Researcher contact email:

The Text and Data Mining should be carried out in the following manner:

- Keeping delays to 10-20 seconds between requests.
- Set the user agent to TDMCrawler, adding contact and project information.



# Procedural / Tech Restrictions

“The Customer and its Authorized Users must access Wiley’s content for Text and Data Mining using an approved API service such as Crossref’s TDM service or a Wiley’s API and must abide by any rate-limiting conveyed in machine readable form from time to time, and may not bypass the API or disrupt the working of Wiley’s server” (License agreement)

[Website](#) further specifies:

- Require API token (same as API key)
- Rate-limit: 3 requests a second
- No creations of local repositories/ parallel services
- No placing results on server *unless* to share with another authorized user
- Must find DOI first (using Crossref)

# RISK EVALUATION



# U.S. TDM RISKS

- **Copyright:** Doing TDM is fair use, but republishing content may not be. Breaking DRM subject to procedural limitations. Violations: Infringement carries potential statutory fines of \$30K/work, or up to \$150K if infringement willful
- **Contract:** Actual damages, but what are they? Possibly lost licensing fees. Also, institutional license suspension or termination. Note that terms of service online have some potential for not being legally binding.

# Canadian TDM RISKS

- **Copyright:** Doing TDM not necessarily Fair Dealing. Statutory damages for non-commercial infringement is \$100-\$5K per work. Further, Canadian anti-circumvention laws very strict, even for personal non-commercial purposes.
- **Contract:** Terms of service agreements create legally binding contracts and the remedies set out therein are enforceable. Institutional licenses may be suspended or canceled. Some limitation (i.e. injunction only) on institutional liability if digital locks violated if institution was not aware

# “Non-quantifiable” RISKS

- **Retraction** if publisher discovers unauthorized use. (See, e.g. <https://retractionwatch.com/2021/07/30/a-very-unfortunate-event-paper-on-covid-19-vaccine-hesitancy-retracted/>)
- **Stymying progress of knowledge** by refraining from research



# EXERCISE

A top-down photograph of exercise equipment on a rustic wooden surface. In the upper left, two black dumbbells are positioned diagonally. In the center, an orange resistance band is coiled. To the right of the band, a black suspension trainer with two handles and a central carabiner is laid out. The wood grain is prominent and runs horizontally across the frame.

[bit.ly/3L9k8MP](https://bit.ly/3L9k8MP)

# Springer-Nature Journals (Canada)

## **TDM authorized with certain limitations; sharing limited**

- i. Materials can be stored only for duration of the project
- ii. Materials can be shared only with authorized users (and Amit's American colleague likely not an authorized user)
- iii. Security measures are required
- iv. Destruction of copies at termination of agreement

# ProQuest eBooks (U.S.)

## Probably no TDM at all

Unlimited Access. Subject to the terms of this Agreement, Licensee and its Authorized Users shall have unlimited access to the Licensed Materials. Notwithstanding the above and in order to protect the Licensed Materials for the research and educational use of Authorized Users, automated searches against ProQuest's systems are not permitted with the exception of nonburdensome federated search services.

**Data mining is prohibited.**



# ProQuest eBooks (U.S.)

## Also restrictive for TDM

Digitally Copy. Licensee and Authorized Users may **download and digitally copy a reasonable portion of the Licensed Materials** so long as each work is retrieved directly from the on-line database system in a manner that causes a “hit” to be registered on the on-line system for each and every print or digital copy. All reproduction and distribution of such printouts, and all downloading and electronic storage of materials retrieved through the Products **shall be for your own internal or personal use as allowed under the doctrines of “fair use” and “fair dealing”**. **Downloading of all or parts of a Product in a systematic or regular manner or so as to create a collection of materials comprising all or a material subset of a Product is strictly prohibited** whether such collection is in electronic or print form.

# ProQuest eBooks (U.S.)

**Plus, probably no sharing underlying materials with Canadian researcher**

Scholarly Sharing. Authorized Users may transmit to a third party in hard copy or electronically, **minimal, insubstantial amounts of the Licensed Materials ...provided that in no case any such sharing is done in a manner or magnitude as to act as a replacement for the recipient's or recipient institution's own subscription**

# ProQuest eBooks (U.S.)

**If TDM had been allowed, nothing expressly precluding breaking DRM / TPM**

In the event that Licensor develops any significant type of digital rights management technology to limit the access or the usage of Licensed Product, Licensor agrees to notify Licensee with a contact name at ProQuest for support as well as provide technical specifications, to the extent such exists, for the digital rights management technology utilized.