Building Legal Literacies for Text Data Mining

Alyssa Fahringer
Digital Scholarship Consultant
Digital Scholarship Center

Copyright

Copyright is a bundle of rights; a limited economic monopoly for authors; and a system "to promote the progress of science and the useful arts."

- Original/creative and fixed
- No registration required
- Grant of rights to owners (the "bundle")
- Wide range of protected works
- Long term of protection
- Exemptions and limitations

Copyright key takeaways

- Copyright law is part of the legal system that covers creation and use
- Copyright is meant to strike a balance between the rights of the copyright holders and users of those of copyrighted works, such as TDM researchers
- The cycle of creation, dissemination, and expiration of rights into the public domain is a critical component of copyright law. Without this balance the system loses its value, or prevents the public from receiving the benefits of the bargain

Public domain

Public domain is the commons of material not protected by copyright. This includes

- Works for which copyright has expired
- Works for which copyright owners failed to comply with "formalities"
- Works that are not copyrightable

Public domain key takeaways

- The public domain has nothing to do with what is readily available for public consumption. Just because something is on the internet does not mean it is in the public domain!
- The public domain is important to the production of creativity. Authors need these essential building blocks with which to work
- If a work you are using is in the public domain, copyright issues do not apply

Fair use

Fair use gives the user a right to exercise one of the exclusive rights of copyright without obtaining the permission of the copyright owner and without the payment of a license fee.

Four fair use factors

1. The purpose and character of the use
2. The nature of the copyrighted work
3. The amount and sustainability of the portion used in relation to the copyrighted work as a whole
4. The effect of the use on the potential market for or value of the copyrighted work

Development of transformative fair use

Fair use key takeaways

- Fair use is for everyone and is useful to the TDM researcher, because TDM involves accessing, copying, and processing works that may be in copyright
- If there were no fair use, and copyright holders could forbid you from using copyrighted worked without permission, this would vastly stifle free expression and scholarship
- Transformative fair use is the life and breath of scholarship, research, and teaching
- Fair use law is adaptable to various scenarios; the purpose of fair use is flexibility

Fair use and TDM

Three core uses that are likely to occur in most TDM research projects

- Copying to create a database for TDM
- Using derived data
- Publishing data sets

TDM researchers have a long line of cases that provide fairly clear support for the kinds of things they do with in-copyright material. Core TDM research methods are well-suited for fair use.

Licenses

A license is a "contract not to sue." A license or a contract is a legal interest created by a titleholder granting use privileges to some non-titleholder. Licenses can determine what a TDM researcher can do within legal bounds.

Sections in licenses where TDM related clauses might reside

- Authorized uses or permitted uses
- Definitions
- Any sections listed as "intellectual property" or "copyright"

Library e-resource licenses / Model licenses

Library e-resource licenses

- Data mining
  - *Subject to any content-specific restrictions, Customer and its Authorized Users may extract and compile data from locally-loaded copies of the Purchased Content for Customer's teaching, learning, and research purposes*
- Restrictions
  - *Except as expressly permitted above, Customer and its Authorized Users shall not: ... text mine, data mine or harvest metadata from the service*

Model licenses

- Text and data mining
  - *Authorized Users may use the Licensed Materials to perform and engage in text and/or data mining activities for academic research*
- Escape clause
  - *Licensee and Authorized Users may make all use of the Licensed Materials as is consistent with United States copyright law, including its Fair Use Provisions*

Beyond the terms of the license

Other than the terms of the license, ask

- Am I bound?
- How does this relate to fair use?
- What happens if I breach?
- Trespass to chattels?
- Risk management

Licensing key takeaways

- Check the [Text and Data Mining Sources infoguide](#)
- Ask us <datahelp@gmu.edu>
- [GMU responsible use of computing policy](#)

Privacy law

Prosser torts

1. Intrusion upon seclusion
2. Public disclosure of embarrassing facts
3. Painting someone in a false light
4. Appropriation of name or likeness

Privacy law key takeaways

- Work on your TDM project can proceed if the following is true
    - If the subject matter of the collections is no longer living, or
    - If the subject matter is newsworthy or of public interest, or
    - If the subject matter is not identifiable, or
    - If the subject matter has released the information themselves--either by giving you permission or posting on social media

Ethics

Questions to ask when working with data

1. What should we do with data that is not private but might be sensitive?
2. When should you impose an ethical framework, and how do you balance that framework with truth-seeking, the public interest, and free expression?

Obtaining consent

- Consent-based framework
- Harms-based framework

Frameworks

Normative

- Deontologist
- Virtue ethicist
- Utilitarian

Alternative

- Ethics of care
- Risk-benefit analysis

Harm

- Psychological
- Physical
- Legal
- Social

Strategies

Strategies to account for ethics

- Consult journal publication or professional association guidelines
- Develop local best practices
- Impose access controls
- Undertake community engagement
- Seek IRB involvement/approval
- Adopt a new ethics/privacy paradigm

Ethics key takeaways

- Ethics is not a one-time consideration. As Dixon and Quirke said, "Ethical issues permeate and unfold beyond the research design stage and throughout the entire research process"
- When doing TDM research, consider different ethical frameworks and the different types of harm, and how best to minimize or weigh that harm
- Start with implementing strategies within your own research, then move outwards

Resources

- University of California Berkeley Library's [Responsible Access Workflows](#)
- Building Legal Literacies Text Data Mining
    - [Institute homepage](#)
    - [OER forthcoming](#)
- Email me <afahring@gmu.edu>