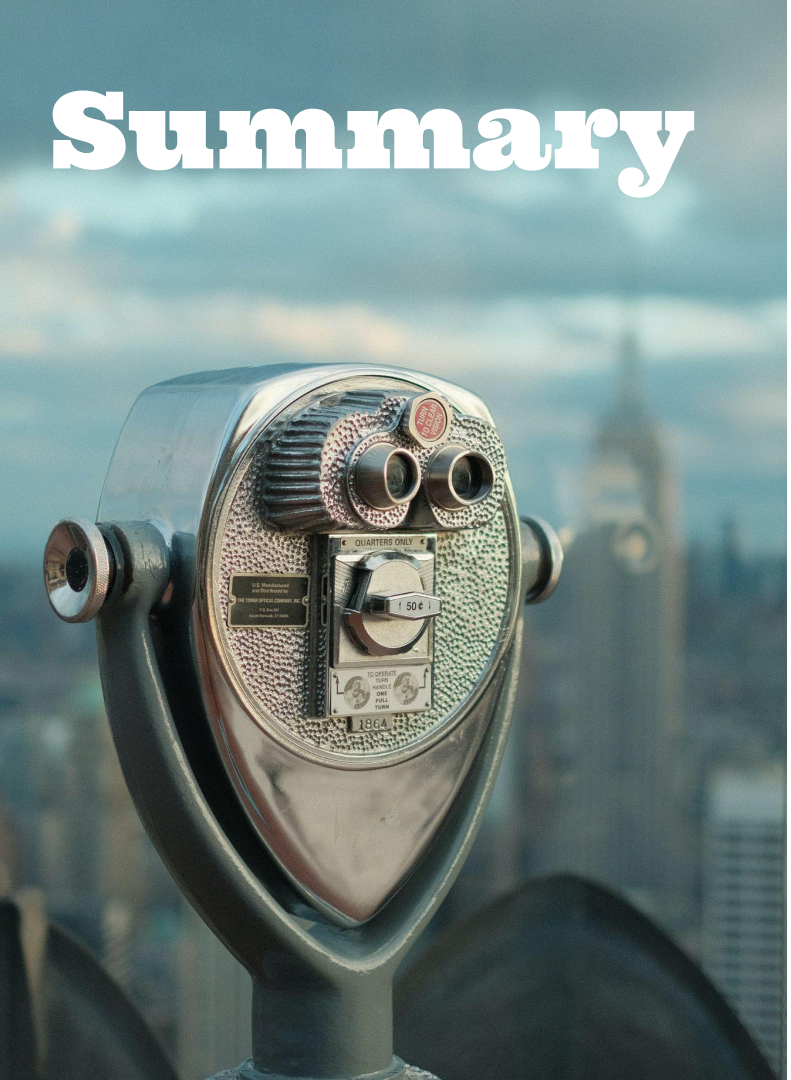
Legal Literacies for Text Data Mining: Cross-Border

NATIONAL ENDOWMENT FOR THE HUMANITIES

Berkeley Library
UNIVERSITY OF CALIFORNIA

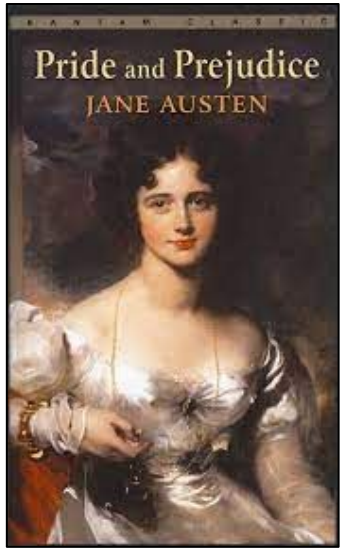
INTERNET ARCHIVE

Summary



- **Origin** - why study these issues?
- **Approach** - writing, virtual round tables, and analysis with researchers and experts
- **Tools and Lessons Learned** - white paper and case study to guide researchers and libraries
- 💫 **Futures** 💫

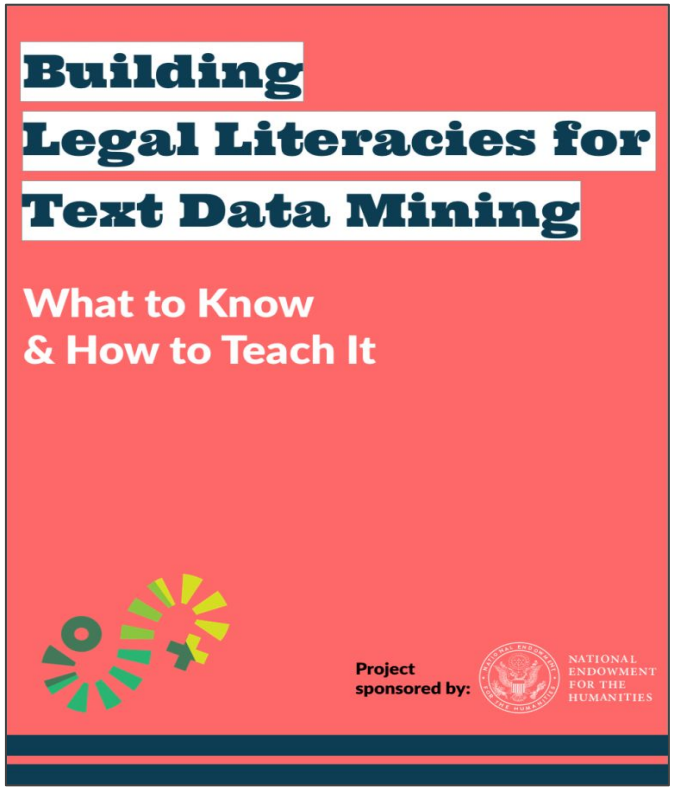
What is text data mining (TDM)?

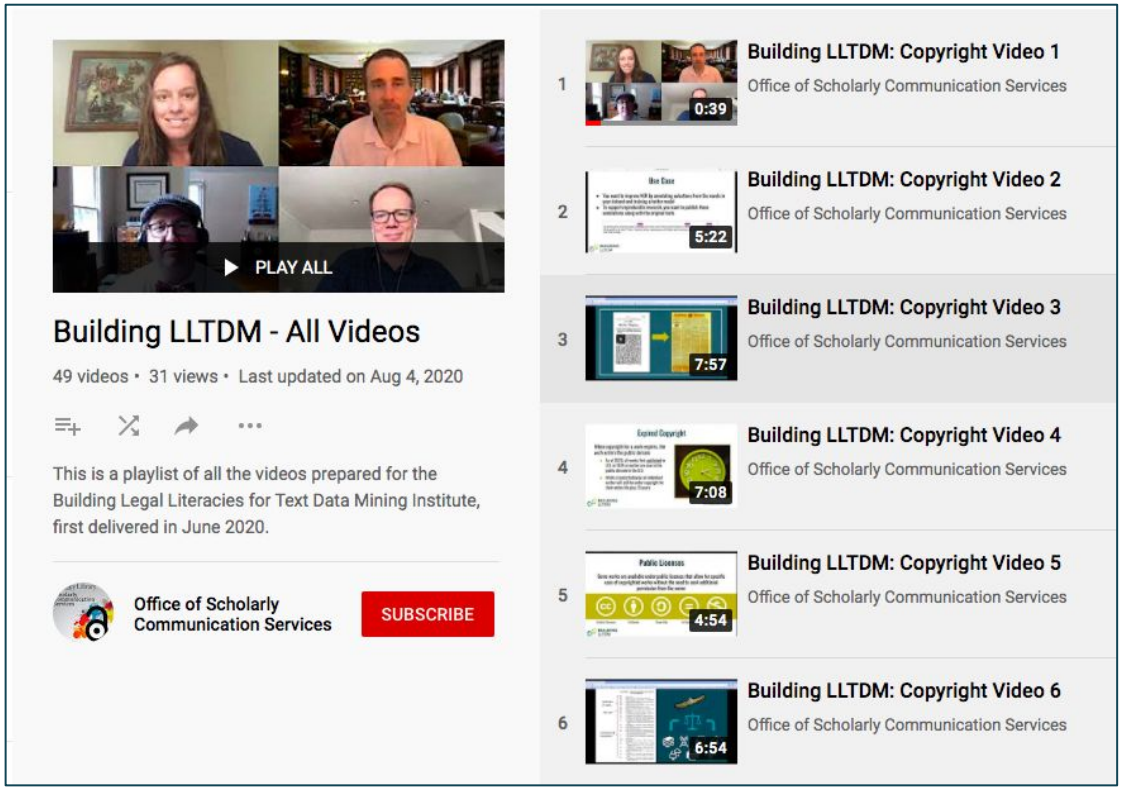


- # of female characters
- # of usages of words like “tender,” “heart,” “sentimental,” and “love” by female characters
- Instances of “fainting” or “whimpering” by males vs. females

Legal Literacies for Text Data Mining

Building Legal Literacies for Text Data Mining Institute





Need for Cross-Border Guidance

1. Materials to be mined are outside of U.S. or subject to foreign licensing
2. Study subjects or content creators live in another country
3. Project collaborators reside abroad

Cross-Border Constraints

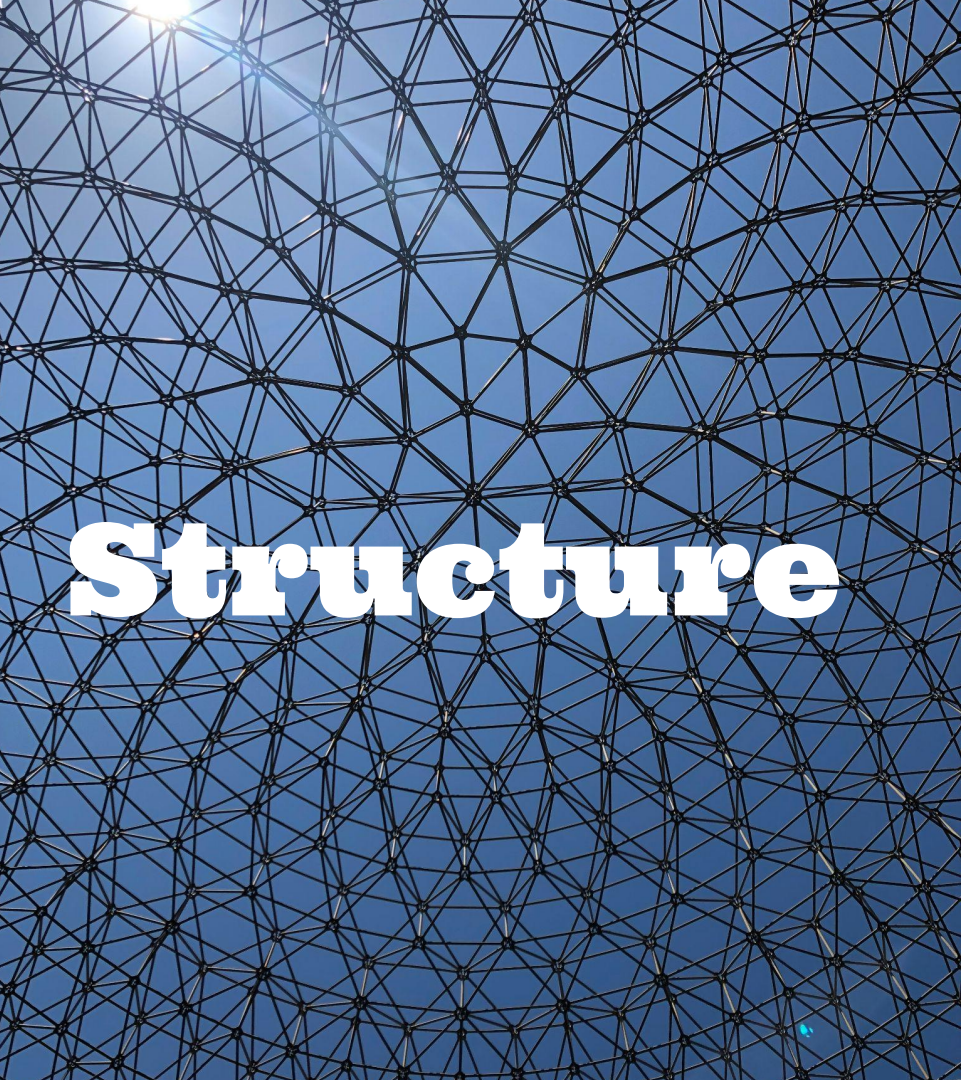
Copyright: 70% reported problems
Licensing: 72% reported problems

Impeded or prevented entirely: 28%
Hesitant to share methods & sources: 40%

- “Slowed down the project”
- Tried “not to ask too many questions” because of the law

Examples of cross-border questions

- Assemble a corpus of foreign-published materials?
- Share TDM corpus with researchers outside of U.S.?
- Circumvent TPMs?
- Privacy laws of other countries?
- How to collaborate with research subjects in authoritarian regimes?
- Involve foreign IRBs?
- Privacy and ethics in personal materials?



Structure

- Pre-round table Practitioner **project sharing**
- Three virtual **round tables**
- Expert **feedback** for Practitioners

White Paper

https://escholarship.org/uc/item/5k91r1s1

Legal Literacies for Text Data Mining – Cross-Border ("LLTDM-X"): White Paper

Uncertainty about cross-border LLTDM issues **hinders TDM researchers**, confirming need for further research and education.

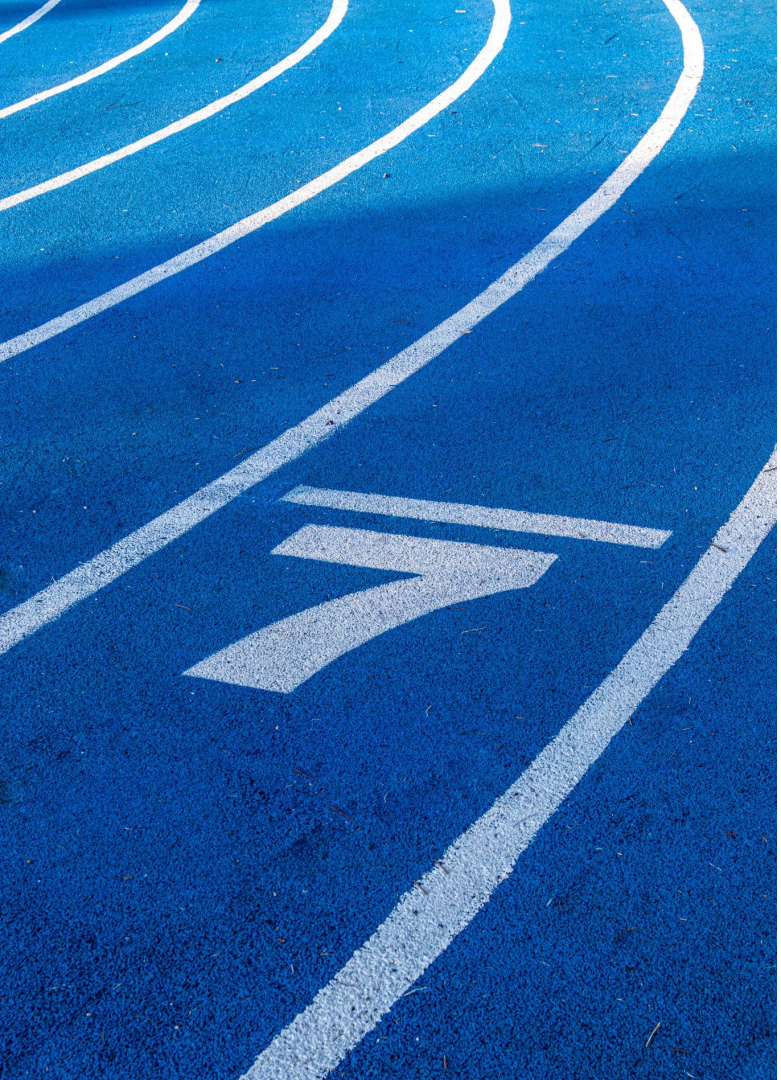
Broader **education regarding U.S.-centric LLTDM literacies** should also continue.

Disparities in national laws may incentivize **TDM researcher “forum shopping”** and exacerbate scholarly bias.

License agreements often dominate analysis of cross-border TDM permissibility.

Lawsuits and potential rulemaking about generative artificial intelligence may impact understanding of fair use and other research exceptions.

Overall **“risk” may remain difficult to quantify**.

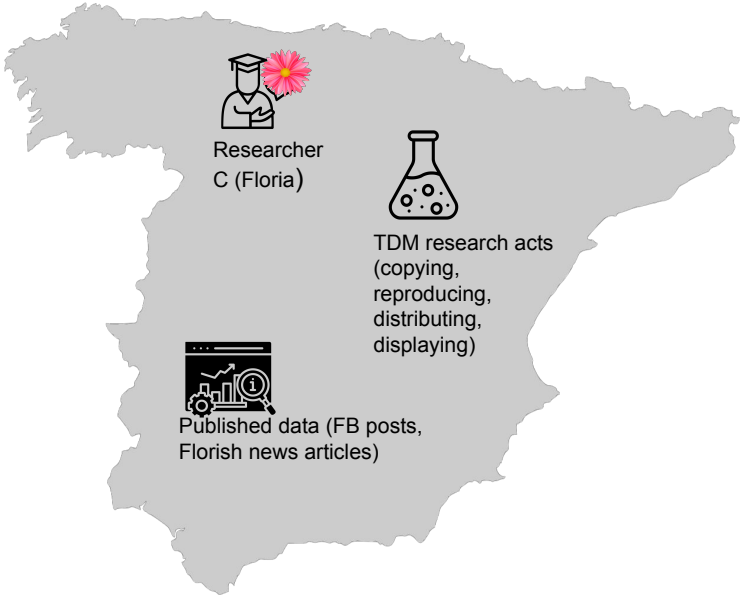


Institutional review boards (IRBs) have an opportunity to build partnerships to support researchers. **Institutions can also adopt researcher-friendly policies.**

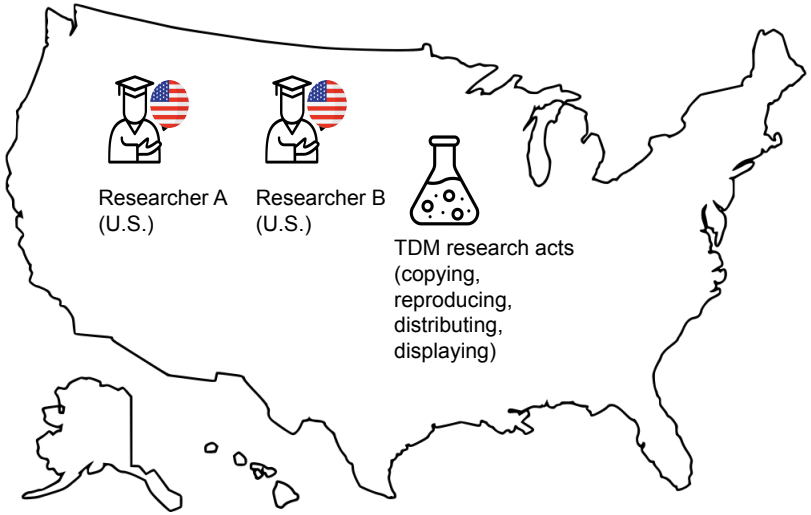
Case Study

https://escholarship.org/uc/item/1w03f9r2

Legal Literacies for Text Data Mining – Cross-Border ("LLTDM-X"): Case Study



Foreign country (Floria)



United States

Hypothetical

Variables by Legal Literacy

Researcher question(s): If the copyrighted materials (e.g. Facebook posts, newspaper articles) to be used for text data mining originated in a foreign country (e.g. Floria), does the foreign country's copyright law apply to the infringement analysis in the U.S.?

Preliminary guidance: No. U.S. courts will apply U.S. law and fair use (17 USC § 107) to acts like reproduction, distribution, display, etc.—i.e. all "exclusive rights" that copyright owners have in copyright protected works—performed in the U.S., regardless of the country of origin of the source material, and regardless of whether the research results are later viewed online outside of the U.S.

Future Directions

- **Engage** additional disciplinary communities (e.g., social sciences, sciences)
- **Align** campus resources to strategically address ethical dimensions of work at policy & practice level (e.g., IRB)
- **Expand** work to address implications of AI

Q & A

Thomas Padilla (Internet Archive)
Rachael Samberg (UC Berkeley)
Timothy Vollmer (UC Berkeley)