

# Legal Literacies for Text Data Mining: Cross-Border



NATIONAL  
ENDOWMENT  
FOR THE  
HUMANITIES

Berkeley Library  
UNIVERSITY OF CALIFORNIA

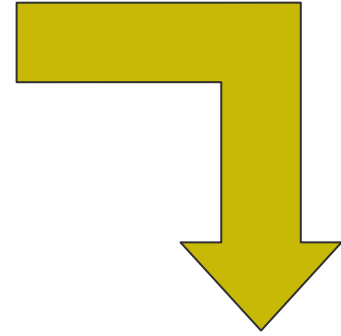
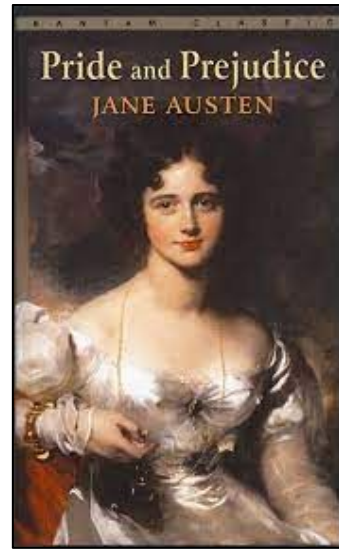


# Summary



- **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

Copyright



Contracts



Privacy



Ethics & Policy



# Building Legal Literacies for Text Data Mining Institute

## Building Legal Literacies for Text Data Mining

What to Know & How to Teach It



Project sponsored by:



NATIONAL  
ENDOWMENT  
FOR THE  
HUMANITIES



### Building LLTDM - All Videos

49 videos • 31 views • Last updated on Aug 4, 2020









This is a playlist of all the videos prepared for the Building Legal Literacies for Text Data Mining Institute, first delivered in June 2020.



Office of Scholarly  
Communication Services

SUBSCRIBE

-  **Building LLTDM: Copyright Video 1**  
Office of Scholarly Communication Services  
0:39
-  **Building LLTDM: Copyright Video 2**  
Office of Scholarly Communication Services  
5:22
-  **Building LLTDM: Copyright Video 3**  
Office of Scholarly Communication Services  
7:57
-  **Building LLTDM: Copyright Video 4**  
Office of Scholarly Communication Services  
7:08
-  **Building LLTDM: Copyright Video 5**  
Office of Scholarly Communication Services  
4:54
-  **Building LLTDM: Copyright Video 6**  
Office of Scholarly Communication Services  
6:54

# 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

<b>Summary</b>	<b>2</b>
<b>Project Origins and Goals</b>	<b>3</b>
Growth of TDM in Digital Humanities	3
Training for U.S. Law and Policy Hurdles	3
Similar Need for Cross-Border Guidance	4
<b>Project Contributors and Activities</b>	<b>6</b>
Project Contributors	6
Project Team	6
Practitioners	7
Experts	7
Financial support	8
Activities	8
Identifying Practitioners and Experts	8
Pre-Round Table Preparation & Statements	9
Round Table 1	10
Round Tables 2 and 3	10
<b>Project Outcomes</b>	<b>11</b>
Expert feedback for Practitioners	11
Case study & white paper	11
Project documentation	11
<b>Takeaways &amp; Recommendations</b>	<b>12</b>
Project Takeaways	12
1. Uncertainty about cross-border LLTDM issues hinders U.S. TDM researchers, confirming the need for further research and education.	12
2. Broader education regarding U.S.-centric LLTDM literacies should also continue.	12
3. Disparities in national laws may incentivize TDM researcher “forum shopping” and exacerbate scholarly bias.	14
4. License agreements often dominate analysis of cross-border TDM permissibility.	16
5. Emerging lawsuits about generative artificial intelligence may impact understanding of fair use and other research exceptions in cross-border TDM.	17
6. Research is needed into issues of foreign jurisdiction, likelihood of lawsuits in foreign countries, and likelihood of enforcement of foreign judgments in the U.S. However, overall “risk” of proceeding with cross-border TDM research may remain difficult to quantify.	18
7. Institutional review boards (IRBs) have an opportunity to explore a new role or build partnerships to support researchers engaged in cross-border TDM.	19
Next steps & Recommendations	20



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.**





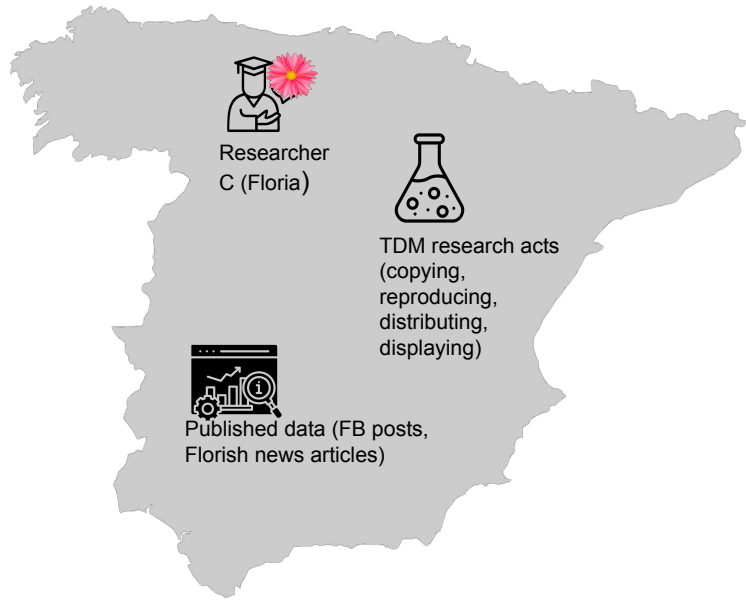
**Institutional review boards (IRBs) have an opportunity** to build partnerships to support researchers. **Institutions can also adopt researcher-friendly policies.**

# Case Study

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

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

<b>Research Scenario</b>	<b>2</b>
<b>Paradigm 1: U.S.-based researchers perform all TDM acts in the U.S.</b>	<b>3</b>
1. Copyright Variables	3
a. Foreign-created materials	3
b. Publication status	3
c. Presence of technological protection measures	4
d. Geographic limitations on data or corpus sharing	4
e. Known foreign infringement; subsequent U.S. reliance / use	4
f. Risk of foreign lawsuit for copyright infringement	5
2. Contractual Variables	6
a. Cross-institutional data or corpus sharing	6
b. Country of origin's impact on license agreement or website Terms of Service	7
c. Contractual impact on data or corpus sharing / republication	7
d. Risk of foreign lawsuit for contractual breach	8
3. Privacy & Ethics Variables	8
a. Applicability of foreign privacy laws to U.S.-based TDM	8
b. Use of data beyond original intent	8
c. Sensitive but not legally private data	9
d. Risk of foreign lawsuit for privacy violations	9
4. Risk assessment	9
<b>Paradigm 2: U.S.-based researchers engage with collaborator abroad, or otherwise perform TDM acts in both U.S. and abroad</b>	<b>11</b>
1. Copyright Variables	11
a. Foreign exercise of protected rights	11
b. Corpus creation abroad, or in U.S. and abroad; sharing data or corpus across borders	13
c. Presence of technological protection measures	14
d. Known foreign infringement; subsequent U.S. reliance / use	14
e. Place of output publication	14
f. Risk of foreign lawsuit for copyright infringement	15
2. Contractual Variables	15
a. Foreign country prohibits contractual override of copyright exceptions, but override permitted in the U.S.	15
b. U.S. contract preserves / authorizes the protected right, but foreign country's copyright laws prohibit it	16
c. Risk of foreign lawsuit for contractual breach in U.S.	16
3. Privacy & Ethics Variables	16
a. Applicability of foreign privacy laws to the U.S. researchers	16
b. Use of data beyond original intent	17
c. Sensitive but not legally private data	17
d. Risk of foreign lawsuit for privacy violations	18
4. Risk assessment	18



Researcher C (Floria)

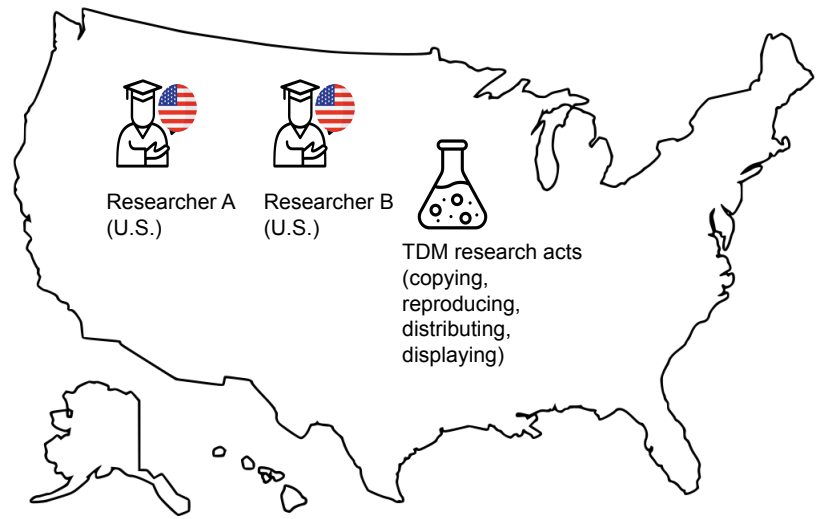


TDM research acts (copying, reproducing, distributing, displaying)



Published data (FB posts, Florish news articles)

Foreign country (Floria)



Researcher A (U.S.)



Researcher B (U.S.)



TDM research acts (copying, reproducing, distributing, displaying)

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)