Building Legal Literacies for Text Data Mining: Institute White Paper

Project Summary **2**

Project Origins & Goals **2**

Growth of Text Data Mining in Digital Humanities 2

Law and Policy Hurdles 3

Past Work Demonstrated Need for Training 4

Goals 6

Project Overview **6**

People 6

Faculty 6

Participant Recruitment 7

Financial support 9

Pre-institute Preparation 10

Institute Schedule & Activities 11

Design Thinking Approach 11

Daily Agenda 12

Post-Institute Meeting 15

Open Educational Resource 15

Impact, Reflections, & Next Steps **16**

Impact 16

Pedagogical Reflections 18

Next Steps 21

Cross-Border Issues Need Future Institutes 21

Need for Documentation Templates 21

Appendices **22**

Project Summary

Until now, digital humanities (DH) researchers conducting text data mining (TDM) in the U.S. have had to maneuver through a thicket of legal issues without much guidance or assistance. Uncertainty about the breadth and contours of TDM rights and obligations has impeded the scope of DH research questions, or unnecessarily exposed scholars to risk. We designed Building Legal Literacies for Text Data Mining (Building LLTDM) to address these questions and barriers to facilitate DH TDM research. Funded as an NEH Institute for Advanced Topics in the Digital Humanities, and hosted by UC Berkeley from June 23-26, 2020, Building LLTDM provided 32 DH TDM researchers, librarians, and professionals with foundational skills to:

1. confidently navigate law, policy, ethics, and risk within DH TDM projects;
2. integrate workflows at their home organizations to provide law and policy support for DH TDM projects;
3. practice sharing these new skills and workflows through authentic consultation exercises;
4. prototype plans for broadly disseminating their knowledge; and
5. develop communities of practice to promote cross-institutional outreach about the DH TDM legal landscape.

While we originally planned Building LLTDM to be held on the UC Berkeley campus, the COVID-19 pandemic necessitated a transition to online teaching. Our faculty of legal experts, librarians, and researchers from across the U.S. provided interactive remote instruction. We presented the substantive content through pre-recorded videos and held live group discussions in a flipped classroom model. We also provided the video transcripts and slides to participants to promote accessibility and accommodate multiple learning styles.

To maximize the reach and impact of Building LLTDM, we compiled the legal literacies covered during the institute into an Open Educational Resource (OER) with a public domain (CC0) dedication. The OER covers copyright (both U.S. and international law), technological protection measures, privacy, and ethical considerations. It also helps other DH professionals and researchers run their own similar institutes by describing in detail how we developed and delivered programming (including our pedagogical reflections and take-aways), and includes ideas for hosting shorter literacy teaching sessions.

Project Origins & Goals

Growth of Text Data Mining in Digital Humanities

If one were to crack open popular English-language novels written in the 1850s—say, ones from Brontë, Hawthorne, Dickens, and Melville—one would find they describe men and women in very different terms. While a male character might be said to “get” something, a female character is more likely to have “felt” it. Whereas the word “mind” might be used when

describing a man, the word “heart” is more likely to be used about a woman. As the 19th Century became the 20th, these descriptive differences between genders diminish within these novels. And we know all this because researchers have used automated techniques to extract information from the novels, and analyzed word usage trends at scale.[1] They crafted algorithms to turn the language of those novels into *data about* the novels.

In fields of inquiry like the digital humanities, the application of such automated techniques and methods for identifying, extracting, and analyzing patterns, trends, and relationships across large volumes of unstructured or thinly-structured digital content is called “text data mining” or “TDM”. (One may also see it referred to as “text and data mining” or “computational text analysis”). TDM is an increasingly important and prevalent research methodology leveraging algorithms to sift, organize, and analyze vast amounts of thinly-structured textual content.[2] For instance, these methods make it possible to: discern racial disparity by evaluating language from police body camera footage;[3] assess visual culture;[4] and examine conversation patterns on Twitter regarding social justice issues such as violence against women.[5] TDM methodologies and tools continue to expand, posing great opportunities for advancements across education, literature, society, politics, and beyond.[6]

Law and Policy Hurdles

Until Building LLTDM, DH researchers conducting TDM faced confusing legal considerations, and a marked absence of community guidance for navigating them. For instance, imagine that researchers wish to digitally crawl and download content about Egyptian tombs and artifacts from online websites, in order to conduct an automated computational analysis on the web-scraped materials. Then imagine the researchers also want to share these content-rich datasets to encourage research reproducibility or enable other scholars to query the datasets with new questions. This kind of work can raise issues of:

- **Copyright** (e.g. Are the images protected by copyright? Does an exception like fair use apply?)

---

[1] Underwood, T., Bamman, D., & Lee, S. (2018). The transformation of gender in English-language fiction. *Journal of Cultural Analytics*. Available at https://doi.org/10.22148/16.019
[2] Hearst, M. (2003, October 17). What is Text Mining? Available at http://people.ischool.berkeley.edu/~hearst/text-mining.html.
[3] Voigt, R., et al., (2017). Language from police body camera footage shows racial disparities in officer respect. *Proceedings of the National Academy of Sciences*, *114*(25), 6521. Available at https://doi.org/10.1073/pnas.1702413114.
[4] Arnold, T., & Tilton, L. (2019). Distant viewing: Analyzing large visual corpora. *Digital Scholarship in the Humanities*. Available at https://doi.org/10.1093/digitalsh/fqz013.
[5] Xue, J., et al., (2019). Harnessing big data for social justice: An exploration of violence against women-related conversations on Twitter. *Human Behavior and Emerging Technologies*, *1*(3), 269–279. Available at https://doi.org/10.1002/hbe2.160.
[6] Hassani, H., et al., (2020). Text Mining in Big Data Analytics. *Big Data and Cognitive Computing*, *4*(1). Available at https://doi.org/10.3390/bdcc4010001.

- **Contracts** (e.g. Are there database license agreements or website terms of use that govern what researchers are permitted to scrape or download? Do these agreements override copyright exceptions?)
- **Privacy** (e.g. Do the images reveal information that could infringe upon the privacy rights of the subjects under federal and state laws? Does downloading images that should not have been made public constitute a further privacy violation?)
- **Ethics** (e.g. Are there social and religious customs, or other circumstances like indigenous knowledge that could impact downloading and use of the materials?)

If researchers are not comfortable navigating these issues or feel that, in doing so, they or their institutions would take on too much risk, they may abandon their projects. Indeed, a study of humanities scholars' text analysis needs found that access to and use of copyright-protected texts was a "frequent obstacle" in participants' ability to select appropriate texts for TDM.[7]

Potential legal hurdles do not just deter TDM research; they also bias it toward particular topics and sources of data. In response to confusion over copyright, website terms of use, and other perceived legal roadblocks, some digital humanities researchers have gravitated to low-friction research questions and texts (e.g. materials exclusively in the public-domain or datasets already compiled) to avoid decisions about rights-protected data. Restricting research to such sources can skew inquiries, leave important questions unanswered, and render resulting findings less broadly applicable. A growing body of research also demonstrates how race, gender, and other biases found in openly available texts have contributed to and exacerbated bias in developing artificial intelligence tools.[8]

Sound guidance from information professionals can help researchers traverse these concerns. Yet, scholars have reported hesitation to seek help from institutional staff whom they fear will question the legality of their TDM methods, or advocate for a more risk-averse approach than the law warrants. Those worries may be validated when libraries sign or enforce license agreements to datasets with unclear or, in some cases, hostile TDM provisions. If equipped with legal and ethical literacies, institutional staff as well as researchers would be better positioned to understand what the law already permits, and negotiate for better usage rights overall.

Past Work Demonstrated Need for Training

For all of these reasons, our project team wanted to help DH scholars and research professionals better navigate the law and policy landscape of TDM—using a pedagogical approach[9] that enables researchers to fully and fairly utilize rights-protected works, and

[7] Green, H., et al., (2016). Scholarly Needs for Text Analysis Resources: A User Assessment Study for the HathiTrust Research Center. *Proceedings of the Charleston Library Conference*. Available at http://dx.doi.org/10.5703/1288284316464.
[8] Levendowski, A. (2018). How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem. 93 Wash. L. Rev. 579. Available at https://digitalcommons.law.uw.edu/wlr/vol93/iss2/2.
[9] An early formulation of that approach was articulated in project team members' 2019 paper: Samberg, R. G., & Hennesy, C. (2019). Law and literacy in non-consumptive text mining: Guiding researchers

disseminate their resulting TDM scholarship broadly. Our intended framework would also need to support TDM researchers in understanding and navigating ethical issues like corpus bias and subject consent.

We began designing our institute by canvassing existing educational programs—which cemented the need for our training. In reviewing a broad sample of digital humanities, humanities, and information science curricula, professional development training programs, and library guides, we found scant trainings or resources that integrate TDM legal literacies into outreach and instruction. While there were a growing number of DH training opportunities on TDM methods and tools, they almost universally omitted copyright and other law or policy concerns. Moreover, our own experiences suggested that DH scholars and professionals face many of the questions that arise around legal issues and TDM at the time of crisis (e.g., when university access to a database is suspended due to systematic downloading). This places undue stress on DH scholars' ability to conduct DH TDM research and may lead institutions to unduly restrict such research via institutional policy.

We understood that addressing this educational need would require cross-organizational training aimed at both (1) the *scholars conducting TDM,* and (2) the *professional staff who assist and collaborate* with them. Digital humanities professionals are people like librarians, consultants, and other institutional staff who conduct digital humanities text data mining or aid researchers in their text data mining research. The DH professional stakeholder group is essential to maximizing the efficacy of legal literacy education for a number of reasons. DH professionals who teach or consult on TDM are well-positioned to incorporate legal literacies into existing trainings. Further, academic libraries, labs, and departments license many of the databases and datasets DH researchers seek to use. Staff are then called upon to provide input on or information about database terms and conditions, and they may be positioned to secure better licensing terms from the start. Many libraries also employ legal experts within scholarly communications or copyright units—some of whom have established TDM training programs and service models that could be adjusted to incorporate law and policy workflows.

There was ample reason to believe that an institute devoted to the development of these legal literacies for DH researchers and professionals would be highly productive. For example, copyright training sessions for librarians have already been found to be effective in building understanding and confidence around copyright research consultations. Educating DH researchers and professionals through a focused institute also offers the benefit of creating shared understanding across the scholarly landscape. This in turn would offer the potential for downstream impact as all participants would be poised to return to their home institutions or professional communities and share what they have learned.

---

through the landscape of computational text analysis. *Copyright Conversations: Rights Literacy in a Digital World* (pp. 289–315). ACRL. Available at https://escholarship.org/uc/item/55j0h74g.

Goals

The law and policy impediments to DH TDM research, coupled with the need for training to help researchers and professionals navigate them, prompted us to provide DH professionals and researchers with foundational skills to:

1. confidently navigate law, policy, ethics, and risk within DH TDM projects;
2. integrate workflows at their home organizations to provide law and policy support for DH TDM projects;
3. practice sharing these new skills and workflows through authentic consultation exercises;
4. prototype plans for broadly disseminating their knowledge; and
5. develop communities of practice to promote cross-institutional outreach about the DH TDM legal landscape.

Project Overview

Our aim is to facilitate the replicability of Building LLTDM institute by others. Accordingly, in this section we detail project design and administration chronologically for easier implementation:

1. Faculty & participant recruitment
2. Provision of financial support
3. Pre-institute preparation
4. Institute schedule & activities
5. Post-institute “catch up”
6. Creation of OER

People

Faculty

Building LLTDM was led by the Office of Scholarly Communication Services at the University of California, Berkeley Library. Rachael Samberg, Scholarly Communication Officer & Program Director, served as Project Director. She oversaw curricular design and execution, as well as the administrative and operational aspects of the institute. Timothy Vollmer, Scholarly Communication & Copyright Librarian, served as Project Manager, and was responsible for coordinating the design and execution of the institute, and streamlining administrative and operational aspects. Both the Project Director and Project Manager also served as faculty instructors for the institute, helping to create and deliver educational materials and training.

The remaining institute faculty hailed from more than a dozen North American universities and institutions, and were each responsible for contributing to institute curricular design and delivery. Faculty were recruited through professional connections and networks, and were composed of legal experts (“LE”), librarians (“L”), and humanities researchers (“HR”). Their real-world roles

straddled these boundaries (e.g. some legal experts are also librarians); yet, the nominal divisions ensured that institute sessions were led by a set of experts who collectively offer a full range of relevant DH expertise. We also had an additional legal expert on call via e-mail during the institute to field any questions that instructors were unable to answer in real time.

Project team members included:

- Scott Althaus, Professor of Political Science & Communication, and Director of the Cline Center for Advanced Social Research at University of Illinois
- David Bamman, Assistant Professor at UC Berkeley's School of Information
- Brandon Butler, Director of Information Policy at the University of Virginia (UVA) Library
- Beth Cate, Associate Professor at Indiana University Bloomington's School of Public and Environmental Affairs (SPEA)
- Kyle K. Courtney, Copyright Advisor for Harvard University, within the Office for Scholarly Communication
- Sean Flynn, Associate Director of the Program on Information Justice and Intellectual Property (PIJIP) and Professorial Lecturer in Residence
- Maria Gould, Research Data Specialist/Product Manager, California Digital Library
- Cody Hennesy, Journalism and Digital Media Librarian at University of Minnesota
- Eleanor Dickson Koehl, HathiTrust Digital Scholarship Librarian at the University of Michigan Libraries, and Associate Director for Outreach and Education, HTRC
- Thomas Padilla, Visiting Digital Research Services Librarian at University of Nevada Las Vegas
- Stacy Reardon, Literatures and Digital Humanities Librarian at UC Berkeley
- Matthew Sag, Professor of Law at Loyola University Chicago School of Law
- Brianna L. Schofield, Executive Director of Authors Alliance
- Glen Worthey, Associate Director for Research Support Services, HathiTrust Research Center
- Megan Senseney, Head of the Office of Digital Innovation and Stewardship at University of Arizona Libraries
- Sara Benson, Copyright Librarian at University of Illinois

Participant Recruitment

We designed the institute to support 32 participants, resulting in what we believed would be a suitable instructor-to-attendee ratio to accommodate the highly immersive and discursive aspects of a design thinking framework (discussed further below).

We sought participation from both DH researchers and professionals. We anticipated that both groups would have mutually beneficial insights and experiences to share. For instance, DH researchers would benefit from LLTDM training that can be applied to their own research projects and publications, and integrated into their teaching and advising—thereby broadening downstream community impact. Conversely, DH professionals are often the first contact point for DH researchers with law-related TDM questions; handle licensing and negotiate access to datasets and digital collections for TDM; and provide training and documentation for DH

researchers on workflows and tools. Educating DH professionals would enable ongoing institute impact as these professionals can bring the skills they have gained back to their own campuses and professional communities. We also aimed for approximately equal numbers of DH researchers and DH professionals to maximize impact—recognizing that these two groups are variously situated in their organizations and thus can provide future advocacy and support in different ways. For similar reasons, we encouraged participation from institutional pairs of participants (e.g. one digital humanities researcher and one professional affiliated with that same organization or project) with the hope that greater representation from a given institution could result in broader literacy implementation at that institution following participant training.

With institute scope and intended reach determined, we developed a project website to host information about the institute application process, timeline, and criteria. We advertised the application process on the Building LLTDM blog, digital humanities and library-related email lists, and social media. The submission window was open for two months, and the application process required individuals to submit two documents: (1) a current CV, and (2) a 2-page (maximum) letter of interest. In their letters of interest, we asked applicants to account for experience with or interest in: the intersection of TDM in DH research and the law; their goals for applying knowledge and skills to be acquired at the institute to their own activities; their goals for sharing knowledge and skills with others at their home institutions/affiliations; and, how they might support the institute's commitment to diversity and equity.

We posted selection criteria prominently on the Building LLTDM website, and gave particular influence to diversity, equity, and inclusion. In particular, the faculty believed that the institute would work best if it reflected the race and gender demographics of the broader population, and not just those of higher education—and we strived to achieve equity by reflecting these more representative demographics. Additionally, we worked to develop a participant group that was representative of different institution types, research advising and support experience, professional roles, levels of experience with DH TDM research, career stages, and disciplinary perspectives.

The selection process took place over two main rounds. First, a subset of the faculty conducted an initial assessment of all applications based on the selection criteria. Our subgroup then met in successive sessions to discuss and normalize rankings, and reached consensus on recommended candidates. We presented our recommendations to the project team for discussion in a full group meeting. The suggested group was composed of 15 DH researchers and 17 DH professionals hailing from 15 different states. We are also pleased to report that all of our selected participants accepted our offer to be institute participants, and included:

- Ilya Akdemir, University of California, Berkeley
- Tara Baillargeon, Marquette University
- Trevor Burrows, Purdue University
- Matthew Cannon, University of California, Berkeley
- Nathan Carpenter, Illinois State University
- Ashleigh Cassemere-Stanfield, University of Chicago

- James Clawson, Grambling State University
- Mark Clemente, Case Western Reserve University
- Quinn Dombrowski, Stanford University
- Alyssa Fahringer, George Mason University
- Heather Froehlich, Penn State University
- Nicole Garlic, Temple University
- Casey Hampsey, New York University
- Devin Higgins, Michigan State University
- Christian Howard, Bucknell University
- Daniel Johnson, Notre Dame University
- Spencer Keralis, University of Illinois
- Sarah Ketchley, University of Washington
- Melanie Kowalski, Emory University
- Barbara Levergood, Bowdoin College
- Jes Lopez, Michigan State University
- Rochelle Lundy, Seattle University
- Jon Marshall, UC Berkeley
- Jens Pohlmann, Stanford University
- Caitlin Pollock, University of Michigan
- Sarah Potvin, Texas A & M University
- Andrea Roberts, Texas A & M University
- Daniel Royles, Florida International University
- Hadassah St. Hubert, Florida International University
- Todd Suomela, Bucknell University
- Nicholas Wolf, New York University
- Madiha Zahrah Choksi, Columbia University

Financial support

On our project website, we made clear to potential applicants that participant stipends would be distributed in advance of the institute. This was designed to promote equity by helping participants avoid having to expend personal funds or await reimbursement. Had the institute been in person, the paid-in-advance stipends would have been sufficient to cover travel, lodging, and related expenses with the aim of eliminating out-of-pocket expenses. As we found ourselves having to rapidly transition the institute online while participant stipends were concurrently being distributed by the university business office, we conferred with our NEH program officer about how to proceed. With NEH guidance, we maintained stipend distribution as awarded in the grant—with stipends being repurposed to compensate for participant time and incentivize participation.

We also offered instructor honoraria to faculty. The honoraria were originally intended to both (1) cover faculty travel costs to the institute, and (2) recognize the substantial contributions project team members were making for developing and teaching curriculum and creating the post-institute OER. (No faculty member time was being charged to the grant, and instead all

efforts were contributed from people's personal time.) As COVID-19 unfolded, and as with participant stipends, we consulted with our NEH program officer and were advised that honoraria should similarly continue, with a focus shifting to rewarding faculty contributions.

Pre-institute Preparation

After participants were chosen, the pre-institute timeline was filled with both substantive and logistical planning:

- ***Four months pre-institute*:** While simultaneously developing instructional content, we also began regular communications with participants, which increased in frequency as the pandemic spread. We began communications through individual and group e-mails. As the start date for the institute approached, we transitioned to Slack for announcements and community information sharing, and to help build familiarity and collegiality. We created a Slack sub-channel for faculty and participant introductions. In addition, faculty and participants created sub-channels to discuss specific TDM research areas, such as social media and oral histories.

- ***One month pre-institute*:** We sent participants a short questionnaire so that the faculty instructors could learn more about participants' research or professional practices related to TDM. This allowed faculty to better understand the participants' real-world experiences and struggles, and tailor the upcoming sessions and exercises to properly meet participant expectations and needs. We also developed a Faculty Facilitation Guide (referred to as the "Faculty Packet") for instructors to help faculty prepare for administering the institute. This Google doc contained faculty and participant contact information, information about how to use Zoom effectively, and guidance about how to support participant contributions and positive interactions during the online institute.

- ***One week pre-institute:*** We distributed pre-reading to participants that provided an overview of the TDM legal and policy environment. However, we kept the amount of required preparation to a minimum—both because we knew the participants were busy individuals with full time jobs and research responsibilities, and also due to the added pressure and stresses of the COVID-19 pandemic. We set the expectation that we hoped the participants would be able to provide as much undivided attention as they could during the actual week of delivery (of course understanding that there might be necessary interruptions due to family or personal responsibilities because of the remote nature of the institute).

  We also distributed a comprehensive guide for participants that we called the Participant Packet—essentially a one-stop-shop to guide participants through the week ahead. The Participant Packet included:
  - Information about how to communicate with faculty and other participants
  - Instructions for how to use Zoom during institute sessions
  - Institute code of conduct (to which the cohort had consented upon acceptance of the offer to participate)

- Information about social media usage and the applicability of the Chatham House Rule to protect participant communications
- Day-by-day agenda for the institute, including assigned meeting groups of various sizes (plenary, small group), free-write activities, and also links to Zoom rooms and shared notes documents for each session
- Links to readings and pre-recorded short videos (with transcripts and slides) so that participants could be prepared for the next day's topics[10]

Institute Schedule & Activities

Design Thinking Approach

We believed that design thinking offered an apt instructional framework to convey the literacies while sufficiently engaging participants. Design thinking relies upon experiential meeting methodologies that foster hands-on learning and allow participants to experiment with developing their own solutions for their TDM hurdles. The institute tracked the five stages of design thinking as follows:

- ***Empathy (Institute Day 1)*:** Building trust and common understanding through experience sharing can foster robust discussion and collaborative inquiry. We thus began the first day of the institute by developing our collective understanding of participants' experiences with TDM through storyboarding sessions. These empathy-supporting activities served as an opportunity for participants to get to know each other and to start learning about each other's hurdles and successes with the LLTDM literacies. The exchanges helped participants discover that they are not alone in their struggles but rather are part of a burgeoning community.

- ***Define & Ideate (Institute Days 2 & 3)*:** For days two and three, we cycled iteratively through the "define" and "ideate" phases of design thinking. Defining and ideation are foundational for developing a shared language to discuss the contours of TDM challenges, and to lay the groundwork for participants to strategize about customized solutions. For these stages, faculty worked with participants to articulate and contextualize TDM issues and literacies through: (1) asynchronous videos conveying the substantive literacies, and (2) synchronous small group time to discuss case studies and undertake "putting it together" exercises (more below under "Daily Agenda") to simulate real-world problems.

- ***Prototype & Test (Institute Day 4; Post Institute)*:** Prototyping involves developing a personalized approach to implementing takeaways and solutions. To model this stage,

[10] In order to provide easy public viewing for all the pre-recorded TDM topical videos, we uploaded them to the UC Berkeley Library's Office of Scholarly Communication Services YouTube account. Viewers can also speed up or slow down the video playback, or turn on closed captions; both features are offered automatically by YouTube. We also created playlists under each topical area (copyright, international copyright, licensing, technological protection measures, and privacy & ethics), as well as a comprehensive playlist containing all the videos.

on the final day of the institute, participants crafted implementation plans regarding how they will integrate the literacies into their work and at their home institutions. Testing would then occur in the months following the institute, as participants put their plans into place. To follow up on testing, we stayed connected through Slack and reconvened the cohort eight months after the institute to learn from each other's outcomes (more below under "Post-Institute Meeting").

Daily Agenda

General Schedule

We adjusted the institute's prime content delivery mechanism to asynchronous (pre-recorded) instructional videos so that we could utilize synchronous sessions for small group discussions and experiential exercises. This minimized sedentary time in front of a computer and allowed participants the opportunity to pace themselves according to their personal schedules and learning styles. We also spaced sessions with intermittent breaks to help participants focus and avoid Zoom burnout. Because participants and faculty were joining from different time zones, we began sessions at 8 a.m. Pacific Time and concluded by 2 p.m. Pacific Time[11], to wrap by the end of normal business hours on the East Coast. This allowed participants sufficient time to prepare for each subsequent day's content and activities irrespective of time zones.

Day 1

1. *Introductions and stage setting*: Faculty instructors used a master slide deck throughout the week. Day 1 began with a welcome, logistical information, explanation of the code of conduct and Chatham House Rule, and a framing for the week's activities. One of the faculty instructors also served as an institute moderator. The moderator's key roles were to: (1) observe and synthesize emerging themes from each day to bolster learning outcomes, and (2) assist with cross pollination of ideas and themes from across small breakout groups. The moderator tuned in to small group discussion sessions and collected individual reflections for sharing at the end of each day.

2. *Empathy building exercise*: Following the moderator's introduction, participants engaged in a virtual white board exercise designed to help them storyboard their own experiences with TDM; build knowledge and understanding among participants; and surface aspects of divergence and convergence across individual experiences. We used the online "sticky note" software tool called Mural for this journey mapping exercise.

3. *Free Write*: Day 1 ended with a free write exercise (the first of three such exercises over the course of the week). Freewriting was intended as an opportunity to reflect on the day's sessions and apply them to one's personal circumstances, research interests, institutional culture, team dynamics, etc. Participants were asked to write without pausing or proofreading and in response to the following prompts:

[11] At the end of each day, we offered optional and informal "Happy Half-Hours" on Zoom. This time was to socialize, decompress, and answer participant questions.

- What did you learn from other participants today about variations in TDM processes and logistical complexities?
- Which pain points highlighted by other participants resonated with you?
- What new questions, concerns, or opportunities emerged during report outs that you didn't capture on the mural board?

Participants e-mailed their text to our shared faculty email group. The institute moderator and several instructors reviewed the submitted responses each evening in preparation for an opening reflection to kick off the next day.

Day 2

1. *Report back from moderator on free write themes:* At the beginning of day 2, the moderator summarized the motifs and lessons evidenced in the previous day's free writes. This practice reminded participants about the themes discussed the day before, and helped them track progress and accomplishments throughout the week.

2. *Substantive literacies—Copyright, international copyright, TPMs:* On day 2, we began to explore the substantive law and policy literacies for text data mining in the digital humanities. We covered copyright (focusing heavily on U.S. law), copyright in the international/cross-border context, and technological protection measures. As mentioned above, participants were asked to watch short pre-recorded videos made by the faculty, as well as view slides and video transcripts.

3. *"Putting it together" exercise:* After the morning substantive sessions, faculty and participants engaged in a real-world simulated exercise. This activity required individual reading and reflection, as well as small- and medium-sized group discussions, on a pre-prepared TDM scenario.

4. *Free Write*: Day 2 ended with another 15-minute free write exercise, with prompts tied to the day's learnings:
    - What copyright concerns do you have about accessing data for your own projects? What about publishing it?
    - How do the projects you've worked on, supported, or encountered differ from the scenario you worked on during the Putting it Together session?
    - What was your biggest "Ah ha!" moment of the day? What do you still find confusing?

Day 3

1. *Report back from moderator on free-write themes*: At the beginning of day 3, the moderator again summarized topics and progress communicated in the previous day's free writes.

2. *Substantive literacies Licensing, Privacy & Ethics:* On day 3, we explored the substantive law and policy literacies for text data mining having to do with licensing, privacy, and

ethics. Participants had watched pre-recorded videos, and synchronous sessions were used for small group discussions.

3. *"Putting it together" exercise*: After the morning substantive sessions, faculty and participants engaged in another "putting it together" exercise. This time, however, the exercise was comprehensive of all literacies—requiring participants to apply not just the day's learnings but also tap into their copyright knowledge from the day before.

4. *Free Write:* Synchronous sessions on day 3 ended with the final 15-minute free write exercise, in which participants reflected on the following prompts:
    - What strategies will you use to evaluate the ethical implications of current and future TDM projects?
    - What licensing issues surfaced for your own work? Where do you see a path forward and where do you feel stuck?
    - What made you feel angry today? What made you feel relieved?

5. *Preparation for Implementation Mapping discussion:* At the conclusion of day 3, we also asked the participants to prepare for day 4 by considering the following questions:
    - How will you provide guidance to others or integrate the literacies in your own practice? What concrete steps or actions will you take? Are there things that you, your institution, or the broader community should *stop* doing?
    - What challenges might you face with implementation of the literacies?
    - How would you like to collaborate with other Building LLTDM participants or other DH researchers / professionals to integrate the literacies into DH TDM practice? What would a high level roadmap look like to achieve this vision? What support or funding would you need to make this vision possible?
    - Are there aspects of the current legal landscape that would benefit from community cooperation and advocacy to better address and enable TDM research?

Day 4

1. *Report back from moderator on free-write themes:* The moderator summarized the free write motifs and lessons.

2. *Implementation mapping:* Faculty and participants convened in small groups to discuss their prepared thoughts on implementation mapping questions. Each group worked to identify next steps, needs, and plans for bringing the literacies to life in their work and at their institutions. We reconvened in a final plenary session to share plans and take-aways from the small group discussion, using the Mural tool to exchange virtual "sticky notes" viewable by all participants. Participants also had an opportunity to post "gratitude" messages to acknowledge or thank other participants, faculty, or recognize a particularly useful or impactful aspect of the institute.

3. *Participant Evaluation:* With impressions and lessons still fresh in their minds, participants completed an evaluation survey prior to attending a final optional "happy half hour."

Post-Institute Meeting

To model the "testing" phase of design thinking, we organized a 1.5-hour check-in meeting eight months after the institute. Our goal was to help the cohort reflect upon their implementation experiences so they could evaluate whether their strategies had been successful.

Approximately two months before the check-in meeting, we re-oriented the cohort to the literacies through a post-institute survey that inquired about their implementation plans and desires for follow-up programmatic resources.

One month before the check-in meeting, we asked participants to share brief (2-minute) videos documenting how they had been supporting TDM legal literacies in their home institutions and projects. We offered the following prompts:

- What have you been thinking about or doing with respect to TDM?
- What's one lasting LLTDM lesson you remember from the Institute?
- What takeaways from the Institute have you been able to implement or share with others?
- What are you still struggling with when it comes to LLTDM?
- What are you proud of with respect to your LLTDM skills?

When we convened for the plenary meeting in February 2021, we began again with the moderator's reflections on themes evidenced in participants' videos. We then transitioned to small group discussions focused on successes, frustrations, or opportunities that the cohort had experienced in implementing the literacies. We concluded with a plenary group exercise to share individual and collective next steps brainstormed during the smaller discussions.

Open Educational Resource

In order to broadly share the materials developed to deliver the institute, we published an openly licensed ebook (open educational resource, or "OER") under the Creative Commons Public Domain Dedication (CC0). This means that the OER can be accessed, reused, and repurposed without restriction.

The OER serves two key purposes:

- **Substantive Literacies:** The first part of the OER covers all the legal literacies covered during the virtual institute, including copyright (both U.S. and international law), technological protection measures, privacy, and ethical considerations. We hope this content will enable any member of the public to gain similar skills and insights as institute participants.

- **Pedagogy:** In the second part, we focus on pedagogy to help anyone who might want to teach the Building LLTDM literacies to others. It describes in detail how we developed and delivered the 4-day institute, and provides ideas and exemplars for hosting shorter instructional sessions. We also include our reflections on both substance and administration to facilitate effective teaching of Building LLTDM literacies by others.

The OER is published on Pressbooks, a web-based platform used to create and share ebooks and other OERs. The ebook is available in a variety of formats, including a web version and downloadable formats such as PDF and EPUB. We are publicizing it through our project website (www.buildingllltdm.org), the UC Berkeley Library blog, via email lists, and through faculty and participants' professional networks.

Impact, Reflections, & Next Steps

Impact

We analyzed participant evaluations and post-institute update videos and survey responses. We observed not only the lasting impact of the LLTDM literacies, but also a persistent sense of shared experience and community.

1. Confidence now abounds

One theme that arose early during the institute was the pervasive feeling of imposter syndrome among participants. It seemed to permeate this work, perhaps because as one participant so rightly observed, no one person can be a deep expert across an entire landscape of issues in text data mining, from corpus building and computation to legal and ethic issues and all of the many technical, intellectual, and labor issues that underpin the work. Yet in post-institute surveys, videos, and discussions, imposter syndrome was absent. Instead, participants commented about how much more confident they felt integrating the literacies into their work. This integration has taken a lot of forms, from licensing negotiations to establishing best practices in their labs. The key struggle transitioned from being unsure of one's skills to finding the time to apply them all.

2. Successful incorporation of ethics into TDM practices

Participants' closing reflections from the institute in June 2020 included a strong desire for taking an ethics-first approach to teaching the literacies and implementing text data mining projects. It has been heartening to see the many ways that participants are living these values by structuring ethics as a key component of their work. For instance:

- One scholar added a dedicated ethics section to a submitted paper involving the use of YouTube data.
- Another centered ethics in application of the literacies to a racial reckoning project at her home institution.
- A librarian has adjusted consultations with researchers to take an ethics-first approach.
- A faculty member has shifted toward an ethics of care framework in working with students in the classroom and in his research lab.

- Several participants developed workshops and related materials that focus on ethical considerations when doing this work.

Participants also turned an eye toward institutional gaps where ethics are concerned. One video update reflected on the lack of oversight of privacy and ethical issues in TDM research, and the need for structures and education that will help with that intervention within our institutions.

Overall, the participants left energized to continue the conversation around ethics and contribute to developing ethics models that might guide TDM researchers in the future.

3. Community education

Across academic institutions, TDM expertise is both shared and distributed. It would be exceedingly rare to find any one person or even any one office prepared to address all of the technical, legal, ethical, and logistical nuances of text data mining. Several participants mentioned that it is difficult to build community due in large part to the dispersed nature of the work. Living and working through a global pandemic has not made that any easier.

Some participants nevertheless made some real gains in community building, and we can celebrate that. One participant described how they initiated conversations across their institution about text data mining to start thinking at an organizational level, and they also noted that they had formed relationships with the sponsored research office and with the faculty working group on data science. Another participant has taken up the idea of the "Data Ombudsperson" and is working to introduce it to the scholarly communication group at their library. Yet another participant has established a new research cluster on "Critical Practice in Text Data Mining" under the auspices of their humanities research center. These kinds of connections hold the potential to make real progress within institutions that are notoriously complex.

4. Struggles with institutional risk aversion

One participant described institutional conservatism and risk aversion as their ongoing struggle. And another had hoped to push their institution to be bolder and braver, but it was not as easy as they had hoped. Seeding institutional change is long durational work and it begins with small acts of relationship building. We reinforced the need to celebrate these gains while striving for much bigger shifts in practice and perception.

5. Efforts to improve institutional licensing

Several participants have been working to break up their institution's licensing routines with various approaches to address TDM. One participant has been looking at the possibility of regularly including TDM language in institutional licenses, which is in keeping with the approach taken in the California Digital Library's model license agreement. Another participant started working on licensing terms and setting up contracts with vendors at their institution, and they ultimately preferred the use of a "Fair Use Escape Clause" rather than outlining specific terms for TDM. They discovered that in an attempt to be explicit, the terms that vendors found acceptable were too confining.

Participants also recognized the need to make the negotiated terms visible to researchers. One participant has been taking that on with a database evaluation to outline who is eligible to use each resource, how the data may be used, and what content is available. Even when full licenses are not readily shared with the campus community, this kind of matrix can help users assess their options when working with content licensed through the libraries.

6. ***Development of workshops***

Another way participants have been working with local communities is by integrating the literacies into their workshops and courses. One participant conducted an hour-and-a-half workshop and shared materials online. Two other participants collaborated on a workshop foregrounding privacy and ethics in DH projects, which is also available online. And yet another participant has put together a suite of relevant workshops associated with their research cluster.

One participant observed that the mere mention of copyright to students can lead to a lot of fear, uncertainty, and doubt, even when the intention is to empower people to understand their rights. It would be helpful to discuss potential strategies for mitigating that effect as part of our ongoing conversations with participants and the research community.

Pedagogical Reflections

The conversations during the institute and the participant feedback gave us much food for thought. We would like to expand our commitment to diversity by ensuring that the demographics of future faculty are as representative as those of the participants, and that the questions and examples that animate discussion sessions themselves engage with issues of ethics, equity, and representation.

We also learned a few specific things that may shape how we approach immersive LLTDM trainings in the future:

1. ***Design Thinking is effective for teaching LLTDM***

The institute empowered participants to understand the basic contours of the legal literacies for text data mining and apply them to their own work, whether that be developing their own TDM projects, advising DH researchers, or working with TDM issues in libraries and archives. The participants' own words from institute evaluations affirm the pedagogical efficacy:

- "I can say with confidence that I understand the four literacies better"
- "I really feel that I am coming out with much more both theoretical and practical knowledge than I expected."
- "I will be much more intentional at the outset of any TDM project about working through all of the pertinent literacies in a systematic way...the way the institute was structured into different literacies provides a repeatable framework to treat potential problems prospectively."
- "I am taking home a lot of new insights from this institute in combination with a feeling of empowerment that will allow me to reach out to the specialists and directors at my

institutions in order to push for more TDM collaboration and a bolder approach concerning materials and datasets for international cooperation. I know now what the important legal issues are and how to use them to form my arguments and that is more than I could have wished for. Also, the institute broadened my perspective with regards to issues that I did not have on the radar that much at the beginning and I am looking forward to engaging with these topics in the future, to integrate them into my teaching, and to advocate for them where I can."

Design thinking can also work in virtual instructive environments, as the pivot from an in-person institute to a virtual one was met with applause. In particular, the participants valued the interactive format with different touch points and small group discussions. Again, in their own words:

- "The deliberately thought through breakdown and mix fostered incredibly valuable discussions and I would hope this kind of framework is used as a best practice for future DH institutes of all kinds going forward. Also, thank you for such an amazing virtual experience which I can only imagine took a tremendous amount of work to coordinate and plan with limited time to shift to an entirely different format--I was overjoyed to critically engage with complex subjects and for the chance to get out of my everyday pandemic routines."
- "I found this to be the best example of how to manage hands-on learning in a virtual environment. I think the planning team did a fantastic job pivoting to a fully online environment without losing the feel of an in-person intensive."
- "The multi-modal communication (Slack, Mural, Zoom) enabled far more interaction than I anticipated."
- "This is by far the best organized event that I have ever attended. The content was by far the most substantive. The faculty were by far the most engaged. A+ across the board."
- "The flipped learning approach, combined with design learning elements, really worked well. The lecture/video materials and reading in particular were well presented and selected, and I really appreciated that we could do that at our own pace. The overall topic of this gathering was well chosen in that it could allow for us to do focused seeking of answers to questions but in a way that had real practical consequences for how we could change the world of TDM research.

2. Copyright is a straightforward literacy to teach

Questions about using material under copyright were at the forefront of participants' minds when they entered the institute, but those concerns evaporated quickly. The copyright portion of the curriculum addressed copyright and the fair use exception extensively, and its applicability to TDM work was solidified. Unexpectedly to many, copyright risk issues turned out to be relatively straightforward and largely confined only to corpus republishing. As a result, participants felt empowered to perform analyses on copyrighted materials. One participant said, "I also feel compelled now to do my own research and take advantage of the expansive idea of fair use to examine contemporary, creative works," and another "was mainly relieved that my TDM project was transformative enough to not violate copyright." The greater challenge the cohort

recognized was finding ways to educate our communities about the full scope of what fair use allows for TDM.

3. Literacies should be woven into research project plans

As scholars and educators, we should be building a legal literacies workflow into DH project planning from the very beginning, and refer to it throughout the project lifecycle. Too often, copyright and other legal considerations are unexamined or brushed aside to the detriment of DH research, partly due to lack of confidence in these areas or fear of institutional or rightsholder reprisal. Institute participants suggested ways of instantiating a lifecycle approach to literacy integration into DH project planning—including intermittent trainings, online guidance about process and sample documentation templates, and building legal questions into the project management process for DH support work. One participant said, "In our library's center for digital scholarship, we need to develop a better charter/MOU/agreement system for digital projects that will at least touch on data management (DMPs), legal implications (copyright, etc), collaborator expectations, and ethics."

4. Institutions need support for adopting TDM-friendly licenses

Licenses with publishers, vendors, museums, and other content providers can further restrict uses that would otherwise be allowed under copyright law. While licensing restrictions can be frustrating when their terms impede the assembly of corpora or application of automated corpora analysis, participants learned what a TDM-friendly license might look like, such as one with terms that specifically allow for TDM uses or that contain a fair use clause. Participants were interested in shaping their institutional licenses—but desired additional instructional materials focused specifically on advocacy and negotiation support.

5. Ethics should be front-and-center

While participants entered the institute focused on questions of copyright, many reported leaving with their copyright questions solved and their ethical questions awakened. As one participant wrote, the institute "erased my anxieties in target areas and introduced whole new considerations in areas like ethics. It answered my questions and left me thinking." We believe questions of ethics loomed large not only because of the critical importance of ethics when addressing data at scale, but also because of the relative absence of guidelines and best practices to help guide us in this area.

We quickly realized that although we discussed ethics as the final substantive literacy during the institute, it was difficult for participants to even begin thinking about copyright, licensing, and other legal issues before ethical considerations were addressed, especially given the institute's care for questions of social justice. As we repurposed the institute training and materials into the OER, we considered additional ways to emphasize and create discussions around ethics, and perhaps foreground ethics as the first step when thinking through DH projects, and in teaching Building LLTDM.

Next Steps

Overall, we are encouraged that the literacies and methodology developed and shared by the institute has empowered DH researchers to build and analyze their text corpora without fear, thanks to their being more secure in their knowledge of the law and ethics. We hope these literacies become rooted more broadly in DH curricula.

In the meantime, we have been considering two specific future courses of action: (1) development of cross-border training, and (2) creation of documentation templates.

Cross-Border Issues Need Future Institutes

Cross-border research collaborations emerged as a clear example of follow-on training that we believe is necessary. Although we had initially intended to focus mainly on U.S. law for most literacies, cross-border and foreign law issues pervaded given the broad range of humanities research in which our cohort engaged: Scholars are working with materials published under different legal frameworks, or are collaborating with others working in those environments. This obviously complicates the legal landscape. Rather than offering clear answers to every question participants raised in the context of cross-border inquiries, we offered strategies for assessing and mitigating risk. Yet, the need for expanding or extending Building LLTDM to international and cross-border contexts is clear.

Need for Documentation Templates

While watching participants' update videos, we also observed their clever use of forms and documentation as tools to help kick start conversations that can ultimately shape practice. One participant described developing an MOU template for use in the digital scholarship lab that includes a section on the legal and ethical implications of the work. The template helps foreground these issues during the negotiation and ensures that they are addressed in the final agreement.

In a similar vein, another participant has been developing a rubric for designing new digital projects that incorporates the literacies and is grounded in the insight that it is best to begin by planning for the end. This presumably helps front-load conversations not just about data collection and corpus building but also representation and distribution for publication and long term preservation. To socialize these practices with graduate students, another participant has started requiring a data management plan for student research projects conducted as part of his research lab to ensure everyone in the lab is thinking deeply about ethics in data collection, dehydration, and eventual destruction for social media research. This approach simultaneously generates deep and thoughtful conversations while also making them expected and routine.

A comprehensive guide or set of customizable templates to document project development choices relative to the literacies is a sound direction for follow-on work.

Appendices

1. Building LLTDM Open Educational Resource
2. Participant Packet
3. Institute Videos, Slides, Transcripts
4. Reading List