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

Research Scenario

Researcher A and Researcher B are scholars in the United States at two different universities, and are planning a "cross-border" text data mining (TDM) project. They intend to analyze the public versus private discourse that occurred in the country "Floria"[1] concerning the 2020 U.S. presidential election. They will do so by performing sentiment analysis on social media group chats, and then compare this analysis to similar sentiment analysis of press coverage.

Methodologically, they will scrape or download Facebook posts made in Floria from 2018-2021, and then run algorithms on the corpus of Facebook posts. They will perform the same process of downloading and analyzing Florish-based newspapers published digitally during the same time period. The Florish newspapers come from three sources: some are licensed through Researcher A's university, some via databases licensed by Researcher B's university, and others are available to the public online.

Researchers A and B are also considering collaborating with Researcher C, a scholar at Floria University, in Europe. In Floria:

- European Union exceptions to copyright law apply, including for research uses and TDM;
- The Florish national copyright law's relevant exception contains a "personal" or "private" use restriction, which has been interpreted to mean that researchers may not reproduce and distribute copyright-protected content to other researchers;
- There is no copyright exception to break technological protection measures;
- Contracts may not override rights and exceptions granted under copyright law; and
- Privacy laws are more stringent than in the U.S. and have extraterritorial reach.

If Researcher C joins the team, Researcher C would:

1. collect and analyze additional posts identified through Florish Facebook groups;
2. provide access to and analyze contents from digital newspaper databases licensed through Researcher C's university; and
3. work with the entire corpus to contribute to algorithmic analysis.

If the research project outcomes prove insightful, Researchers A and B wish to replicate the entire project focusing instead on a geographical region impacted by government-regulated or government-controlled speech and media, such as China or Myanmar.

Researchers A and B have come to their respective university libraries with questions about copyright, licenses, privacy, and ethics in planning their cross-border TDM research project. They would like these questions answered both as to the actions they can undertake as part of the research, and what content or outputs they can ultimately share with other researchers or disseminate to the public from the corpora they compile.

[1] "Floria" is a fictitious country situated in Europe, developed for illustrative purposes in this hypothetical.

Paradigm 1: U.S.-based researchers perform all TDM acts in the U.S.

For Researcher A and B performing all text data mining research and publication in the U.S., and with no engagement of Researcher C, the following "cross-border" variables may influence research design or be useful considerations as part of research guidance:

1. Copyright Variables

a. Foreign-created materials

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)[2] to acts like reproduction, distribution, display, etc.—i.e. all "exclusive rights" that copyright owners have in copyright protected works[3]—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.

b. Publication status

Researcher question(s): Does the publication status (i.e. published vs. unpublished)[4] of the materials in the foreign country affect the U.S. infringement analysis?

Preliminary guidance: Probably not. U.S. copyright law applies to the research activities performed in the U.S. This means that the four factors of the U.S. fair use analysis should apply. If the materials are unpublished in Floria, this arguably could affect U.S. fair use determinations under Factor 2 (which gives preference to published works because an author has the right to control the first public appearance of their expression), but likely with de minimis impact on the fair use balancing test overall.

---

[2] *17 U.S. Code § 107 - Limitations on exclusive rights: Fair use*. (n.d.). LII / Legal Information Institute. Retrieved August 18, 2023, from https://www.law.cornell.edu/uscode/text/17/107
[3] *17 U.S. Code § 106 - Exclusive rights in copyrighted works.* (n.d.). LII / Legal Information Institute. Retrieved August 18, 2023, from https://www.law.cornell.edu/uscode/text/17/106
[4] For a discussion of what "publication" means, see U.S. Copyright Office, *Circular 1*, available at https://www.copyright.gov/circs/circ01.pdf#page=7. For a discussion of how to determine publication status (i.e. published vs. unpublished), see U.S. Copyright Office, *Compendium*, Publication, available at https://www.copyright.gov/comp3/chap1900/ch1900-publication.pdf. And for a discussion of the overall impact of publication status on the fair use analysis, see, e.g., the second fair use factor "The Nature of the Copyrighted Work; https://fairuse.stanford.edu/overview/fair-use/four-factors/.

c. Presence of technological protection measures

Researcher question(s): If the Florish newspaper databases licensed through Researcher A and B's institutions impose technological protection measures (TPMs), can Researchers A and B circumvent these TPMs in the U.S. even if breaking TPM would otherwise be prohibited for text data mining in the newspapers' country of origin?

Preliminary guidance: TPMs can be circumvented in the U.S. on motion pictures and literary works (which includes periodicals) regardless of where the underlying materials were created.[5] However, the database license agreements signed by each institution may contain provisions that override this right under copyright law. While there is dispute as to the efficacy of copyright override provisions, it is generally held that the United States does not prohibit contractual override of fair uses.

d. Geographic limitations on data or corpus sharing

Researcher question(s): What can the researchers share / distribute of their analysis or the corpus, including in the U.S., international publications, and with international colleagues?

Preliminary guidance: Because U.S. law applies to the infringement analysis, researchers can share only to the extent that doing so would be considered fair use. This means they would generally be able to share outputs like: metadata, frequencies / word cloud, nGrams, Word2Vec, XML mark-ups, machine learning models, and other extractions and annotations. But they may exceed the limits of fair use if they instead wish to distribute the corpus, itself—particularly to researchers beyond their research team, and without appropriate protections in place controlling downstream access or use. In addition, the database license agreements signed by Researcher A & B's institutions likely also address or restrict sharing content, and may also restrict sharing excerpts or extractions from that content. The impact of these agreements is addressed separately in Paradigm 1, Section 2(c) below.

e. Known foreign infringement; subsequent U.S. reliance / use

Researcher question(s): If the Researchers A and B rely on a "shadow" library or corpus of materials known to be unlawfully made available in another country (e.g. SciHub), do they undermine their own ability to rely on fair use by conducting TDM on these materials within the U.S.?

Preliminary guidance: If someone in a foreign country has unlawfully reproduced, scraped, or "liberated" (e.g. cracked TPM) on an underlying work or set of works,

[5] 37 CFR § 201.40—Exemptions to prohibition against circumvention. (n.d.). LII / Legal Information Institute. Retrieved August 18, 2023, from https://www.law.cornell.edu/cfr/text/37/201.40. Note, however, that this exemption excludes "compilations that were compiled specifically for text and data mining purposes," such as literary products that were designed specifically to facilitate TDM within the product.

arguably this could impact the "fairness" of the use for the U.S. researchers making use of the materials in the U.S.; however, there is not much existing case law to be determinative on this point either way. Some scholars have argued that a known foreign infringement does not preclude a finding of fair use.[6] In all events, U.S. researchers should not encourage researchers in foreign countries to violate the law as this arguably could be inducement to infringement even in the U.S.

f. Risk of foreign lawsuit for copyright infringement

Researcher question(s): Can the U.S. researchers be sued in the foreign jurisdiction (e.g. Floria) if their TDM research or publication violates foreign copyright law, even if all of their research functions are performed in the U.S.?

Preliminary guidance: It is possible that a foreign court could attempt to assert jurisdiction over a researcher for acts performed in the U.S. that somehow "invoke" foreign copyright law (e.g. if the underlying works being allegedly unlawfully reproduced or distributed in the U.S. are still protected under a foreign country's copyright law, as in The Matter of Fischer). *However*, if a U.S.-based researcher performs all complained-of acts in the U.S., then that researcher can likely defeat the foreign court's assertion of jurisdiction should the researcher raise this issue as a defense. Retaining counsel can be vital to preserving and protecting one's rights in this regard.

If for some reason the foreign court is found to have proper jurisdiction *and* ultimately enters a judgment against the U.S.-based researcher, that still does not mean that the foreign judgment will actually be enforced against the researcher. The foreign copyright owner would next need to come to a U.S. court and ask that a U.S. court enforce this foreign judgment. As further research would demonstrate, it is even less likely that a U.S. court would agree to enforce a foreign judgment when the underlying acts were both performed *and* otherwise permitted in the U.S.

For instance, in California, the "California Recognition Act" allows a California court to decline to recognize a foreign-country money judgment if the "judgment or the cause of action or claim for relief on which the judgment is based is repugnant to the public policy of [California] or of the United States." Cal. Civ. Proc. Code § 1716(c)(3). However, the bar for satisfying "repugnancy" (and thus declining to enforce the foreign judgment) is very high. *Ohno v. Yasuma*, 723 F.3d 984 (9th Cir. 2013). As explained most relevantly in *De Fontbrune v. Wofsy*, 39 F.4th 1214 (9th Cir. 2022), "The issue is not simply whether the 'foreign judgment or cause of action

[6] *See*, *e.g.,* Carroll, M.W., *Copyright and the Progress of Science: Why Text and Data Mining Is Lawful*, 53 UC Davis Law Review 893 (2019). https://lawreview.law.ucdavis.edu/issues/53/2/articles/files/53-2_Carroll.pdf. Further, it may also be difficult to assess whether a given corpus was infringing when created because of jurisdictional variations; for instance, some countries have far more liberal limitations and exceptions in general or for certain classes of uses (e.g., Japan for TDM, India for some educational uses), and easy to make incorrect presumptions about the lawfulness of a given corpus.

is contrary to our public policy,' Rather, the question is whether either is 'so offensive to our public policy as to be prejudicial to recognized standards of morality and to the general interests of the citizens.' [citations omitted]." Under these standards, in *De Fontbrune v. Wofsy*, the Ninth Circuit ruled that the disputed act did not constitute a fair use to begin with, and thus the California court could indeed enforce the foreign judgment because the act would not have been permitted even in the United States. However, the court declined to rule on whether the foreign judgment would have similarly been enforced or rejected as repugnant if the disputed act had instead been considered a fair use in the United States. 39 F.4th at 1223.

Thus, it remains unclear whether text and data mining (which is considered a fair use in the United States) could ever be a proper subject for enforcement of judgment from a foreign jurisdiction that does not recognize fair use. *De Fontbrune* may be further distinguished because the disputed acts (sales of infringing works) took place abroad, whereas in the hypothetical described here, all disputed research acts arguably take place within the United States—thus further limiting the propriety of any foreign court's jurisdiction over Researchers A & B from the start.

2. Contractual Variables

a. Cross-institutional data or corpus sharing

Researcher question(s): Can researchers from different U.S. institutions share access to the foreign newspapers with each other?

Preliminary guidance: The terms of each institution's license agreement will dictate what materials (including even derived data from those materials), if any, can be shared with collaborators at other institutions.[7] The country of origin of the underlying content is immaterial to that question; what matters instead is what the contract says about the sharing of that content.

Subject to the caveat that there is some marginal possibility that U.S. contracts cannot properly circumscribe fair use: If the license agreements do not permit content sharing between the researchers, then each researcher may need to either (i) be responsible for mining only that content to which their own institution has access, or (ii) renegotiate terms with the publishers / vendors to permit multi-institution collaborations.

[7] For more on how institutional license agreements can "bind" or obligate researchers to comply, see the Licensing chapter of Building LLTDM, https://berkeley.pressbooks.pub/buildinglltdm/chapter/licensing/. Generally speaking, researchers may be bound by license agreements through several mechanisms: (1) directly by the agreements if they have signed them personally or for/by their research team; (2) directly by being presented with a "Terms of Use," "click-through," or other end user license agreement, or (3) indirectly by being third party beneficiaries or otherwise falling within contractual "privity" through their university. In this last scenario, however, there is some possibility that only the university, and not the individual researchers, are potentially liable for the acts of the researchers, and often universities negotiate to have their license agreements disclaim liability for the acts of their users.

b. Country of origin's impact on license agreement or website Terms of Service

Researcher question(s): Does either the original publication of the Facebook posts via Facebook users in Floria, or the original publication of newspaper articles online in Floria, result in a foreign country's Terms of Service (also sometimes called "Terms of Use") and foreign laws being applied to the subsequent U.S. mining / scraping of the posts and online articles? If the foreign country's Terms of Service prohibit acts like research uses, scraping, or downloading, are they enforceable against research activities undertaken in the U.S.?

Preliminary guidance: Potentially, but unlikely. Facebook Terms of Service (or other terms of service for Internet content like the online Florish newspapers) typically dictate their territorial reach. They will also likely address which jurisdiction's law applies to their interpretation and enforcement (see also Researcher question(s) Paradigm 1, Section 2(d) below). The most important terms of service for researchers to be aware of are the ones they agree to / assent to by the use of the content at issue.

However, this does not entirely preclude a public policy argument from being made that a U.S. court should account for foreign terms of service because the authors of the posts or news articles relied on those terms in deciding to share their materials online. For instance, if a Florish Facebook poster was informed through the Florish Facebook Terms of Service that no research uses could be made of the content they were about to post, then they had a reasonable expectation that no such research uses would be made. This might mean that even if Facebook U.S. Terms of Service do permit research uses, a U.S. court *might* enforce the version or terms that the posting author relied on deciding to post to begin with. Further research could confirm the likelihood of such an argument succeeding. And guidance regarding overall risk would be needed to help researchers understand any application of foreign law in this context.

c. Contractual impact on data or corpus sharing / republication

Researcher question(s): What can the researchers share or republish of their findings or the corpus, either in the U.S. or abroad?

Preliminary guidance: Both the newspaper license agreements and the Facebook Terms of Service will likely govern what or how much of the underlying content may be shared or republished. Findings / results / analyses that do not distribute or disseminate copyright-protected content should raise no issues (e.g. frequencies / word cloud, nGrams, Word2Vec, XML mark-ups, machine learning models, and other extractions and annotations).

d. Risk of foreign lawsuit for contractual breach

Researcher question(s): Can the U.S. researchers be sued in the foreign jurisdiction if the TDM research or publication violates a license agreement or Terms of Service, even if all research functions are performed in the U.S.?

Preliminary guidance: Potentially, though research is needed into the likelihood. The license agreement or Terms of Service will likely suggest which country's law applies. This could mean that foreign law is applicable and jurisdiction is bestowed to a foreign court. However, it is not guaranteed that a U.S. court would enforce a foreign judgment, particularly if the underlying acts were performed or permitted in the U.S.

3. Privacy & Ethics Variables

a. Applicability of foreign privacy laws to U.S.-based TDM

Researcher question(s): Do the privacy laws of the foreign country apply to text data mining research collection, analysis, and dissemination where the content is authored by or pertains to individuals in a foreign country, but the TDM research is performed in the U.S.?

Preliminary guidance: Depending on the jurisdiction, foreign privacy laws (e.g. General Data Protection Regulation[8], China's Personal Impact Protection Law[9]) may have an extra-territorial effect. Many of these laws govern only the acts of certain types of large data aggregators or providers, meaning that as a practical matter, most of the impact and risk of privacy law violations of such laws would likely be borne by the entity *providing* or *licensing* already public data, rather than the institution or researcher using or independently collecting it. Nevertheless, research into applicable countries' laws and consultation with one's Institutional Review Board is advised.

b. Use of data beyond original intent

Researcher question(s): If the content was created abroad and originally intended for a limited foreign audience or a certain purpose, what considerations should be made when utilizing it for TDM beyond that original purpose or geographic boundary? Should foreign content creators have the right to determine if or how their content will be used in U.S.-based TDM?

[8] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). (n.d.). Retrieved August 18, 2023, from https://eur-lex.europa.eu/eli/reg/2016/679/oj
[9] Personal Information Protection Law of the People's Republic of China. (2023). In Wikipedia. https://en.wikipedia.org/w/index.php?title=Personal_Information_Protection_Law_of_the_People%27s_Republic_of_China&oldid=1156099013

Preliminary guidance: From a legal perspective, contracts (like Facebook Terms of Service) explain that posts may be used for purposes like research. Arguably, this puts content creators on notice that their posts may be used in a manner beyond what they might have initially expected. But it is also known that users do not necessarily read or understand the Terms of Service, and their data could be shared in ways that the creators of those materials were never able to imagine. There are no cross-border ethical requirements as a matter of law, but researchers might decide that there are certain situations in which they want to take a more nuanced ethical position in their use of such content. There are a variety of sources (general and discipline-specific) to which they may look for guidance. Many are cataloged in the Ethics chapter of *Building Legal Literacies for Text Data Mining*.

c. Sensitive but not legally private data

Researcher question(s): What are the implications of using research data and findings that are not technically private (either under foreign law or in the U.S.), but that might be sensitive? How should researchers address data usage relative to the political or social regimes operating within the country in which the data originated?

Preliminary guidance: TDM researchers may want to evaluate how or whether to publish data or analysis if ethical norms either in the content's origin country *or* in the U.S. suggest that publication could lead to exploitation of people, resources, or knowledge. For example, researchers may determine not to publish materials if there is substantial concern that content is culturally treated as "confidential" or "traditional knowledge" in a different geographic region, or if it threatens the safety of individuals who could be punished for having spoken out against a political regime.

d. Risk of foreign lawsuit for privacy violations

Researcher question(s): Can the U.S. researchers be sued in the foreign jurisdiction if the TDM research or publication violates foreign privacy laws, even if all research functions are performed in the U.S.?

Preliminary guidance: Potentially, though research is needed into the likelihood. The foreign privacy statutes will dictate both their extraterritorial applicability and the place of suit for violations. Research is needed to evaluate the extent to which a U.S. court would subsequently enforce a foreign judgment when the underlying acts were both performed and permitted in the U.S.

4. Risk assessment

Researcher question(s): What other risks might be posed other than risks of lawsuits?

Preliminary guidance: As a preliminary matter, lawsuits may impose either injunctions (i.e. orders to stop behavior), or damages (i.e. monetary sanctions), or

both. Researchers may perceive the threat of injunctions to be less “risky” than the potential for damages. Unfortunately, laws relating to damages and injunctions, and their availability and scope, vary by state and country—making universal guidance difficult.

In addition, there are other types of risks that arise in cross-border TDM research:

- *Risks to researchers*: There could be reputational harms associated with violating agreements or knowingly infringing. Some publishers may also refuse publication or retract papers when violations come to light.
- *Risks to institutions*: Institutions could face litigation costs and loss of access to key resources (e.g. if access for the campus is terminated as a result of an individual’s violation of a license agreement)
- *Risks to subjects / third partie*s: Rights holders, vulnerable or marginalized communities, and data subjects may face varying types and degrees of harm (e.g. danger, shame, ridicule) if their expectations of privacy or obscurity are breached or exceeded.

Paradigm 2: U.S.-based researchers engage with collaborator abroad, or otherwise perform TDM acts in both U.S. and abroad

For Researcher A and B performing some text data mining research and publication in the U.S., but with engagement of Researcher C and some TDM acts or corpus-sharing across borders, the following "cross-border" variables may influence research design or be useful considerations as part of research guidance:

1. Copyright Variables

a. Foreign exercise of protected rights

Researcher question(s): (i) If the international collaborator (Researcher C) undertakes the reproduction, distribution, or other copyright-protected act abroad, which country's infringement analysis applies to the activities conducted in that foreign country? (ii) If Researcher C infringes in the foreign country, can Researchers A and B be liable for contributory infringement, or whatever equivalent culpability theory may exist in that foreign country?

Preliminary guidance:

(i) The law of the country in which the research work is performed should govern the copyright infringement analysis for acts performed in that country. Stated another way: A court's inquiry into whether something constitutes infringement should be decided by applying the law of the country in which the copyright-protected acts were performed.

All countries have implemented copyright exceptions to support activities like scientific or scholarly research. Some of these exceptions—like fair use in the United States—may authorize TDM research. However, approximately only one fifth of countries' research exceptions are broad enough to permit the full range of TDM research, which requires the ability to copy, share, and analyze whole works in collaboration with others.[10] As explained by Flynn et al. (2022),[11] "some countries have research exceptions that permit uses only of excerpts of a work (e.g., Argentina), do not apply to uses of books or other kinds of works (e.g., most post-Soviet countries), or require membership in a specific research institute (e.g., Sweden)."

[10] Flynn, S., Schirru, L., Palmedo, M., & Izquierdo, A. (2022). Research Exceptions in Comparative Copyright. *Joint PIJIP/TLS Research Paper Series*. https://digitalcommons.wcl.american.edu/research/75
[11] Fiil-Flynn, S. M., Butler, B., Carroll, M., Cohen-Sasson, O., Craig, C., Guibault, L., Jaszi, P., Jütte, B. J., Katz, A., Quintais, J. P., Margoni, T., de Souza, A. R., Sag, M., Samberg, R., Schirru, L., Senftleben, M., Tur-Sinai, O., & Contreras, J. L. (2022). Legal reform to enhance global text and data mining research. *Science*, *378*(6623), 951–953. https://doi.org/10.1126/science.add6124

Here, if the digitization and downloading (both of which constitute "reproduction") of the newspapers and Facebook posts are performed by Researcher C in Floria (governed by European Union law), then these acts are first governed by the Directive on Copyright in the Digital Single Market (DCDSM)[12] which generally supports the research and TDM uses being described here, provided that among other things there is no dissemination of the underlying corpus publicly.

But applying the DCDSM is not the end of the inquiry. A directive like the DCDSM "is a legislative act that sets out a goal that all EU countries must achieve. However, it is up to the individual countries to devise their own laws on how to reach these goals."[13] And although the DCDSM imposes minimum requirements, "national laws still have a margin of discretion on how they implement the different elements of the legal regime, especially at this early stage of implementation, before the Court of Justice of the EU steps in."[14]

As such, the next step of the inquiry is to apply the national law of the country. In this case, the copyright law of Floria limits copyright exceptions to a "personal" or "private" right or use, which has been interpreted in similar jurisdictions to mean that the Florian researcher would be restricted from reproducing and distributing the copyright-protected corpus to research colleagues like Researchers A and B.[15] This result might discourage a U.S.-based researcher from partnering with a Florish colleague for TDM due to the distribution restriction, and incentivize the U.S. researcher to partner instead with a scholar from England or Germany, which have open research exceptions that would allow the desired corpus sharing within the research group.[16]

Overall, the implications of these differences in national copyright laws and exceptions may exacerbate bias in the nature of research questions being studied (e.g. perhaps leaving research questions affecting countries like Floria underexplored relative to those affecting countries like England) or the types of materials being used to study them (e.g. perhaps favoring use of public domain works not protected by copyright).[17]

---

[12] Directive 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC. (n.d.). Retrieved August 18, 2023, from https://eur-lex.europa.eu/eli/dir/2019/790/oj

[13] *Types of legislation | European Union*. (n.d.). Retrieved August 29, 2023, from https://european-union.europa.eu/institutions-law-budget/law/types-legislation_en

[14] Written project feedback from João Pedro Quintais, Institute for Information Law, University of Amsterdam. On file with authors.

[15] "The most common of these exceptions extend to research uses as a category of "private" or "personal" use. By virtue of the use of the term "private" or "personal," we assume that none of these exceptions authorizes sharing with other researchers…" Flynn, S., Schirru, L., Palmedo, M., & Izquierdo, A. (2022). Research Exceptions in Comparative Copyright. *Joint PIJIP/TLS Research Paper Series*. p. 26, Table 4. https://digitalcommons.wcl.american.edu/research/75

[16] Ibid, at p. 17, Table 1.

[17] Levendowski, A. (2018). How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem. *Washington Law Review*, *93*(2), 579. https://digitalcommons.law.uw.edu/wlr/vol93/iss2/2/

(ii) While Researcher C's country may assert that Researchers A and B face contributory liability for Researcher C's infringement, the likelihood of Researchers A and B being "hauled into court" in Researcher C's country is possible but not likely (see Researcher question(s) Paradigm 2, Section 1(f) below). This point would benefit from further research.

b. Corpus creation abroad, or in U.S. and abroad; sharing data or corpus across borders

Researcher question(s): If the international collaborator digitizes, downloads, or otherwise compiles a portion of the corpus in Floria and then digitally shares that corpus with the U.S. researchers for the TDM analysis to be performed in the U.S., whose law applies to that infringement analysis? What if all collaborators create portions of a corpus (by reproducing and displaying content), and the corpus is then generally shared (i.e. distributed) within the research team for TDM activities to be conducted wherever collaborators are located?

Preliminary guidance: The law of the country in which the acts are performed should govern the infringement analysis for those acts. Stated another way: A court's inquiry into whether something constitutes infringement should be decided by applying the law of the country in which the copyright-protected acts were performed. If the researchers build a shared digital corpus for use by the entire research team—here, with the corpus consisting of the U.S. researchers' mined content *and* the Florish researcher's mined content—then reproduction, display, and distribution would need to be permitted in both countries in order for the Researcher A and B's acts to be lawful in the U.S., and Researcher C's acts to be lawful in Floria.

Applying this rule to this scenario: U.S copyright law may allow corpus sharing amongst the researchers under fair use principles (though this would be a case-by-case determination), so Researchers A and B may be able to provide "their" portions of the corpus to Researcher C. But while distribution of the corpus by Researcher C to Researchers A and B might fall within the European Union's research exceptions[18], Florish national law appears to restrict distribution to others (as it affords only a "personal" reproduction and distribution exception); thus, Researcher C may not be able to rely on a copyright exception to reproduce and distribute "their" portion of the corpus to Researchers A and B.

[18] See Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society, EP, CONSIL, 167 OJ L (2001). http://data.europa.eu/eli/dir/2001/29/oj/eng and Directive 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC. (n.d.). Retrieved August 18, 2023, from https://eur-lex.europa.eu/eli/dir/2019/790/oj

c. Presence of technological protection measures

Researcher question(s): If the newspaper databases at Researcher C's foreign institution impose technological protection measures (TPMs), can Researcher C provide the encrypted files to the U.S.-based researchers for use in the U.S., even if breaking TPM would be prohibited in Researcher C's country?

Preliminary guidance: Research is needed. It would first need to be generally lawful (i.e. it would need to fall within a copyright exception) for Researcher C to provide / distribute the copyrighted content to Researchers A and B regardless of whether the content is also protected by TPM. Assuming it would generally be lawful to distribute copyright-protected materials for research, then: If it is unlawful to break TPM for TDM research in Researcher C's country, that country's TPM-related statute or provision may further dictate whether it is lawful to "export" that TPM-protected content to be mined in a country in which it *is* lawful to break for TDM. We would anticipate that liability would more likely attach to the researcher in whose country the acts are unlawful, rather than the U.S. recipients who receive and decrypt the content in accordance with U.S. law, unless the U.S. researchers could also be said to have induced the infringement.

d. Known foreign infringement; subsequent U.S. reliance / use

Researcher question(s): If the TDM research activity (e.g. scraping) is not authorized in the foreign country but is permitted in the U.S., can it still be undertaken in the foreign country and "provided" (reproduced and distributed) to the U.S. researchers under the auspices of the enterprise being predominantly based in the U.S.?

Preliminary guidance: If content is unlawfully downloaded, scraped, or reproduced in Floria, under the DCDSM that content is no longer an appropriate subject for the TDM research exemption in the E.U., though national law may vary as to this point. Although some scholars argue that downstream use in the U.S. of unlawfully acquired content abroad may still be a fair use,[19] institutions could not reasonably advise researchers in foreign jurisdictions to intentionally (or in legal parlance, "knowingly") provide unlawfully scraped, downloaded, or liberated content to researchers in countries with more permissive TDM jurisprudence.

e. Place of output publication

Researcher question(s): If the research team publishes their findings or corpus on a website hosted in Floria or within / hosted by a Florish journal, but the content can

[19] See, e.g., Carroll, M.W., *Copyright and the Progress of Science: Why Text and Data Mining Is Lawful*, 53 UC Davis Law Review 893 (2019).
https://lawreview.law.ucdavis.edu/issues/53/2/articles/files/53-2_Carroll.pdf

be accessed online anywhere around the world, which country's law applies to alleged infringement of the distribution right for this content?

Preliminary guidance: The country in which the copyright-protected acts (e.g. reproduction, distribution, display, etc.) are performed should be the country whose copyright law governs those acts for infringement analysis—even if the resulting content or output can be viewed or is distributed globally.

f. Risk of foreign lawsuit for copyright infringement

Researcher question(s): Can the U.S. researchers be sued in the foreign jurisdiction if the TDM research or publication violates foreign copyright law, even if all research functions are performed in the U.S.?

Preliminary guidance: Research is needed. For the reasons set forth in Paradigm 1, Section 1(f), we believe it is unlikely that a foreign court would be able to properly assert jurisdiction, and even less likely that a U.S. court would enforce a foreign judgment when the underlying acts were both performed and permitted in the U.S.

2. Contractual Variables

a. Foreign country prohibits contractual override of copyright exceptions, but override permitted in the U.S.

Researcher question(s): (i) If Researcher C is within a country that prohibits license agreements from overriding copyright exceptions like TDM (i.e. meaning Researcher C's institutional license agreements cannot prohibit research uses authorized in the EU), can Researcher C compile, download, reproduce, or distribute the corpus content for the U.S. researchers to mine, or for all three researchers to mine collectively?

(ii) In reverse, can Researchers A and B compile database content in the U.S. to provide to Researcher C for Researcher C to perform the analysis in Floria, if contractual override is inapplicable in Floria?

Preliminary guidance:

(i) If Researcher C is within a jurisdiction like the EU that prohibits override of certain copyright exceptions (such as the exception that enables TDM by research organizations for the purposes of scientific research), then Researcher C's institutional license agreements for the Florish newspaper databases should not preclude Researcher C from conducting TDM. That said, in this case Florish national law bears certain distinctions from the EU's DCDSM, in that the right of reproduction and distribution is a "private" right (i.e. personal to the researcher). This may mean that while Researcher C's database license agreement cannot override the right for Researcher C to conduct the actual TDM, Researcher C might still be precluded from

distributing the corpus content to Researchers A and B unless a license agreement authorizes it.

(ii) Researchers A and B compiling content in the U.S. are governed by their institutional license agreements for which contractual override of underlying copyright exceptions *is* a possibility. Therefore, whether Researcher C's country prohibits contractual override of TDM rights is irrelevant to whether Researchers A and B can reproduce and distribute content to Researcher C for TDM.

b. U.S. contract preserves / authorizes the protected right, but foreign country's copyright laws prohibit it

Researcher question(s): If Researchers A and B's database license agreements permit them to distribute copyright-protected content with other research collaborators, can Researchers A and B provide that content to Researcher C for TDM even if TDM is not authorized in Researcher C's country?

Preliminary guidance: The place where the research acts are performed matters for infringement analysis. So, Researcher C likely cannot make TDM uses of the content from Researchers A and B, even if Researcher A and B's license agreement had authorized TDM by or sharing content with research colleagues in foreign countries.

c. Risk of foreign lawsuit for contractual breach in U.S.

Researcher question(s): Can the U.S. researchers be sued in the foreign jurisdiction if they perform TDM acts in the U.S. that violate a foreign license agreement or Terms of Service?

Preliminary guidance: Potentially, though research is needed into the likelihood. The license agreement will likely suggest which country's law applies and its territorial reach. This could mean that foreign law is applicable and jurisdiction is bestowed to a foreign court. However, it is not guaranteed that a U.S. court would enforce a foreign judgment, particularly if the underlying acts were performed and permitted in the U.S.

3. Privacy & Ethics Variables

a. Applicability of foreign privacy laws to the U.S. researchers

Researcher question(s): If the data is authored by or pertains to individuals in Researcher C's country, can the research team avoid liability for violating foreign privacy laws if only Researchers A and B in the U.S. perform the acts that would otherwise violate the privacy laws of Researcher C's country?

If the data is protected under the privacy laws of Researcher C's country but was lawfully made publicly-available (either voluntarily through Facebook or the news

databases), would the researchers still or further violate these laws by reusing publicly-available data?

Preliminary guidance: Depending on the jurisdiction, foreign privacy laws (e.g. General Data Protection Regulation,[20] China's Personal Impact Protection Law[21]) may have an extra-territorial effect. The foreign privacy laws should set forth what constitutes a violation and the laws' territorial reach. This means Researchers A and B might indeed violate the laws in the country of Researcher C even if they perform the acts in the U.S. Additionally, Researcher C might violate the privacy laws of the foreign country by procuring the data to provide to Researchers A and B.

The foreign privacy laws should also address whether reuse of publicly-available data constitutes a violation, particularly if or where that reuse is done for a purpose beyond its original disclosure or without obtaining additional consent if needed (see also Paradigm 2, Section 3(b) below). Research into applicable countries' laws and consultation with one's institutional review board is advised.

b. Use of data beyond original intent

Researcher question(s): If the data is protected under the privacy laws of Researcher C's country but was lawfully made publicly-available (either voluntarily through Facebook or the news databases), would the researchers still or further violate these laws by reusing publicly-available data?

Preliminary guidance: The foreign privacy laws should also address whether reuse of publicly-available data constitutes a violation, particularly if or where that reuse is done for a purpose beyond its original disclosure or without obtaining additional consent if needed. Research into applicable countries' laws and consultation with one's institutional review board is advised.

c. Sensitive but not legally private data

Researcher question(s): What are the implications of using research data and findings that are not technically private (either under foreign law or in the U.S.), but that might be sensitive? How should researchers address data usage relative to the political or social regimes operating within the country in which the data originated?

Preliminary guidance: TDM researchers may want to evaluate how or whether to publish data or analysis if ethical norms either in the content's origin country *or* in the U.S. suggest that publication could lead to exploitation of people, resources, or

[20] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). (n.d.). Retrieved August 18, 2023, from https://eur-lex.europa.eu/eli/reg/2016/679/oj
[21] Personal Information Protection Law of the People's Republic of China. (2023). In Wikipedia. https://en.wikipedia.org/w/index.php?title=Personal_Information_Protection_Law_of_the_People%27s_Republic_of_China&oldid=1156099013.

knowledge. For example, researchers may determine not to publish materials if there is substantial concern that content is culturally treated as "confidential" or "traditional knowledge" in a different geographic region, or if it threatens the safety of individuals who could be punished for having spoken out against a political regime.

d. Risk of foreign lawsuit for privacy violations

Researcher question(s): Can the U.S. researchers be sued in the foreign jurisdiction if the TDM research or publication violates foreign privacy laws, even if all research functions are performed in the U.S.?

Preliminary guidance: Potentially, though research is needed into the likelihood. The foreign privacy statutes will dictate both their extraterritorial applicability and the place of suit for violations. Research is needed to evaluate the extent to which a U.S. court would enforce a foreign judgment when the underlying acts were both performed and permitted in the U.S.

4. Risk assessment

Researcher question(s): What other risks might be posed, other than risks of lawsuits?

Preliminary guidance: As a preliminary matter, lawsuits may impose either injunctions (i.e. orders to stop behavior), or damages (i.e. monetary sanctions), or both. Researchers may perceive the threat of injunctions to be less "risky" than the potential for damages. Unfortunately, laws relating to damages and injunctions, and their availability and scope, vary by state and country, making universal guidance difficult.

In addition, there are other types of risks that arise in cross-border TDM research.

- *Risks to researchers*: There could be reputational harms associated with violating agreements or knowingly infringing. Some publishers may also refuse publication or retract papers when violations come to light.
- *Risks to institutions*: Institutions could face litigation costs and loss of access to key resources (e.g. if access for the campus is terminated as a result of an individual's violation of a license agreement), or uphill battles in negotiating future license agreements.
- *Risks to subjects / third partie*s: Rights holders, vulnerable or marginalized communities, and data subjects may face varying types and degrees of harm (e.g. danger, shame, ridicule) if their expectations of privacy or obscurity are breached or exceeded.

Fear of some of these risks may have a distorting and negative effect on research (e.g., by excluding important sources or topics of inquiry even though doing so is not methodologically sound or required). And so, researchers need to accurately assess

them and weigh the potential negative harm with the potential negative harm caused by not using moving forward.