POLICY FORUM

DATA ACCESS

Legal reform to enhance global text and data mining research

Outdated copyright laws around the world hinder research

By **Sean M. Fiil-Flynn[1], Brandon Butler[2], Michael Carroll[1], Or Cohen-Sasson[3], Carys Craig[4], Lucie Guibault[5], Peter Jaszi[1], Bernd Justin Jütte[6], Ariel Katz[7], João Pedro Quintais[8], Thomas Margoni[9], Allan Rocha de Souza[10], Matthew Sag[11], Rachael Samberg[12], Luca Schirru[13], Martin Senftleben[8], Ofer Tur-Sinai[14], Jorge L. Contreras[15,16]**

Researchers engaged in text and data mining (TDM) research collect vast amounts of digitized material and use software to analyze and extract information from it. TDM is a crucial first step to many machine learning, digital humanities, and social science applications, addressing some of the world's greatest scientific and societal challenges, from predicting and tracking COVID-19 to battling hate speech and disinformation (*1*–*3*). Although applications of TDM often occur across borders, with researchers, subjects, and materials in more than one country, a patchwork of copyright laws across jurisdictions limits where and how TDM research can occur. With the World Intellectual Property Organization (WIPO) Standing Committee on Copyright and Related Rights, and legislatures around the world, deliberating the harmonization of copyright exceptions for various research uses, we discuss policy measures that can ensure that TDM research is unambiguously authorized under copyright law.

Most text, images, and other materials that TDM researchers use are subject to copyright law. Copyright law gives the owner of a protected work the legal right to prohibit reproduction, distribution, modification, and other forms of exploitation of that work without the owner's permission. These rights apply even if the material is readily accessible—for example, published on the internet or available in a library.

The justifications for copyright are grounded both in the rights of individual authors in their creations and in instrumentalist incentives for the creation and dissemination of new works. Copyright thus gives the author of a work, or the author's assignee (e.g., a publisher), the exclusive right to reproduce, transmit, and make derivatives of the protected work and to prevent the unauthorized appropriation of these rights (*4*). Although copyright originally subsisted solely in textual works of authorship, today it has expanded to cover graphical and visual works and, in some countries, data and databases [though some countries, including member states of the European Union (EU), have separate statutes protecting databases].

Each stage of a TDM project is potentially constrained by copyright depending on how the scope of protection is interpreted. Copyright prohibitions on unlicensed reproduction may be implicated when sources are digitized, formatted, and compiled into a corpus that can be mined for analysis. Copyright may also limit the application of an algorithm to a TDM corpus, which may make additional temporary copies in computer memory. Copyright restrictions on transmitting and reproducing works may be implicated when researchers collaborate, when examiners validate, and when publishers report results. Thus, without copyright permission, or the application of exceptions under copyright law, much of the world's copyrighted material may be off limits to TDM use.

Some publishers make limited copyright licenses available for TDM uses, often for additional fees charged to libraries or researchers. But paid licensing is not an affordable or viable option for many critical TDM projects. TDM research often requires use of massive datasets with works from many publishers, including copyright owners that cannot be identified or are unwilling to grant licenses. Forcing researchers to use only licensed or public domain content (i.e., content in which there is no enforceable copyright) can restrict topics of study, hamper reproducibility and validation (*5*), bias results (*6*), and dissuade researchers from undertaking projects (*7*). A lack of a license need not, and should not, be an absolute barrier to TDM research.

The rights granted by copyright are not absolute. All international copyright treaties permit, and all countries have, exceptions from copyright protection for various purposes, some of which may authorize TDM research. In the US, for example, a flexible exception exists for "fair use" for purposes such as education and research and has been interpreted by courts to permit at least some TDM uses. Copyright laws in many other countries contain exceptions for research (or "scientific") uses that can be interpreted to apply to TDM uses (*4*). But only about a fifth of these 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 (*8*) (see the figure and table). For example, 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).

Empirical studies show that copyright exceptions for research matter—with correlations between more permissive research exceptions and higher production of citable works of scholarship (*9*) and increased academic use of TDM methodologies (*10*). But until legal enabling environments for TDM research can be harmonized, the full benefits from this new research frontier will remain inadequately explored.

LEGALLY ENABLING TDM RESEARCH

Ideally for researchers, a minimum standard for global uses of TDM would be implemented everywhere. There are a number of avenues that policy-makers can take to promote more harmonization of copyright exceptions for TDM uses.

[1]Program on Information Justice and Intellectual Property, Project on the Right to Research and International Copyright, American University Washington College of Law, Washington, DC, USA. [2]University of Virginia Library, Charlottesville, VA, USA. [3]Zvi Meitar Center for Advanced Legal Studies, Faculty of Law, Tel Aviv University, Tel Aviv, Israel. [4]Osgoode Hall Law School, York University, Toronto, Canada. [5]Dalhousie University, Schulich School of Law, Law and Technology Institute, Halifax, Canada. [6]Sutherland School of Law, University College Dublin, Belfield, Ireland. [7]Faculty of Law, University of Toronto, Toronto, Canada. [8]University of Amsterdam, Institute for Information Law (IViR), Amsterdam, Netherlands. [9]Centre for IT & IP Law (CiTiP), Faculty of Law, University of Leuven (KUL), Leuven, Belgium. [10]Federal University of Rio de Janeiro (UFRJ), Rio de Janeiro, Brazil. [11]Emory University, School of Law, Atlanta, GA, USA. [12]University of California, Berkeley, Berkeley, CA, USA. [13]Pontifical Catholic University of Rio de Janeiro, Rio de Janeiro, Brazil. [14]Faculty of Law, Ono Academic College, Kiryat Ono, Israel. [15]University of Utah S. J. Quinney College of Law, Salt Lake City, UT, USA. [16]Department of Human Genetics, University of Utah School of Medicine, Salt Lake City, UT, USA. Email: sflynn@wcl.american.edu

International treaties

Nearly all countries provide a high baseline level of copyright protection as the result of several widely adopted multilateral treaties, beginning with the Berne Convention for the Protection of Literary and Artistic Works (Berne, 1886) and continuing through the WIPO "Internet Treaties" (Geneva, 1996). A key feature of these treaties is a requirement that signatory countries impose high standards for protecting copyright but leave exceptions, such as those permitting research, largely to the discretion of national legislatures and courts. The result is the fragmented landscape of exceptions shown in the figure. But the tide is turning. The WIPO's Standing Committee on Copyright and Related Rights is now deliberating over the harmonization of exceptions for uses that include research. Coalitions of researchers and academics are proposing that this forum draft a treaty that would permit cross-border and other uses of research materials to permit TDM everywhere (*11*).

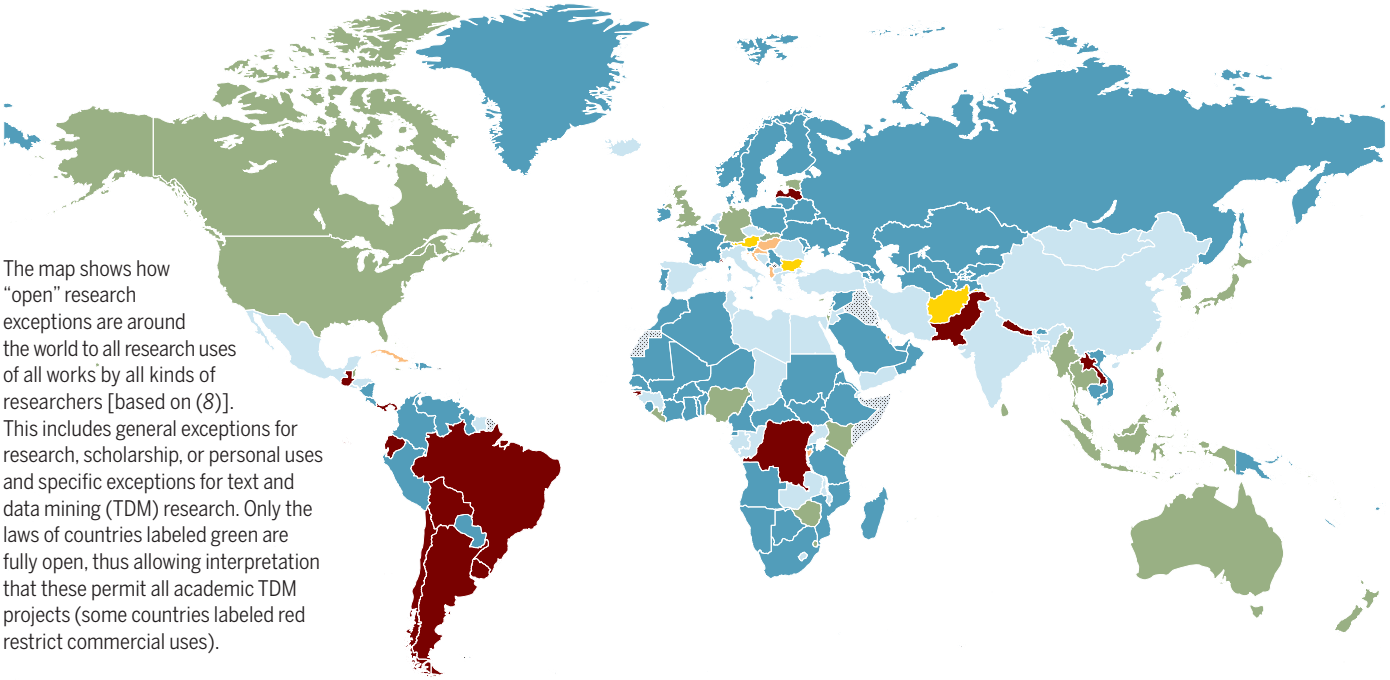
There are important precedents for an international treaty that imposes uniform copyright exceptions around the world. For example, WIPO's last major treaty, the Treaty to Facilitate Access to Published Works for Persons Who Are Blind, Visually Impaired or Otherwise Print Disabled (Marrakesh, 2013), harmonized copyright exceptions for people with visual impairments. In addition, the EU recently enacted an extensive new directive including a requirement that national copyright laws permit at least some TDM research (EU, 2019/790).

Domestic law reform

Domestic legislatures can independently amend their laws to permit TDM and other research uses without any action at the international level. Such legal adjustments are not unprecedented—copyright exceptions to permit TDM have recently appeared in fair use case law in the US (*12*) and in legislative changes in the EU, Singapore, Japan, Switzerland, and the UK (*8*). It is important for a TDM exception to apply to uses of all kinds of works (including audiovisual works used in media monitoring, for example) and enable sharing of materials at least for the purpose of collaboration and validation. Some scholars propose clarifying that all "nonexpressive" (not shared publicly) uses of works in TDM and other research should be deemed to be outside of copyright regulation (*13*). Japan recently implemented this approach into its law, adopting an exception from copyright control "where such exploitation is not for enjoying or causing another person to enjoy the ideas or emotions expressed in such work," such as "in a data analysis" (2018, Article 30-4).

The extension of TDM exceptions to commercial uses may be controversial. On one hand, many commercial users might be capable of paying licensing fees and other transaction costs, and copyright exceptions that simply transfer wealth from copyright owners to commercial TDM users, might seem arbitrary and unjustified. On the other hand, many socially beneficial uses of TDM—including the BlueDot program that originally tracked COVID-19 (*1*) or internet search engines that copy and mine the entire internet (*3*)—would likely not exist if commercial uses were excluded from copyright exceptions. Some countries see commercial TDM as a way to invest in domestic innovation and technology transfer. The EU recently adopted a rule that, although not fully tested, permits copyright holders to opt out of commercial (but not "scientific" or "cultural") TDM uses.

Research exceptions in copyright laws around the world



The map shows how "open" research exceptions are around the world to all research uses of all works by all kinds of researchers [based on (*8*)]. This includes general exceptions for research, scholarship, or personal uses and specific exceptions for text and data mining (TDM) research. Only the laws of countries labeled green are fully open, thus allowing interpretation that these permit all academic TDM projects (some countries labeled red restrict commercial uses).

- **Open research exception:** permits all research uses (including sharing) of all works by all users
- **Restrictions on sharing:** permit reproductions of whole works for research but do not extend to communications with others for purposes such as collaboration or validation
- **Restrictions to private reproduction:** provide a research exception only for use by individuals for a "private" or "personal" use
- **Restrictions to institutional users:** provide a research exception that applies only to institutions, such as nonprofit libraries
- **Restrictions on types of works:** permit research uses of whole works only of certain kinds, often prohibiting uses of whole books and databases
- **TDM restricted:** confine research exceptions to short excerpts, effectively prohibiting TDM uses
- Not mapped

GRAPHIC: ADAPTED BY N. CARY/*SCIENCE*

Another recent TDM exception, enacted by Singapore in 2021, offers a model linking generality and specificity. In addition to a broad fair use exception similar to US law (2021, Part 5, Division 2), Singapore enacted a specific TDM exception for "computational data analysis," including the ability to share reproduced works with others "for the purpose of (i) verifying the results of the computational data analysis" and for "(ii) collaborative research or study relating to the purpose of the computational data analysis" (2021: Part 5, Division 8). Unfortunately, some TDM exceptions in other countries fall short by failing to authorize the sharing of works in collaborative research or otherwise restricting the full scope of TDM methods (see the table).

Policy guidance

Guidance in the interpretation and amendment of copyright law could help policymakers evaluate their options. Even in countries with permissive legislation, there is likely to be value in clarifying the application of national law to TDM research. Such guidance can be provided, for example, through statements of best practices developed by the research community in collaboration with legal experts. Statements of best practices in fair use have been successful in enabling filmmakers, educators, research librarians, and other user-creators to confidently use copyright materials in their work (*14*).

LIBERATE AND REGULATE

The resistance to TDM exceptions in copyright comes primarily from the multinational publishing industry, which is a strong voice in copyright debates and tends to oppose expansions to copyright exceptions. But the success at adopting exceptions for TDM research in the US and EU already—where publishing lobbies are strongest—shows that policy reform in this area is possible. Publishers need not be unduly disadvantaged by TDM exceptions because publishers can still license access to their databases, which researchers must obtain in the first instance, and can offer products that make TDM and other forms of research more efficient and effective.

Policy-makers and the public may fear that expanding TDM rights will empower technology companies and models of surveillance capitalism that are under increasing scrutiny by regulators. But copyright permission does not trump privacy, consumer protection, or other regulation of the activity of technology conglomerates. Countries can liberate TDM research and still regulate these other areas.

Failing to authorize TDM research everywhere aggravates harmful disparities in our global research system. As shown in the figure, the most open regimes for TDM research are concentrated in some of the wealthiest countries and regions, whereas many poorer countries have the most restrictive copyright laws. To ensure that the needs of TDM researchers are heard in the local and national forums where copyright laws can be modified to enable this research, researchers themselves must speak out to voice their concerns and needs. It is time that copyright laws around the world are adapted to enable TDM research. ■

TDM exceptions in copyright laws

The table shows the small number of countries that had enacted specific copyright exceptions for text and data mining (TDM) research as of July 2021 [based on (*8*)]. It applies the same color scheme as that in the figure, thus showing that only a small number of countries with TDM exceptions at the time had legislated to allow all academic TDM research. Some countries (such as the UK, Switzerland, and various EU countries) are different colors in the figure than in this table because the map in the figure identifies the most open research exception, including any general exception for a research use, and this table only analyzes specific TDM exceptions. The UK, Switzerland, and some EU countries have more open general research exceptions than they provide specifically for TDM research (for example, their general research exceptions apply to all uses, not only reproduction). Whether one can apply the more general exception above and beyond the uses allowed by the TDM exception is a local legal question that this study did not attempt to answer. EU DSM Art 3, Article 3 of the EU Directive on Copyright in the Digital Single Market.

COUNTRY	COMMERCIAL	USES PERMITTED	USERS	WORKS	TYPOLOGY
Japan	Yes	Use	All	All	Open to TDM
Singapore	Yes	Reproduction, communication	All	All	Open to TDM
Germany	No	Reproduction, communication, storage	All	All	"Open to TDM (noncommercial)"
Estonia	No	"processing"	All	All	"Open to TDM (noncommercial)"
UK	No	Reproduction	All	All	Reproduction only
Switzerland	Yes	Reproduction	All	All	Reproduction only
EU DSM Art 3	Maybe	Reproduction, storage	Cultural institutions	All	Cultural institutions only
France	No	Reproduction, communication (decree)	All	Scientific writings	Limitation on works
Ecuador	Maybe	Safe harbor for liability of libraries for TDM "acts carried out by their users"	Libraries, archives (safe harbor)	All	Lacks TDM right; safe harbor only

REFERENCES AND NOTES

1. M. Prosser, "How AI Helped Predict the Coronavirus Outbreak Before It Happened," *Singularity Hub* (2020); https://singularityhub.com/2020/02/05/how-ai-helped-predict-the-coronavirus-outbreak-before-it-happened/.
2. W. Knight, "Researchers Will Deploy AI to Better Understand Coronavirus," *WIRED*, 17 March 2020.
3. OpenMinted, TDM Stories, http://openminted.eu/blog/ (2018).
4. P. Goldstein, B. Hugenholtz, *International Copyright: Principles, Law, and Practice* (Oxford Univ. Press, 2019).
5. V. Stodden, "Enabling Reproducibility in Big Data Research: Balancing Confidentiality and Scientific Transparency" in *Privacy, Big Data, and the Public Good: Frameworks for Engagement*, J. Lane, V. Stodden, S. Bender, H. Nissenbaum, Eds. (Cambridge Univ. Press, 2014), chap. 5.
6. A. Levendowski, *Wash. Law Rev.* **93**, 579 (2018).
7. R. G. Samberg, C. Hennesy, "Law and Literacy in Non-Consumptive Text Mining: Guiding Researchers Through the Landscape of Computational Text Analysis" in *Copyright Conversations: Rights Literacy in a Digital World*, S. R. Benson, Ed. (Association of College and Research Libraries, 2019), chap. 17.
8. S. Flynn, L. Schirru, M. Palmedo, A. Izquierdo, "Research Exceptions in Comparative Copyright" (PIJIP/TLS Research Paper Series no.75, 2022); https://digitalcommons.wcl.american.edu/research/75).
9. M. Palmedo, *EFIL J. Econ. Res. J.* **2**, 114 (2019).
10. C. Handke, L. Guibault, J. J. Vallbé, *MDE. Manage. Decis. Econ.* **42**, 1999 (2021).
11. S. Flynn *et al.*, *Eur. Intellect. Prop. Rev.* **42**, 393 (2020).
12. M. W. Carroll, *UC. Davis Law Rev.* **53**, 893 (2019).
13. M. Sag, *Northwest. Law Rev.* **103**, 1607 (2009).
14. M. Jacob, P. Jaszi, P. S. Adler, W. Cross, Facilitators, "Code of Best Practices in Fair Use for OER: A Guide for Authors, Adapters & Adopters of Openly Licensed Teaching and Learning Materials" (American University Washington College of Law, 2021); https://auw.cl/oer.

ACKNOWLEDGMENTS

We acknowledge the contribution of members of the Global Expert Network on Copyright User Rights and the Access to Knowledge Coalition to the research and ideas informing the drafting of this article. Funding was provided by a grant from the Arcadia Fund, a Charitable Fund of Lisbet Rausing and Peter Baldwin, for a project of the Network on the Right to Research in International Copyright. Data used for the figure and table are published in (*8*) and in the GitHub repository at https://github.com/pijip.rtr.

10.1126/science.add6124

Legal reform to enhance global text and data mining research

Sean M. Fiil-FlynnBrandon ButlerMichael CarrollOr Cohen-SassonCarys CraigLucie GuibaultPeter JasziBernd Justin JütteAriel KatzJoão Pedro QuintaisThomas MargoniAllan Rocha de SouzaMatthew SagRachael SambergLuca SchirruMartin SenftlebenOfer Tur-SinaiJorge L. Contreras

Science, 378 (6623), • DOI: 10.1126/science.add6124