AMERICAN UNIVERSITY
WASHINGTON COLLEGE OF LAW

PROGRAM ON INFORMATION JUSTICE AND INTELLECTUAL PROPERTY

WIPO/IP/AI/2/GE/201

WIPO Conversation on Intellectual Property (IP) and Artificial Intelligence (AI)

Sean Flynn, Counsel of Record
Professorial Lecturer & Director
Program on Information Justice
and Intellectual Property
American University
Washington College of Law
SFlynn@wcl.american.edu

Michael Carroll
Professor of Law
American University
Washington College of Law

Matthew Sag
Professor of Law
Loyola University, Chicago

Prof. Lucie Guibault
Associate Dean
Schulich School of Law
Dalhousie University

Dr. Thomas Margoni
Senior Lecturer
School of Law - CREATe Centre
University of Glasgow

Brandon Butler
Director of Information Policy
University of Virginia Library

Allan Rocha de Souza
Lawyer, Professor & Researcher
Federal University of Rio de Janeiro

Dr. Maja Bogataj Jančič, LL.M.
Founder & Member
Intellectual Property Institute

Peter Jaszi
Professor of Law Emeritus
American University Washington
College of Law

Dr. João Pedro Quintais
Post-doctoral Researcher
Institute for Information Law (IViR)
University of Amsterdam

Christoph Geiger
Professor & Director
Director of the Research Department
CEIPI - Université de Strasbourg

Caroline Ncube
Professor of Law
University of Cape Town

Ben White
Doctoral Researcher
Centre for Intellectual Property Policy &
Management
Bournemouth University

Arul George Scaria, Ph.D.
Professor of Law
National Law University, Delhi

Carolina Botero
Executive Director
Karisma Foundation - Colombia

Dr. Carys Craig
Associate Professor of Law
Osgoode Hall Law School
York University, Toronto

We submit this comment in response to the World Intellectual Property Organization request in relation to its work on the impact of artificial intelligence (AI) on intellectual property (IP). We are members of the Global Expert Network on Copyright User Rights with particular interest in the application of copyright to the use of text and data mining technology, including for the purposes of machine learning and artificial intelligence (AI).

We comment here only on the copyright related questions in section 13. Some of our comments with regard to the framing of the questions and defining the differences between AI, machine learning and text and data mining may apply more broadly to the entire document.

I. NEW QUESTIONS

We first address elements that we propose WIPO add to the existing set of questions.

Defining Text and Data Mining, Machine learning, and AI

As a threshold matter, all of the questions in this section (and perhaps the rest of the questionnaire) meld the definitions of text and data mining with machine learning and AI. As a result, many of the questions are confusing and difficult to answer accurately.

Text and data mining (TDM) should be used to refer to applying computational processes to materials (which could include copyrighted works) to derive data about those works. Machine learning and AI involve applying programing techniques to data (often derived from text and data mining) to enable machines to dynamically "learn" from the data inputted. Text and data mining have many other applications, including in medicine, humanities, and social science, that do not necessarily involve machine learning for the purpose of AI. Many of the copyright rules discussed in this section of questions would potentially affect text and data mining research that is both used to train AI and text and data mining research that may be unrelated to AI.

New question on WIPO's Role

Before moving to specific comments on the questions asked -- we propose a question on the WIPO role on this issue:

What actions may WIPO take that may help balance the proper role of the copyright system in promoting creativity, disseminating knowledge, and fostering

technological development in relation to the development of machine learning, artificial intelligence, and text and data mining?For example:

- *Should WIPO help explain the proper interpretation of existing law's scope of protection as including permission to run queries and otherwise apply processes to a lawfully produced corpus of copyrighted materials, and*
- *Should WIPO help facilitate the development of international norms and guidance on permitting the cross-border uses of materials and tools lawfully created in one member country to another?*

II. COMMENTS ON PROPOSED QUESTIONS

13(i). Should the use of the data subsisting in copyright works without authorization for machine learning constitute an infringement of copyright? If not, should an explicit exception be made under copyright law or other relevant laws for the use of such data to train AI applications?

We suggest that this question be rephrased as follows for the reasons expressed below:

> *Should existing law (including relevant exceptions of general applicability) be understood to permit applying computational processes to copyrighted works without authorization to derive data about those works, including for for the purposes of machine learning and AI, assuming the reproductions do not express the work to the public and even if such processes involve making temporary or ephemeral reproductions of the works studied?*
>
> *Should existing law (including relevant exceptions of general applicability) be understood to permit the technical reproduction and storage of copyrighted works to enable the application of computational processes to derive data about those works, including for the purposes of machine learning and AI, assuming the reproductions do not express the work to the public?*

Our proposed redrafted question focuses on the descriptive issue about the current state of the law because the normative question of what the law should be is addressed below.

The phrase "use of the data subsisting in copyright works without authorization" needs to be clarified throughout the questions in this section. The word "data" should be used with more precision. Text and data mining uses copies of copyrighted works as the "data" that is being analyzed or "mined". The outputs of text

and data mining analysis is data about the copyrighted works. That "data" does not "subsist in" those works, but rather is a product of observation of them.

The most relevant copyright question is whether and when temporary or more permanent copies of works may be made to enable text and data mining processes, including to train machines for AI. For this reason, the question should distinguish between at least two relevant categories of research using copyrighted works required in machine learning and AI, which may have very different treatment under copyright law:[1]

- First, the use of a data mining or other research tool to search or query a database of protected works, including to train machines for AI. Conducting a search on the Internet or querying the Google Books database are common examples. This use would literally involve "use of the data" derived from copyright works without authorization. However, the mere use of a tool to extract data from works is often not an act regulated by copyright given the fact/expression dichotomy in the law.
- Second, the making of a database (or "corpus") to be mined may involve making and storing copies of works requiring a copyright exception.

As currently phrased, the question could be answered negatively ("No, use of the data alone should not be considered infringement") by parties who nevertheless believe that making of a corpus to facilitate machine learning and AI tools may require authorization or operation of a copyright exception.

13(ii) If the use of the data subsisting in copyright works without authorization for machine learning is considered to constitute an infringement of copyright, what would be the impact on the development of AI and on the free flow of data to improve innovation in AI?

We offer the following reformulation of the question:

> *13(ii). If copyright law in some or all countries were understood to prohibit applying computational processes to copyrighted works without authorization, or were understood to prohibit the making and storing of reproductions of works to create corpora to be mined, what would be the impact on development of text and data mining research, machine*

[1] Cf. Michael Carroll, *Copyright and the Progress of Science: Why Text and Data Mining Is Lawful,* 53 U.C. DAVIS L. REV. 893 (2019), https://lawreview.law.ucdavis.edu/issues/53/2/articles/files/53-2_Carroll.pdf (distinguishing between four types of copies: "Researchers make multiple copies of the data during the TDM process. They make copies when they: (1) collect and compile the data; (2) format the data for computational processing; (3) process the data in a computer's active memory; and (4) store or archive the data to enable reanalysis or to enable validation through reproducing the analysis.").

learning and AI?

WIPO could ask more specific sub-questions to draw attention to specific impact areas, e.g.:

- New or small businesses
- Researchers, including academics, journalists, and others
- Equity and ethical issues, such as transparency, accountability, algorithmic discrimination, black box v. ethically trained AI[2]
- Interaction with other laws, such as the EU publishers right
- Complications that may arise in the use of out of commerce works
- The impacts of a globally fragmented legal system to the extent different national laws took different approaches to answering 13(i).

13(iii) If the use of the data subsisting in copyright works without authorization for machine learning is considered to constitute an infringement of copyright, should an exception be made for at least certain acts for limited purposes, such as the use in non-commercial user-generated works or the use for research?

We suggest the following reformulation:

> *13(iii). If copyright laws were understood to prohibit applying computational processes to copyrighted works without authorization to derive data about those works, or were understood to prohibit the making and storing of reproductions of works to create corpuses to be mined, including for the purposes of machine learning, should new exceptions be made under copyright law or other relevant laws to enable such activities, and subject to what restrictions, if any?*

We reiterate our concerns above about the "use of the data subsisting" formulation.

The current question asks about "limited purposes, such as the use in non-commercial user-generated works or for research."

A canvassing of the existing research exceptions that may apply to allow text and data mining activities, including to train machine learning and AI, display at least nine different categories of internal limits, with different possible impacts on the field. The questions could ask what the benefits or drawbacks may be from including such limits in research rights as compared to the models that are more open.

[2] *See* Amanda Levendowski, *How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem*, 93 WASH. L. REV. 52 (2018).

Open exceptions with "fair" practice limits. U.S. and other fair use (e.g. Israel) or open fair dealing exceptions (e.g .Singapore, Malaysia), are "open" in the sense of potentially applying to any *purpose* -- commercial and non-commercial; any *use* implicating an exclusive right (e.g. reproduction, storage, making available, etc.); all kinds of *works*; and uses by all kinds of *users*.

The operative limitation in open exceptions is that the particular use must be "fair" to the rights holder. The fairness criteria includes assessment of any impact on the market for the work.

In a line of recent cases, the fair use right in U.S. law has been interpreted to permit the reproduction of copyrighted works to create a corpus for computational uses (including of the kind that could train AI), and to making the data from the corpus available to other researchers through a search tool, as long as the process used does not re-express works to the public in a way that could compete in the market for the work.[3]

Purpose restrictions. There is variation in how the purposes of exceptions are drafted between countries. Canada, and many other fair dealing countries have exceptions broadly applying to "research."[4] Japan's exceptions cover any non-expressive use[5] or "information analysis."[6] The EU Copyright in the Digital Single Market" Directive (2019) allows acts of reproduction and extraction "for the purposes of text and data mining" by research organisations for scientific research purposes Article 3), or for any purposes but with the possibility to opt-out Article 4.[7]

Commercial use restrictions. Some research exceptions -- including text and data mining exceptions passed in the EU before the most recent directive -- are limited in their application to "non-commercial" research. WIPO should inquire into the application and impact of commercial use restrictions. How do these restrictions

---

[3] *See Authors Guild v. Google, Inc.*, 804 F.3d 202, 215 (2d Cir. 2015); *Authors Guild, Inc. v. HathiTrust* 755 F.3d 87, 105 (2d Cir. 2014).

[4] Copyright Act, 2019, Sec. 29 (Canada), *reprinted in* https://laws-lois.justice.gc.ca/eng/acts/c-42/page-9.html#h-103270. *See also Sec. 52(1)(a) of Copyright Act, Sec. 52(1)(a) (India) (fair dealing for "private or personal use, including research") http://copyright.gov.in/Documents/CopyrightRules1957.pdf*

[5] *See* Copyright Law of Japan, Article 47, *reprinted in* https://www.cric.or.jp/english/clj/doc/20161018_October,2016_Copyright_Law_of_Japan.pdf.

[6] *Id.; See also* Canada House of Commons, 2019, Standing Committee on Industry, Science and Technology, Statutory Review of the Copyright Act, Recommendation 23 *reprinted in* https://www.ourcommons.ca/DocumentViewer/en/42-1/INDU/report-16/page-189#49 (recommending amendments to Canada's Copyright Act "to facilitate the use of a work or other subject-matter for the purpose of *informational analysis*.")

[7] Bernt Hugenholtz, *The New Copyright Directive: Text and Data Mining (Articles 3 and 4)*, Kluwer Copyright Blog (July 24, 2019), http://copyrightblog.kluweriplaw.com/2019/07/24/the-new-copyright-directive-textand-data-mining-articles-3-and-4/?print=print.

impact the growth of public-private partnerships[8] or public interest commercial activities like journalism? How can the line between commercial and non-commercial activities be drawn in regard to many broadly socially beneficial commercial text and data mining products, such as Internet search, language translation, and projects that seek to harness AI for the public good?[9]

Uses implicating exclusive rights. Many research exceptions, especially those based in fair use or fair dealing, potentially apply to any use that implicates an exclusive right. Others specify the uses that are authorized, thus potentially excluding application to other uses. Specified authorized uses included in some but not all current research exceptions include:

- reproduction of the corpus[10]
- making the corpus available to other researchers[11]
- adaptation[12]
- storage[13]
- extraction[14]
- reuse[15]

WIPO may ask what the implications may be of authorizing some, but not all, uses that may be needed in data mining and machine learning. For example, in many cases, researchers need to access works from a distance. Providing such access may involve the making available right, not only the reproduction right.

---

[8] *See* NGRAM VIEWER, https://books.google.com/ngrams (last visited on Feb. 7, 2020) (a 'text mining experience' offered to all internet users through a graphic tool created in collaboration between Google and Harvard University researchers).

[9] *AI For Good with Microsoft*, MICROSOFT.COM, https://www.microsoft.com/en-us/ai/ai-for-good (last visited Feb. 6, 2020).

[10] *See* Digital Republic Act, Loi Pour Une République Numérique, 2016, Art. 38 (France), *reprinted in* https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000033202746&categorieLien=id.

[11] *See* Act on Copyright and Related Rights, Urheberrechtsgesetz, 2017, Sec. 60d (Germany), *reprinted in* https://www.gesetze-im-internet.de/englisch_urhg/englisch_urhg.html#p0431 (prohibiting "transfer" of the corpus, but authorizing "making available to a limited circle").

[12] Copyright Law of Japan, Article 47, *reprinted in* https://www.cric.or.jp/english/clj/doc/20161018_October,2016_Copyright_Law_of_Japan.pdf.

[13] EU, Germany; *See* Bernt Hugenholtz, *The New Copyright Directive: Text and Data Mining (Articles 3 and 4)*, Kluwer Copyright Blog (July 24, 2019), http://copyrightblog.kluweriplaw.com/2019/07/24/the-new-copyright-directive-text-and-data-mining-articles-3-and-4/?print=print. (explaining the necessity of storage rights for corroboration purposes).

[14] Digital Republic Act, Loi Pour Une République Numérique, 2016, Art. 38 (France), *reprinted in* https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000033202746&categorieLien=id.

[15] *Id.*

Works. All the specific exceptions, except France's current law (which may need to be changed to comply with the DSM),[16] apply their research exception to all kinds of copyrighted works. WIPO may ask for examples where data mining is useful outside the strict confines of photographs and written text that most of the literature focuses on. For example, text and data mining of audiovisual works and broadcasts are used for a variety of purposes from media monitoring to the development of language translation tools.

Transfer and sharing. Germany is the only law to explicitly address uses needed to share a data mining corpus with other researchers. It permits the making available of a corpus only to a "specifically limited circle of persons for their joint scientific research, as well as to individual third persons" for quality assurance.[17] It does not appear to permit the making available of the corpus more broadly. Art. 3(2) and 4(2) of the EU DCDSM have different wording on the need for replicability, e.g. in order to ensure that the AI has been trained in a fair, transparent, and accountable manner. WIPO may ask about the circumstances when rights to reproduce and share a corpus are necessary to accomplish machine learning and digital research ends as well as to ensure public interest regulatory objectives.

Lawfully accessed source. Three of the specific exceptions for research require that the materials used to create a corpus be "lawfully accessed." Other provisions are silent on this matter.[18] WIPO may ask what the implications of a restriction or silence may be on this matter.

Cross-border rights. Perhaps most importantly for WIPO, there is little legal certainty on whether and when a researcher can transfer, share, make available, or otherwise allow the use of a lawfully created research corpus in another country from that in which it was lawfully created. WIPO may ask whether and when cross border rights,including rights to reproduce and transfer a corpus, may be necessary for some kinds of beneficial research activities, including in the training of machines for AI.

Contract and TPM override. Notable examples of non-copyright barriers to digital research which could impede uses for machine learning and AI include contract law (e.g. purchasing or licensing restrictions on research uses) and

---

[16] *Id.* Digital Republic Act (restricting TDM rights to use of "scientific writings").

[17] Christophe Geiger et al., *The Exception for Text and Data Mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market - Legal Aspects*, CENTRE FOR INTERNATIONAL INTELLECTUAL PROPERTY STUDIES (CEIPI) RESEARCH PAPER NO. 2018-02, 23 (2018), https://ssrn.com/abstract=3160586.

[18] Michael Carroll, *Copyright and the Progress of Science: Why Text and Data Mining Is Lawful,* 53 U.C. DAVIS L. REV. 893, 951-8 (2019), https://lawreview.law.ucdavis.edu/issues/53/2/articles/files/53-2_Carroll.pdf (arguing that "copying from an infringing source necessary for TDM research is still fair use").

prohibitions on the circumvention of technological protection measures.[19] WIPO should ask about such issues.

13(iv). If the use of the data subsisting of copyright works without authorization for machine learning is considered to constitute an infringement of copyright, how would existing exceptions for text and data mining interact with such infringement?

We propose deleting this question as it would be answered in response to our reformulated question 13(i).

13(v) Would any policy intervention be necessary to facilitate licensing if the unauthorized use of data subsisting in copyright works for machine learning were to be considered an infringement of copyright?

We would reformulate this question as follows:

> *13(v). In the absence of applicable exceptions, are there policy interventions that could facilitate licensing works for text and data mining research, including to train machines for AI? What would be the strengths and weaknesses of those interventions, and how could they be made to work across borders?*

The essential problem for licensing solutions in this area is that "[t]raining data sets are likely to contain millions of different works with thousands of different owners," such that "allowing a copyright claim is tantamount to saying, not that copyright owners will get paid, but that no one will get the benefit of this new use."[20] Crafting a licensing mechanism to respond to these massive transaction costs would be exceedingly complicated. WIPO could ask specifically about some of those complications, e.g.:

- Would a collective society for mandatory collective management for all works and all uses help facilitate this issue?
- How would such a collective licensing solution work across jurisdictional boundaries?

[19] Matthew Sag, *The New Legal Landscape for Text Mining and Machine Learning,* 66 J. OF THE COPYRIGHT SOC'Y OF THE USA, 3 (2019), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3331606 (discussing contract, TPM, and cross-border issues); Thomas Margoni & Giulia Dore, *Why We Need a Text and Data Mining Exception (But it is Not Enough)*, 3 (2016), https://zenodo.org/record/248048#.WXdf2oiGNEY (stating that "a TDM exception, not limited to non-commercial purposes . . . should be implemented as soon as possible"); Ian Hargreaves, *Digital Opportunity: A Review of Intellectual Property and Growth,* (2011) https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/32563/ipreview-finalreport.pdf (recommending TPM exception for data mining).

[20] Mark A. Lemley & Bryan Casey, *Fair Learning* (2020), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3528447.

- What would be the normative basis for such a system, particularly as applied to non-expressive uses of works that do not compete in any market with the original author?
- Who would or should benefit from such a system (authors, publishers, or CMOs themselves)?
- How would the system avoid over-licensing, for example in cases where non-expressive elements or merely functional elements of the copyrighted works are used for data mining and machine learning purposes?

13(vi). How would the unauthorized use of data subsisting in copyright works for machine learning be detected and enforced, in particular when a large number of copyright works are created by AI?

The question should be edited to make its phrasing consistent with other questions in regard to eliminating the "use of data subsisting in copyright works" formulation.

We propose adding the following question:

> *What would be the impact of different enforcement regimes including, for example, the overdeterrence that may result from application of statutory damages in cases of infringement of potentially millions of works in the act of training machine learning?*

References

1. David Vaver & Pierre Sirinelli, *Principles of Copyright: Cases and Materials,* (2002), https://www.wipo.int/edocs/pubdocs/en/copyright/844/wipo_pub_844.pdf.
2. Matthew Sag, *The New Legal Landscape for Text Mining and Machine Learning,* 66 J. OF THE COPYRIGHT SOC'Y OF THE USA (2019), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3331606.
3. Thomas Margoni & Giulia Dore, *Why We Need a Text and Data Mining Exception (But it is Not Enough)*, (2016), https://zenodo.org/record/248048#.WXdf2oiGNEY.
4. Eleanora Rosati, *Copyright as an Obstacle or an Enabler? A European Perspective on Text and Data Mining and its Role in the Development of AI Creativity,* ASIA PACIFIC L. REV. (2019), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3452376.

5. Ian Hargreaves, *Digital Opportunity: A Review of Intellectual Property and Growth,* (2011) https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/32563/ipreview-finalreport.pdf.

6. Michael Geist, *Why Copyright Law Poses a Barrier to Canadian AI Ambitions*, The Globe and Mail (May 17, 2017), https://www.theglobeandmail.com/report-on-business/rob-commentary/why-copyright-law-poses-a-barrier-to-canadian-ai-ambitions/article35019241/.

7. Amanda Levendowski, *How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem*, 93 Wash L. Rev. 579 (2018), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3024938.

8. Christian Handke et al., *Is Europe Falling Behind in Data Mining? Copyright's Impact on Data Mining in Academic Research*, *in* New Avenues for Electronic Publishing in the Age of Infinite Collections and Citizen Science: Scale, Openness and Trust 120–130 (Brigit Schmidt & Milena Dobreva eds., 2015), https://pure.uva.nl/ws/files/2657677/168581_STAL9781614995623_0120.pdf.

9. Jerome H. Reichman & Ruth L. Okediji, *When Copyright Law and Science Collide: Empowering Digitally Integrated Research Methods on a Global Scale*, 96 Minn. L. Rev. 1362, 1362–2182 (2012), https://scholarship.law.duke.edu/faculty_scholarship/2675/.

10. Bernt Hugenholtz, *The New Copyright Directive: Text and Data Mining (Articles 3 and 4),* Kluwer Copyright Blog (July 24, 2019), http://copyrightblog.kluweriplaw.com/2019/07/24/the-new-copyright-directive-text-and-data-mining-articles-3-and-4/?print=print.

11. Christophe Geiger et al., *The Exception for Text and Data Mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market - Legal Aspects*, Centre for International Intellectual Property Studies (CEIPI) Research Paper No. 2018-02 (2018), https://ssrn.com/abstract=3160586.

12. Rachael G. Samberg & Cody Hennesy, *Law and Literacy in Non-Consumptive Text Mining: Guiding Researchers Through the Landscape of Computational Text Analysis*, Copyright Conversations: Rights Literacy in a Digital World (2019), https://escholarship.org/uc/item/55j0h74g.

13. Niva Elkin-Koren, *The New Frontiers of User Rights*, 32 Am. U. Int'l L. Rev. 1 (2016).

14. Thilla Rajaretnam, *Data Mining and Data Matching: Regulatory and Ethical Considerations Relating to Privacy and Confidentiality in Medical Data*, 9 J. Int'l Com. L. & Tech. 294 (2014).

15. Dennis S. Karjala, *Copyright Protection Of Computer Program Structure*, 64 Brook. L. Rev. 519, 532 (1998),

https://brooklynworks.brooklaw.edu/cgi/viewcontent.cgi?article=1782&context=blr

16. Benjamin Sobel, *Artificial Intelligence's Fair Use Crisis*, 41 COLUM. J.L. & ARTS 45 (2017), https://www.bensobel.org/files/articles/41.1_Sobel-FINAL.pdf.

17. Arul George Scaria, *Should Indian Copyright Law Prevent Text and Data Mining?,* Spicy IP (August 21, 2019), https://spicyip.com/2019/08/should-indian-copyright-law-Prevent-text-and-data-mining.html.

18. Ariel Katz,*The Orphans,The Market, and the Copyright Dogma: A Modest Solution to a Grand Problem*, 27 BERKELEY TECH. LAW J. 1285 (2012), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2118886.

19. Michael Carroll, *Copyright and the Progress of Science: Why Text and Data Mining Is Lawful,* 53 U.C. DAVIS L. REV. 893, 951-8 (2019), https://lawreview.law.ucdavis.edu/issues/53/2/articles/files/53-2_Carroll.pdf (arguing that “copying from an infringing source necessary for TDM research is still fair use”).

20. Mark A. Lemley & Bryan Casey, *Fair Learning* (2020), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3528447.