CSP CADERNOS DE SAÚDE PÚBLICA
REPORTS IN PUBLIC HEALTH

Text and data mining in health research: reflections on copyright

Mineração de textos e dados na pesquisa em saúde: reflexões sobre direitos autorais

Minería de textos y datos en la investigación en salud: reflexiones sobre los derechos de autor

Allan Rocha de Souza [1,2]
Luca Schirru [2,3]
Miguel Bastos Alvarenga [1,2]

doi: 10.1590/0102-311XEN169023

[1] Programa de Pós-graduação em Políticas Públicas, Estratégias e Desenvolvimento, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brasil.
[2] Instituto Brasileiro de Direitos Autorais, Rio de Janeiro, Brasil.
[3] Centre for IT & IP Law (KU Leuven), Leuven, Belgium.

Correspondence
M. B. Alvarenga
Rua do Humaitá 282, apto. 1306, bloco 2, Rio de Janeiro, RJ 22261-004, Brasil.
miguel.alvarenga@pped.ie.ufrj.br

The usage of computer systems in the collection, organization, and analysis of large volumes of data and texts and, in this context, text and data mining (TDM) [1] techniques is critical to contemporary data-intensive studies, such as those related to the COVID-19 pandemic [2,3,4,5,6,7,8]. In addition to the ethical issues concerning research data appropriation [8], copyright brings specific challenges regarding the use of TDM, due to the objects to which protection applies, its temporal extension, broad scope of assigned rights, and scarcity of explicit limitations. Because of their importance for research, our study is focused on these obstacles and ongoing regulatory actions.

Copyright does not protect facts, information or data, nor even the content of a work, as its object of protection is the literary or artistic form in which they are expressed and communicated. However, when these elements are combined, organized or systematized in a database that is minimally original in terms of content selection, organization or arrangement, this material will have its access and use controlled by its owner and then protected by copyright [9,10,11].

Also, legal protection by copyright of databases is just one of the layers of control over access and use of data and information [11]. A second layer involves technological restrictions to access or use of databases [12], and a third layer refers to legal prohibitions to circumvent these mechanisms, even when there is no direct violation of any right over the material [13]. It has a direct impact on essential research activities, such as verification, reproducibility, and communication of results, by allowing or limiting who can perform what type of research, use what type of material, and under what conditions [14,15,16]. This is highlighted in TDM, as it requires uses from as many available sources as possible, which do not always have open access. Therefore, using only this type of source with this method can lead to exacerbated biases in the results [15].

There is an ongoing demand for regulatory reforms at a global level, given the normative, technological, and contractual extension of control powers of data holders, which imposes legal challenges on numerous data-intensive research activities [17,18,19] – a fact particularly observed in countries of the Global South [19,20,21]. Given the above, several legal systems around the world started in the last decade reforms in their copyright legislation to include standards with a focus on ensuring the legality of TDM [22], such as the European Union [23,24] and countries like Japan [25,26] and Singapore [27].

In Brazil, potential challenges and opportunities have been identified regarding the use of data-intensive technologies in the texts of the Brazilian Artificial Intelligence Strategy (EBIA, acronym in Portuguese) [28] and the Brazilian National Intellectual Property Strategy (ENPI, acronym in Portu-

guese) [29]. However, so far, TDM has been the most extensively discussed by the Committee of Jurists on Artificial Intelligence (CJUSBIA, acronym in Portuguese), responsible for supporting the drafting of a new law proposition for Artificial Intelligence [30].

The CJUSBIA Final Report, now *Bill n. 2,338/2023* of the Brazilian Federal Senate, proposes the definition of text and data mining for regulatory purposes, in addition to limitation (or exception) to copyrights of holders of databases and other works involved in this process [31]. The limitation proposed in article 42 of *Bill n. 2,338/2023* has a clear and delimited scope, only authorizing access and use by entities whose mission is connected to the public interest such as research and journalism institutions, and to the extent required for the intended objectives. Limitations also describe that the use, for protected works, should not imply the creation of competing products or that somehow impact the reasonable exploitation of the works, for example [31].

Some argue that it would not require limitations to text and data mining [24,32], because it is a non-expressive use [33,34] or because what is extracted from works – factual elements and patterns, for example – would not be covered by the scope of copyright protection [24,32,34]. However, multiple projects have already resorted to the direct use of copyrighted content and caused legal disputes, even when the material had been legally accessed [33]. Thus, an express, permissive norm is extremely important to ensure legal security for research institutions and organizations, journalism, museums, archives, libraries, and for their employees, to help them pursue their institutional objectives and missions in the contemporary context of research, innovation and technological development [35].

More broadly, the discussion about the legitimacy and importance of TDM for research purposes has involved a broader debate about the recognition, contours, and effectiveness of the right to research as a fundamental guarantee. Regularly linked in normative texts to the right to education and access to knowledge (article 27 of the *Universal Declaration of Human Rights* and article 206, II and article 208, V of the *Brazilian Federal Constitution*) [36,37], or more broadly to science and innovation (article 218 et seq. of the *Brazilian Federal Constitution*) [37], its particularities demand its own independence and autonomy; and the development of its own legal framework, normative functions, and legal effects is still unsder construction [38].

However, the implementation and enforcement of this right have not kept up with its social relevance. The recent pandemic has exposed the centrality and importance of research and science in the functioning of society. Recent years have highlighted the political and social vulnerability to which research activities are subject. Recognizing the legitimacy of text and data mining – particularly, but not only, in research purposes – is an important and necessary step, although not sufficient, for the development and consolidation of the right to research, which we understand as essential to foster innovation, of economic development, and technological autonomy of the country.

Contributors

A. R. Souza contributed to the study conception, data analysis, writing and review; and approved the final version. L. Schirru contributed to the study conception, data analysis, writing and review the text; and approved the final version. M. B. Alvarenga contributed to the study conception, data analysis, writing and review the text; and approved the final version.

Additional information

ORCID: Allan Rocha de Souza (0000-0002-6549-0085); Luca Schirru (0000-0002-4706-3776); Miguel Bastos Alvarenga (0000-0002-4079-7785).

References

1. Han J, Pei J, Kamber M. Data mining: concept and techniques. 3rd Ed. Waltham: Morgan Kaufmann; 2012.
2. Agrela L. Inteligência artificial previu epidemia do coronavírus da China. Exame 2020; 28 jan. https://exame.com/tecnologia/inteligencia-artificial-previu-epidemia-do-coronavirus-da-china/.
3. Stieg C. How this Canadian start-up spotted coronavirus before everyone else knew about it. CNBC Make it 2020; 3 mar. https://www.cnbc.com/2020/03/03/bluedot-used-artificial-intelligence-to-predict-coronavirus-spread.html.
4. Souza AR, Schirru L, Alvarenga MB. Direitos autorais e mineração de dados e textos no combate à COVID-19 no Brasil. Liinc em Revista 2020; 16:e5536.
5. You J, Expert P, Costelloe C. Using text mining to track outbreak trends in global surveillance of emerging diseases: ProMED-mail. J R Stat Soc Ser A Stat Soc 2021; 184:1245-59.
6. Safdari R, Rezayi S, Saeedi S, Tanhapour M, Gholamzadeh M. Using data mining techniques to fight and control epidemics: a scoping review. Health Technol (Berl) 2021; 11:759-71.
7. Pedroso MM, Lima JC, Assef Neto VB. Ciência de dados aplicada ao arca: disponibilização de ferramentas para recuperação da informação no repositório institucional da Fundação Oswaldo Cruz. RECIIS 2017; 11 Suppl. https://www.reciis.icict.fiocruz.br/index.php/reciis/article/view/1417/pdf1417.
8. Chiruvella V, Guddati AK. Ethical issues in patient data ownership. Interact J Med Res 2021; 10:e22269.
9. Brasil. Lei nº 9.610, de 19 de fevereiro de 1998. Altera, atualiza e consolida a legislação sobre direitos autorais e dá outras providências. Diário Oficial da União 1998; 20 feb.
10. Organização Mundial do Comércio. Acordo sobre Aspectos dos Direitos de Propriedade Intelectual Relacionados ao Comércio (TRIPS). https://www.wto.org/english/docs_e/legal_e/27-trips.pdf (accessed on 10/Jan/2024).
11. Derclaye E. The legal protection of databases: a comparative analysis. Cheltenham: Edward Elgar; 2008.
12. Toth AK. Algorithmic copyright enforcement and al: issues and potential solutions, through the lens of text and data mining. Masaryk University Journal of Law and Technology 2019; 13:361-87.
13. Brown K. Digital rights management: trafficking in technology that can be used to circumvent the intellectual property clause. Houston Law Rev 2003; 803:803-36.
14. Reichman JK, Okediji RL. When copyright law and science collide: empowering digitally integrated research methods on a global scale. Minn Law Rev 2012; 96:1362-480.
15. Levendowski A. How copyright law can fix artificial intelligence's implicit bias problem. Washington Law Review 2018; 93:579-630.
16. Ducato R, Strowel A. Limitations to text and data mining and consumer empowerment: making the case for a right to machine legibility. CRIDES Working Paper Series 2018; 31 oct. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3278901 (accessed on 30/Jun/2023).
17. Caspers M, Guibault L. A right to 'read' for machines: assessing a black-box analysis exception for data mining. Proceedings of the Association for Information Science and Technology 2016; 53:1-5.
18. Flynn S, Butler B, Carroll M, Cohen-Sasson O, Craig C, Guibault L, et al. Legal reform to enhance global text and data mining research: outdated copyright laws around the world hinder research. Science 2022; 378:951-3.

19. Souza AR. COVID-19, Text and data mining and copyright: the Brazilian case. In: World Intelectual Property Organization; World Trade Organization, editors. WIPO-WTO Colloquium Papers. v. 11. s/l: World Intelectual Property Organization/World Trade Organization; 2020. p. 1-14.
20. Bertón M. Text and data mining exception in South America: a way to foster AI development in the Region. GRUR International 2021; 70:1145-57.
21. Flynn S, Schirru L, Palmedo M, Izquierdo A. Research exceptions in comparative copyright. 2022. https://digitalcommons.wcl.american.edu/cgi/viewcontent.cgi?article=1077&context=research (accessed on 10/Jan/2024).
22. Palmedo M, Alvarenga M, Imran M, Le D, Schirru L. Measuring change in copyright exceptions for text and data mining. https://digitalcommons.wcl.american.edu/cgi/viewcontent.cgi?article=1100&context=research (accessed on 10/Jan/2024).
23. European Union. Regulation (EU) 2023/2854 of the European Parliament and of the Council of 13 December 2023 on harmonised rules on fair access to and use of data and amending Regulation (EU) 2017/2394 and Directive (EU) 2020/1828 (Data Act). http://data.europa.eu/eli/reg/2023/2854/oj (accessed on 12/Jan/2024).
24. Margoni T, Kretschmer M. A deeper look into the EU text and data mining exceptions: harmonisation, data ownership, and the future of technology. GRUR International 2022; 71:685.
25. World Intellectual Property Organization. Copyright Act (Act No. 48 of May 6, 1970, as amended up to Act No. 72 of July 13, 2018). https://www.wipo.int/wipolex/en/text/504411 (accessed on 08/Apr/2024).
26. Ueno T. The flexible copyright exception for 'non-enjoyment' purposes - recent Amendment in Japan and its implication. GRUR International 2021; 70:145-52.
27. World Intellectual Property Organization. Copyright Act 2021 (Revised edition 2020, Act No. 22 of 2021). https://www.wipo.int/wipolex/en/text/584840 (accessed on 30/Jun/2023).
28. Ministro de Estado da Ciência, Tecnologia e Inovações. Portaria GM nº 4.617, de 6 de abril de 2021. Institui a Estratégia Brasileira de Inteligência Artificial e seus eixos temáticos. Diário Oficial da União 2021; 9 apr.
29. Brasil. Decreto nº 10.886, de 7 de dezembro de 2021. Institui a Estratégia Nacional de Propriedade Intelectual. Diário Oficial da União 2021; 8 dec.
30. Schirru L, Souza AR, Chamas C. Building a text and data mining limitation: the Brazilian case. GRUR International 2024; 73:217-22.
31. Brasil. Projeto de Lei nº 2.338/23. Dispõe sobre o uso da Inteligência Artificial. https://legis.senado.leg.br/sdleg-getter/documento?dm=9347593&ts=1684441712901&disposition=inline&_gl=1*9y8waw*_ga*NzYwNzg2OTY4LjE2ODgxMzg3NTQ.*_ga_CW3ZH25XMK*MTY4ODE0MDg1NC4yLjAuMTY4ODE0MDg2Ni4wLjAuMA (accessed on 30/Jun/2023).
32. Senftleben M. Compliance of national TDM rules with International Copyright Law: an overrated nonissue? International Review of Intellectual Property and Competition Law 2022; 53:1477-505.
33. Carroll MW. Copyright and the progress of science: why text and data mining is lawful. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3531231 (accessed on 30/Jun/2023).
34. Sag M. The new legal landscape for text mining and machine learning. Journal of the Copyright Society of the USA 2019; 66:1-34.
35. Alvarenga MB. Mineração de dados, Big Data e direitos autorais no Brasil [Doctoral Dissertation]. Rio de Janeiro: Universidade Federal do Rio de Janeiro; 2019.
36. Fundo das Nações Unidas para a Infância. Declaração Universal dos Direitos Humanos (1948). https://www.unicef.org/brazil/declaracao-universal-dos-direitos-humanos (accessed on 25/Jan/2024).
37. Brasil. Constituição Federal da República Federativa do Brasil, 05 de outubro de 1988. Diário Oficial da União 1988; 5 oct.
38. Geiger C, Jütte BJ. The Right to Research as Guarantor for Sustainability, Innovation and Justice in EU Copyright Law (June 19, 2022). In: Pihlajarinne T, Mähönen J, Upreti P, editors. Rethinking the role of intellectual property rights in the post pandemic world: an integrated framework of sustainability, innovation and global justice. Cheltenham/Northampton: Edward Elgar; 2023. p. 138-69.

Submitted on 06/Sep/2023
Final version resubmitted on 29/Jan/2024
Approved on 15/Feb/2024