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) techniques is critical to contemporary data-intensive studies, such as those related to the COVID-19 pandemic. In addition to the ethical issues concerning research data appropriation, 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. Legal protection by copyright of databases is just one of the layers of control over access and use of data and information. A second layer involves technological restrictions to access or use of databases, 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. 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.

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.

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 – a fact particularly observed in countries of the Global South. 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, such as the European Union and countries like Japan and Singapore.

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) and the Brazilian National Intellectual Property Strategy (ENPI, acronym in Portuguese).
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.

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. 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.

Some argue that it would not require limitations to text and data mining, because it is a non-expressive use or because what is extracted from works – factual elements and patterns, for example – would not be covered by the scope of copyright protection. 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. 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.

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), or more broadly to science and innovation (article 218 et seq. of the Brazilian Federal Constitution), its particularities demand its own independence and autonomy; and the development of its own legal framework, normative functions, and legal effects is still underway construction.

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


