REGLAB CONTEXT

Global Mapping of Copyright Laws and Exceptions for Data Mining

August 2025

reglab
center for strategy & regulation

observatório
direitos autorais e tecnologia

About Reglab

We are a private **research center specializing in the media and technology sector**, supporting companies, associations, and policymakers in making strategic decisions based on data and evidence.

About the Context Series

RegLab's **Context** Series presents **concise briefs that synthesize specific topics or emerging trends**. They are designed to deliver information in a clear and accessible way, incorporating visual elements such as charts, tables, and infographics, combining analytical rigor with practicality

Acknowledgements


Authors: Natália Góis Ribeiro and Marina Gonçalves Garrote

Researchers: Natália Góis Ribeiro

Cartographic Design: Amanda Silva Almeida

Final Design: Eliza Natsuko Shiroma, Camyla Romão

suggested citation

RIBEIRO, N.; GARROTE, M. *Global Mapping of Copyright Laws and Exceptions for Data Mining*. Reglab Context, no. 1. Reglab, 2025.

observatório
direitos autorais e tecnologia

This study is part of the **Observatory on Copyright and Technology**.
The Observatory is an initiative of **RegLab and its partners**, dedicated to producing studies and reports on the subject, with the goal of strengthening the debate through **clarity and evidence**.

learn more at http://observatorio.cc

executive summary

- **Text and Data Mining (TDM)** — also known as computational analysis — is a technique used to extract patterns from large volumes of content, playing a key role in training AI models and conducting scientific research.
- This study mapped the **copyright frameworks of 50 countries**, focusing on legal exceptions applicable to TDM, to understand how different jurisdictions approach the intersection between data mining and copyright law.
- **The international landscape is highly diverse**. There is no global standard regarding copyright exceptions for TDM, with significant variations across countries.
- **Unequal access to TDM and the global nature of digital networks may deepen global asymmetries in AI development**. High-income countries tend to have the most permissive legal frameworks, while others still operate under more restrictive regimes.
- **Brazilian law remains ambiguous regarding the legality of TDM on protected works**, which may create legal uncertainty for technical, scientific, and commercial uses.
- Introducing explicit exceptions for TDM in Brazil, rather than a mere alignment with foreign models, **could serve as a strategic move to strengthen the country's technological sovereignty**.

reglab
center for strategy & regulation

Clickable index

introduction

Amid the global race to accelerate the development of Artificial Intelligence (AI), countries have adopted different approaches regarding what can — and cannot — be used to train these systems.

Large-scale data extraction and analysis techniques, such as Text and Data Mining (TDM), are essential for the advancement of AI. However, the legal framework governing their use remains uncertain, particularly in light of copyright laws that were designed for a different era.

While some countries have introduced more flexible copyright exceptions to enable computational uses, others maintain strict interpretations. In Brazil, this discussion has reached Congress through Bill No. 2338/2023, which currently includes specific provisions on the subject.

To understand the risks and opportunities of AI regulation in Brazil from a copyright perspective, it is essential to examine how different countries are addressing the issue. This summary, part of the Reglab Context series, aims to **translate comparative evidence on copyright exceptions into a visual and accessible format,** offering a starting point for Brazil's regulatory debate.

We acknowledge that the topic involves legitimate tensions, and it is from this understanding that we seek to contribute to the discussion with fresh perspectives.

> “Regulation can both protect creators, by reducing barriers to knowledge access, and perpetuate economic asymmetries, turning the full use of AI into a privilege for few”

reglab
center for strategy & regulation

Clickable index

what is text & data mining?

Training AI models involves complex stages — from data collection and curation to parameter tuning and performance evaluation — using various techniques to analyze large volumes of information. Among these, Text and Data Mining (TDM) stands out as a key tool designed for the automated analysis of large-scale content.

> TDM is a technique that predates the generative AI boom and has been widely used across fields to identify patterns, trends, and correlations within massive datasets.

TDM is broadly applied across sectors and serves a wide range of purposes. During the COVID-19 pandemic, for example, it was crucial in accelerating scientific discoveries, enabling the automated analysis of thousands of medical research papers. Governments and public organizations have also used TDM to detect disinformation and hate speech on social media, to support the design of public policies, and to monitor climate change. In the private sector, the technique is applied in search engines, machine translation, voice recognition, and even in investigative journalism, helping to cross-reference large volumes of documents to identify patterns and inconsistencies.

Stages Involved in Data Mining

Data Collection
Gathered from various sources

Pre-processing
Preparation and transformation of data

Data Organization
Fast access and storage of data for retrieval and mining

Mining
Algorithmic inference and information extraction

Analysis
User analysis and navigation

reglab
center for strategy & regulation

Clickable index

so, is TDM compatible with copyright law?

It depends. That's precisely why this study set out to examine which national legal frameworks already provide for copyright exceptions that enable AI training.

We began with the 2022 article "*Legal Reform to Enhance Global Text and Data Mining Research*" to identify whether and how countries include in their legislation exceptions that authorize the use of protected works for TDM, whether in scientific, institutional, or commercial contexts.

We recognize that not all uses of TDM are for AI training, and not all AI training depends solely on TDM. However, our research revealed that many countries have adopted a broad legal definition of TDM, one that may also encompass its use in AI training.

In light of this, and due to the absence of a formal definition in Brazilian law, we adopted Directive (EU) 2019/790 as our methodological reference. It defines Text and Data Mining as:

> "any automated analytical technique aimed at analyzing text and data in digital format in order to generate information such as patterns, trends, and correlations, among others (EU, 2019).

We then analyzed national copyright laws available in the WIPO repository, supplementing with secondary data where necessary. Our sample included 50 countries, which were classified into three levels of permissiveness regarding copyright exceptions for TDM: high, medium, and low.

High Permissiveness	This category includes countries that, cumulatively: (i) explicitly authorize TDM; (ii) do not restrict user profiles; and (iii) allow commercial use of TDM outputs. It also includes common law jurisdictions, such as the United States—where the fair use doctrine has been interpreted by courts as permissive of TDM in multiple cases to date.
Medium Permissiveness	This group includes countries whose laws: (i) limit TDM exceptions to specific groups (e.g., universities, libraries); **and/or** (ii) restrict permissions to non-commercial uses only.
Low Permissiveness	Jurisdictions in this category **do not provide** explicit exceptions for TDM under current copyright laws.

Copyright & Text and Data Mining (TDM)

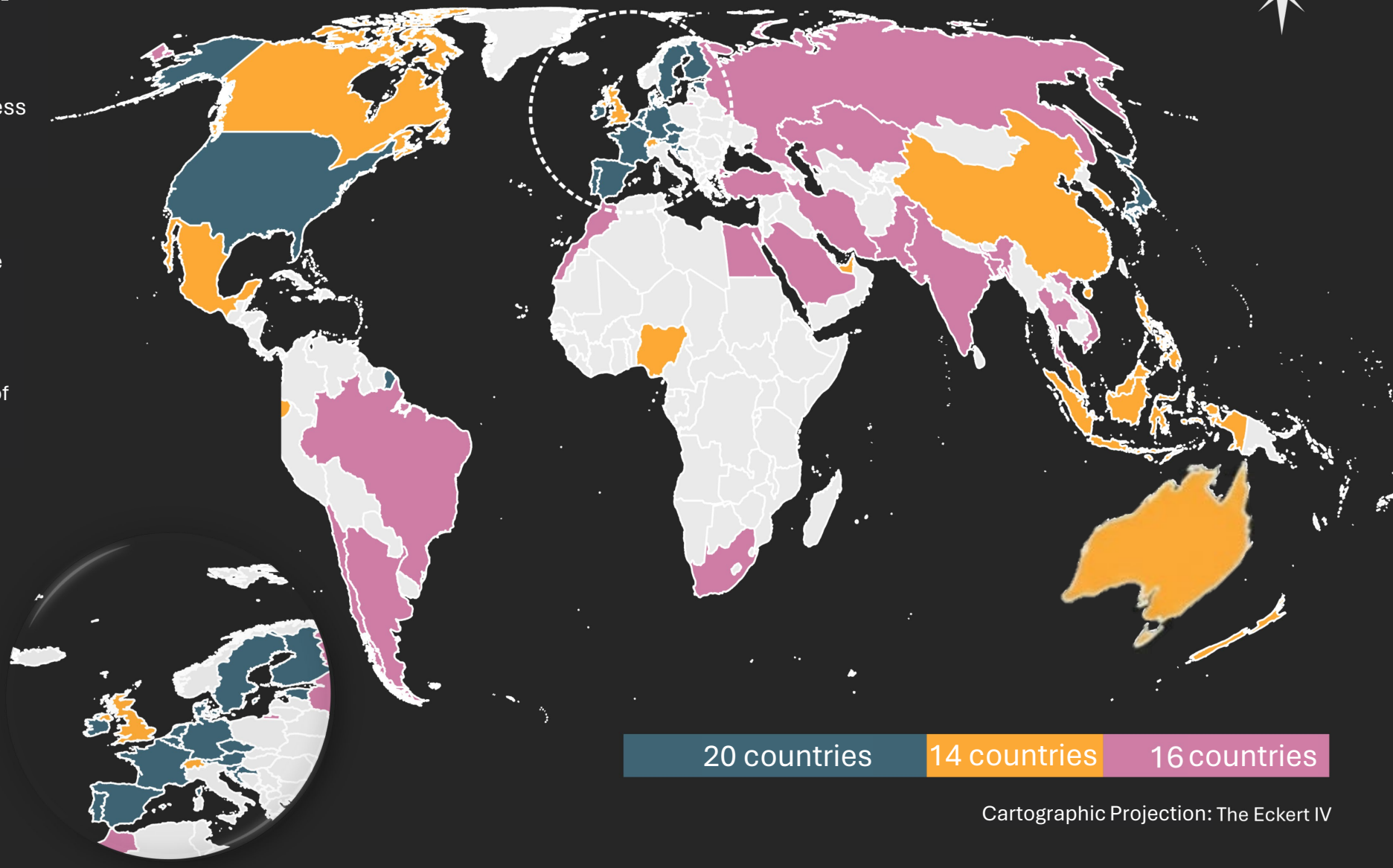
International Comparison of Permissiveness Classification Across Countries

High Permissiveness: (i) explicitly authorize TDM; (ii) do not restrict user profiles; and (iii) allow commercial use of TDM outputs. It also includes common law jurisdictions, such as the United States—where the fair use doctrine has been interpreted by courts as permissive of TDM in multiple cases to date.

Medium Permissiveness: (i) limit TDM exceptions to specific groups (e.g., universities, libraries); and/or (ii) restrict permissions to non-commercial uses only.

Low Permissiveness: Jurisdictions in this category do not provide explicit exceptions for TDM under current copyright laws.

Not mapped



20 countries | 14 countries | 16 countries

Cartographic Projection: The Eckert IV

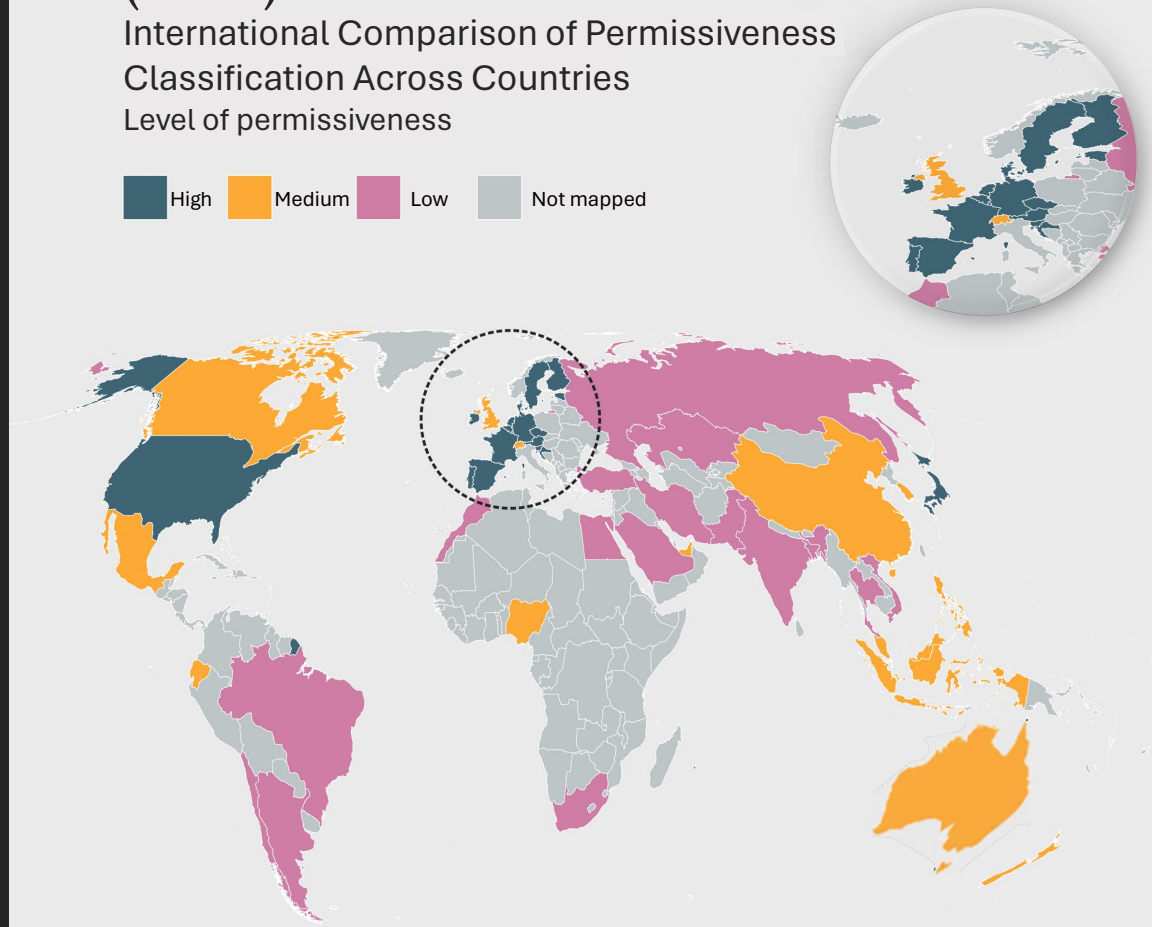
reglab
center for strategy & regulation

Clickable index

Copyright & Text and Data Mining (TDM)

International Comparison of Permissiveness Classification Across Countries

Level of permissiveness



High Permissiveness: (i) explicitly authorize TDM; (ii) do not restrict user profiles; and (iii) allow commercial use of TDM outputs. It also includes common law jurisdictions, such as the United States—where the fair use doctrine has been interpreted by courts as permissive of TDM in multiple cases to date.

Medium Permissiveness: (i) limit TDM exceptions to specific groups (e.g., universities, libraries); and/or (ii) restrict permissions to non-commercial uses only.

Low Permissiveness: Jurisdictions in this category do not provide explicit exceptions for TDM under current copyright laws.

High Permissiveness	Austria, Belgium, Croatia, Czech Republic, Denmark, Estonia, Finland, France, Germany, Ireland, Japan, Luxembourg, Malta, Netherlands, Portugal, Singapore, Slovenia, Spain, Sweden, United States.
Medium Permissiveness	Australia, Canada, China, Ecuador, Indonesia, Malaysia, Mexico, New Zealand, Nigeria, Philippines, South Korea, Switzerland, United Arab Emirates, United Kingdom.
Low Permissiveness	Argentina, Bangladesh, Brazil, Chile, Egypt, India, Iran, Kazakhstan, Morocco, Pakistan, Russia, Saudi Arabia, South Africa, Thailand, Turkey, Vietnam.

reglab
center for strategy & regulation

Clickable index

deep dive

In the United States, the fair use doctrine has thus far allowed data mining practices. In cases preceding the generative internet boom, U.S. courts generally authorized TDM — including for commercial purposes — when the use was lawful, transformative, and did not harm the original market.
Notable cases such as Authors Guild v. Google (2011), HathiTrust (2014), and Fox v. TVEyes (2014) illustrate this favorable interpretation.
However, this understanding is being challenged in recent lawsuits, including The New York Times v. OpenAI (2023) and Authors v. Anthropic (2025), which explore the limits of TDM practices in training generative AI systems.
In a 2025 report, the U.S. Copyright Office emphasized that there is still no definitive answer regarding the application of the fair use doctrine in such cases. The Office stressed that each situation must be assessed individually, based on the four factors of fair use.

Since 2018, Japanese law has allowed the use of copyrighted works without prior authorization, including for commercial purposes, as long as the use is aimed at information analysis rather than the creative reproduction of the original work. This exception effectively permits TDM techniques to be used without the need for licensing or payment to rights holders.

More recently, Japan's Agency for Cultural Affairs reaffirmed that, in its interpretation, the use of copyrighted works for AI training is, in principle, allowed without the prior consent of the rights holder.

reglab
center for strategy & regulation

Clickable index

deep dive

The **European Union** adopted **Directive 2019/790**, which introduced specific exceptions for TDM, both for scientific research purposes and for general purposes, including commercial use. As a directive, its implementation depends on transposition into national law, which in practice allows for variations across member states in how the directive is applied.

The EU also has a dedicated AI law: the **AI Act.** While the law's text does not directly address copyright issues, Recital 105, which provides interpretative guidance, reinforces that TDM must comply with the rules set out in the Directive, including the opt-out mechanism. This mechanism allows copyright holders to prohibit the use of their works for TDM purposes if they explicitly opt out.

In **Singapore**, copyright law explicitly authorizes the copying and use of legally accessed content for the purposes of computational analysis, including pattern recognition through software.
Commercial use is also permitted under the country's legal framework.

reglab
center for strategy & regulation

Clickable index

deep dive

The United Kingdom has recognized TDM as a copyright exception since 2014, allowing individuals to make copies for computational analysis, provided the use is non-commercial.

In June of this year, the country passed the Data (Use and Access) Act. One of its sections requires the government to produce reports evaluating the economic and legal impacts of different regulatory models, such as the feasibility of exception regimes with opt-out options for rights holders, and the adoption of transparency mechanisms regarding the data used to train AI models.

Although there is no specific legal exception for TDM, South Korean copyright law allows the use of protected works for education, research, citation, personal use, and use within libraries, provided the use remains within reasonable limits and does not cause harm to rights holders.

In 2021, a bill was proposed to create a broad and specific exception for TDM, but the proposal is still under discussion.

China demonstrates a medium level of permissiveness regarding data mining practices. The Copyright Law, revised in 2020, allows a range of exceptions for educational purposes, non-profit scientific research, library preservation, and accessibility. However, it does not explicitly mention TDM practices.

reglab
center for strategy & regulation

Clickable index

deep dive

Nigerian legislation does not explicitly reference data mining, but it incorporates the concept of fair dealing, which allows the use of copyrighted works without prior authorization from the rights holder, provided certain conditions are met, such as reasonable and proportionate use that does not cause undue harm to the market for the original work.
These exceptions apply to both institutional users and the general public.

Ecuador is the only country in South America whose copyright legislation explicitly mentions TDM, authorizing “text mining” within the scope of libraries and archives.
However, the scope of this exception is limited to institutional use, such as by libraries and archives, and is conditioned on the users’ good faith and adherence to “honourable uses”, a standard that considers factors such as purpose, extent of use, and market impact on the original work.

reglab
center for strategy & regulation

Clickable index

deep dive

South Africa has a **more restrictive approach** to the use of copyrighted works. The current legislation permits limited uses under the fair dealing doctrine, which allows excerpts of works to be used for private study, teaching, or news reporting, provided such use does not harm the market for the original work.

India follows a fair dealing regime, with copyright exceptions limited to non-commercial uses, such as study, research, and journalism, and does not currently provide a specific exception for TDM.

The country is working toward developing a broader AI governance framework. In 2025, a Multi-Stakeholder Group was established to design a regulatory framework, following the publication of a report with recommendations for building a responsible AI ecosystem, which was submitted for public consultation. The document includes a dedicated section on copyright, raising regulatory questions and suggesting the need for clearer guidance on the use of protected data in the context of AI.

reglab
center for strategy & regulation

Clickable index

what about Brazil?

Brazil, like **most Latin American countries, is characterized by a more restrictive copyright framework**.

The Copyright Law (Law No. 9,610/1998) has not been updated to address TDM practices and does not provide broad exceptions. Although there are ongoing debades about the scope of rights granted under the law, the current legal landscape creates a scenario of legal uncertainty.

Bill No. 2,338/2023, currently under discussion in the Chamber of Deputies, proposes a limited exception for text and data mining, allowing such activity only when carried out for the purpose of research and development of AI systems, and exclusively by scientific and educational institutions, museums, archives, and libraries.

In these cases, protected content may be used without prior authorization from rights holders, provided that copyright restrictions (opt-out) are not expressly stated. The text of the bill also allows rights holders to retroactively claim against the use of their works, even after the system has already been trained. Outside these specific circumstances, rights holders would be entitled to prohibit the use of their works for data mining purposes. The bill also establishes that AI agents using protected content for mining, training, or development purposes must compensate the rights holders of such works.

to learn more about the limits and implications of Bill No. 2,338/2023, access our full study on the subject!

access our study here

reglab
center for strategy & regulation

Clickable index

analysis and commentary

authors' perspective

The map reveals a fragmented landscape: there is no regulatory convergence on copyright exceptions for data mining, which affects legal certainty for research and AI training. However, attention must be given to the geographic distribution: in general, more permissive regimes are concentrated in developed countries, while restrictive legislation prevails in Global South economies.

The absence of exceptions may negatively impact academic output and local innovation capacity, further widening the gap between countries. Moreover, the **cross-border nature of the internet allows companies to relocate AI training activities to more permissive jurisdictions, reducing the practical reach of restrictive national laws.**

Although part of the resistance to exceptions may come from sectors reliant on the exploitation of intellectual property rights (e.g., the publishing industry), the experience of the EU and the U.S. shows that **well-designed exceptions can coexist with sustainable business models**, including opportunities for database licensing under opt-out regimes.

Finally, concerns about technological sovereignty and the dominance of large technology companies are legitimate. However, restricting TDM exceptions may have the opposite effect of what is intended. **Well-structured exceptions can enable the local development of AI systems and the creation of competitive alternatives, including by small developers, universities, and research centers**. This is why the debate on copyright can, and should, advance without disregarding other legitimate discussions, such as data protection, content moderation, and competition policy.

reglab
center for strategy & regulation

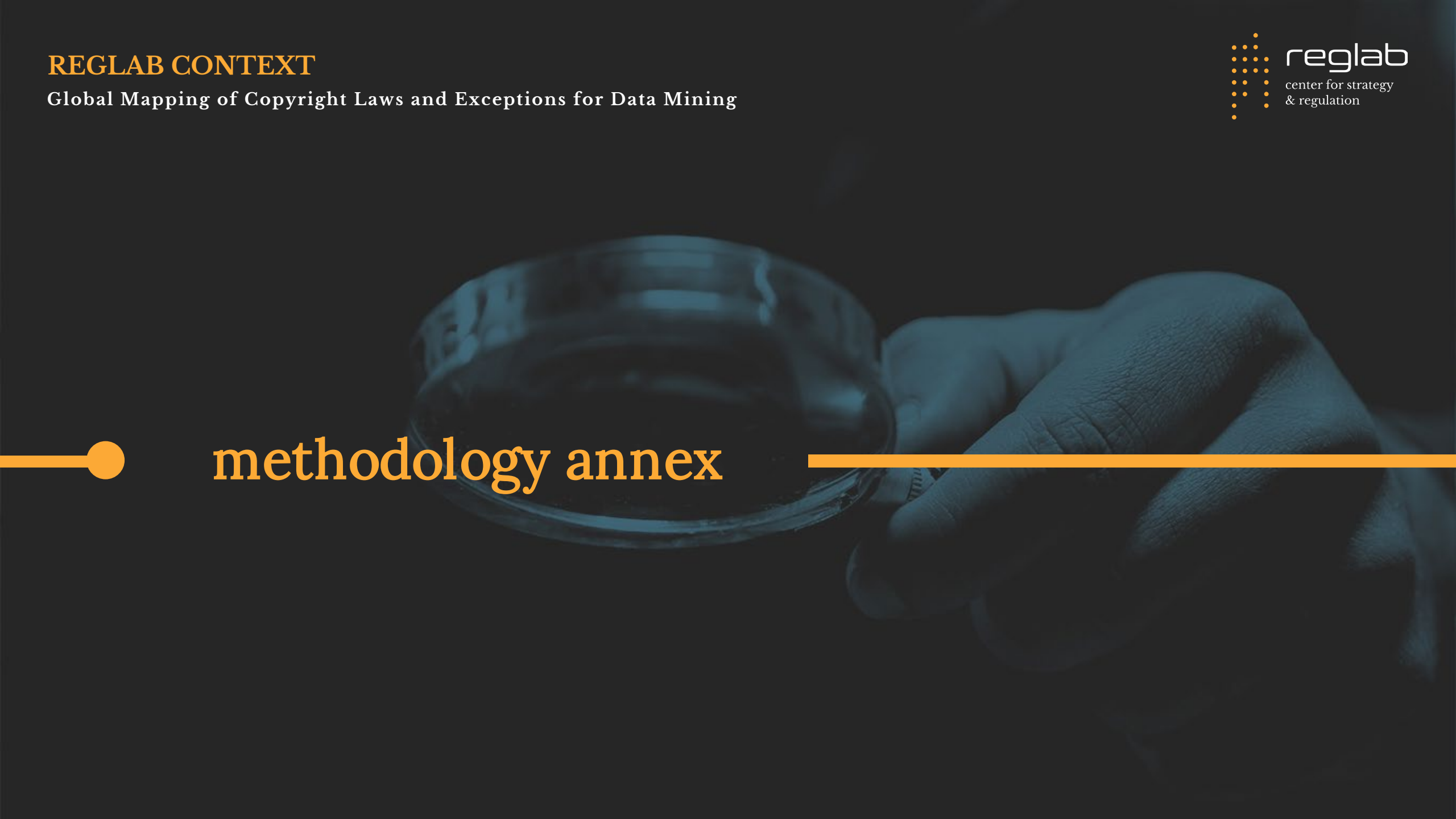
Clickable index

direction for future studies

The context series consists of concise, non-exhaustive studies designed to spark debate and encourage further research. Future studies could adopt different approaches to complement this work, such as:

- Comprehensive mapping of countries, including the analysis of bills currently under discussion;
- Assessment of social costs and benefits, grouping different stakeholders and measuring externalities;
- Economic analysis of the effects of different regulatory alternatives, including the actual gains or losses for copyright holders in each scenario;
- Scenario simulation, using methods such as gaming, Delphi, or Shell frameworks to estimate the social consequences of different regulatory decisions.

reglab
center for strategy
& regulation



methodology annex

reglab
center for strategy & regulation

Clickable index

Reglab Methodology Annex

FORMAT: CONTEXT STUDY

Title of the Study	**Global Mapping of Copyright Laws and Exceptions for Data Mining**
Research Question	How do copyright laws around the world currently address exceptions that enable the practice of text and data mining (TDM)?
Methodology Summary	This study is a synthesis of research with an exploratory, comparative, and non-exhaustive character. Data collection relied on documentary analysis techniques, based on Flynn et al. (2022), with data processing conducted through deductive matrix content analysis.
Data Collection	Documentary analysis of legislation from the World Intellectual Property Organization (WIPO) repository. The sample comprised 50 countries, selected as follows: (i) 30 countries with the largest number of internet users (Statista, 2025); and (ii) 20 countries selected through purposive sampling, based on legal diversity, leadership in technological innovation, and explicit references to TDM practices.
Data Analysis	Through content analysis, information was systematized in spreadsheets using standardized codes (e.g., type of copyright exception, legal basis, etc.). Based on these initial codes, countries were classified into three levels of permissiveness (high, medium, and low), defined deductively. This classification was visually represented in a world map, created using the Geographic Information System (GIS) Quantum GIS (QGIS) 3.40.4 for operationalization.
Bias Reduction Procedures	• **Method triangulation:** In addition to research in the WIPO database, exploratory searches were conducted using search engines and AI tools to find supplementary references, such as judicial decisions and academic literature, to confirm categorizations. Any caveats to classifications were noted in the deep dive section. • **Double validation and cross-validation**: At least two researchers reviewed the collected data and categorizations. We thank researcher Isabella Cristina Pereira, who kindly provided data from her ongoing research for cross-validation, ensuring verification through an alternative analytical method.
Other Methodological Limitations	• **Dependence on Open-Access Sources:** The study relied on public databases, news portals, and open-access academic journals, which may limit the breadth of analysis. • **Non-Exhaustive Nature:** The Context series delivers concise, exploratory analyses, subject to methodological delimitations regarding scope and duration. • **Data Collection Period:** The study considered only data available up to July 25, 2025.
Software used	MS Office Suite, ChatGPT, Perplexity, WIPO Lex, The Eckert IV.
Ethical Guidelines	• **Research Funding**: This study is part of the Observatory on Copyright and Technology, an initiative of RegLab supported by AWS, Google, Meta, Microsoft, and YouTube. RegLab retained full editorial control and autonomy in defining the scope, objectives, and methodology. The authors bear sole responsibility for the content and analyses presented in this work. • **Use of Public and Documentary Data**: The research relied exclusively on public and documentary sources, such as national legislation available in the WIPO repository and official country documents. No personal data was collected, nor were interviews with individuals conducted.

reglab

center for strategy
& regulation