

TDM impact study – Legal analysis and use cases

Remy Chavannes, Niels Arnold, Daniël de Weerd (Brinkhof), July 2026

The commercial text and data mining (TDM) exception was created in EU copyright law in 2019 in order to provide legal certainty about the conditions under which such commercial TDM is permitted, and to spur innovation in the European private sector. Now that the exception has been implemented in Member State laws, it is indeed enabling a wide range of use cases across European industries, including pharma, fashion, healthcare and education. Such use cases, and the innovation which they enable, would be under threat if the TDM exception were weakened or abolished.

1 - Introduction

Directive (EU) 2019/790 on Copyright in the Digital Single Market ("**CDSMD**") introduced a harmonized exception for text and data mining ("**TDM**"). TDM was recognized at an early stage as a relevant tool for both researchers and commercial actors. The emergence of large generative AI models developed primarily in the United States, such as ChatGPT and Gemini, has since changed the terms of the debate around this framework.

This brief provides a legal analysis of the TDM exception, and descriptions of use cases deployed by a range of European industries which depend on TDM. Chapter 2 traces the TDM exception from the original 2016 Commission Proposal to the final Article 4 CDSMD, which broadened beneficiaries from research organisations to any lawfully accessing user subject to a rightholder opt-out. Section 2.2 examines criticism that generative AI training may fall outside the scope of the exception and that the opt-out mechanism is ineffective at scale. Section 2.3 explains the argument for retaining the TDM exception from Article 4 CDSMD. Chapter 3 discussed a number of TDM-dependent use cases from major European industry players in sectors ranging from pharmaceutical research to financial services and cyber security. The analysis and use cases are based on public sources.

2 - The text and data mining exception

“Text and data mining” (TDM) is an umbrella term that loosely refers to the processing of large amounts of data with the aim of extracting useful information. In a report commissioned by the European Parliament in 2018, the authors note that the TDM process usually consists of four steps: (1) selecting material to be analyzed, such as works or datasets that have been collected separately or are already organized in a database; (2) copying substantial amounts of material, including pre-processing it by converting it into machine-readable format suitable for the TDM technology used (so that structured data can be extracted) and, depending on the method chosen, uploading this material in a platform; (3) extracting relevant data; and (4) merging and analyzing that data to recognize patterns and arrive at a final result.¹

¹ C. Geiger, G. Frosio & O. Bulayenko, *The Exception for Text and Data Mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market - Legal Aspects*, February 2018, p. 5, https://www.europarl.europa.eu/RegData/etudes/IDAN/2018/604941/IPOL_IDA%282018%29604941_EN.pdf

To provide a basis to enable such processing under EU copyright law, Article 4 CDSMD establishes a general, mandatory exception: Member States shall provide for an exception permitting reproductions of copyrighted works to the extent they are carried out for text and data mining.

2.1 - Background and scope of the exception

EU copyright law – partially harmonized with Directive 2001/29/EC – aims to safeguard works through rightholders' core exclusive rights, in particular the right of reproduction and the right of communication to the public. TDM engages the reproduction right in particular. Accordingly, acts such as copying or storing protected content require the rightholder's authorisation unless an exception applies..

The breadth of the reproduction right under EU law was well established before the introduction of a dedicated TDM exception. In *Infopaq I*, the CJEU held that a data capture process involving the scanning of copyrighted newspaper articles, optical character recognition, and the storing and printing of short text extracts constituted an series of acts of reproduction.² In *InfoPaq II*, the CJEU further held that even temporary reproductions arising in such processes can carry economic significance. These rulings demonstrated that TDM-related copying falls within the scope of the reproduction right.

Academic work has questioned the underlying logic. Strowel argues that the reproduction right has become an "all-encompassing and formalistic notion" that, through its technology-neutral definition, treats analogue and digital copies as equivalent even where the digital copies are purely technology-driven and lack any commercial value. Still, due to this closed system of copyright exceptions and the expansive scope of the reproduction right as interpreted by the CJEU, a new exception was needed to ensure that TDM could be conducted legally.

2.1.1. - Legislative background

Prior to the introduction of the CDSMD and the TDM exception, the EU Commission's Expert Group published a report on text and data mining.³ The report outlines the various fields in which TDM is already applied. Mentioned examples of TDM applications include historians using digital archives to compare how often specific terms appear across different periods; large scale video mining for purposes such as meteorology and police forensics; retail analytics combined with datasets of shoppers' leisure time and health; and medical research that cross-examines unprecedented quantities of neurological and lifestyle data from patient records across multiple countries. The report concluded that TDM was an "important research technique which is certain to become more important" and recommended that the EU "re-consider and reform the EU's legal framework" to facilitate the use of TDM tools.⁴ A later report added that the Commission was "investigating

² Court of Justice of the European Union, Case C-302/10 (*Infopaq I*).

³ European Commission (Expert Group), *Standardisation in the area of innovation and technological development, notably in the field of Text and Data Mining: Report from the Expert Group*, April 2014, https://pure.uva.nl/ws/files/2478182/157208_TDM_report_from_the_expert_group_042014.pdf.

⁴ European Commission (Expert Group), *Standardisation in the area of innovation and technological development, notably in the field of Text and Data Mining: Report from the Expert Group*, April 2014, https://pure.uva.nl/ws/files/2478182/157208_TDM_report_from_the_expert_group_042014.pdf, p.65.

ways in which data-driven innovation based on data-mining, including text-mining, might be enhanced” and proposed reviewing copyright rules with the overall aim of stimulating a “thriving data-driven economy”.⁵

On 14 September 2016, the European Commission presented its first proposal (the “**Proposal**”) for the CDSMD, which included an exception for text and data mining.⁶ In the Proposal, TDM is defined as “any automated analytical technique aiming to analyse text and data in digital form in order to generate information such as patterns, trends and correlations.”⁷ The TDM exception is set out in Article 3 and requires Member States to introduce an exception to the relevant reproduction and database rights for reproductions and extractions made by research organisations, to the extent necessary to carry out TDM on material to which they have lawful access. It solely applies for the purpose of scientific research.

In the drafting process that followed, key stakeholders proposed to expand the scope of the exception. On 21 April 2017, months after the publication of the Proposal, the European Economic and Social Committee (“**EESC**”) published its opinion on the Proposal and recommended specific amendments to the TDM framework. The EESC argued that in order to foster innovative enterprises, the scope of application of TDM should include researchers and businesses that operate for profit.⁸ In addition, the EESC recommended codifying the principle that mere facts and data are not protected by copyright, and therefore should be explicitly reflected in the text of the Proposal. In its opinion, the EESC refers to the The Hague Declaration (2014), which highlights the enormous potential of TDM for innovation and research. The EESC describes TDM as an important tool for researchers, small and medium-sized enterprises, and large technology companies, but it is underutilized in the EU due to legal, technological, and contractual constraints.⁹ Several Member States subsequently shared opinions providing feedback on the Proposal. For example, the Romanian opinion recommended that text and data mining for research purposes should also be made possible for commercial entities, particularly SMEs, in order to strengthen the EU’s competitiveness.¹⁰

A joint proposal by several Member States also argued for a more permissive exception. In this proposal submitted by The Netherlands, Belgium, Germany, Finland and the Slovak Republic, these Member States argue that the initial Proposal would create legal uncertainty, because TDM for other purposes than scientific research might require additional licenses for activities to continue. The Member States note that TDM is already widely used beyond scientific research—for example in data analytics, machine learning, and autonomous driving. Limiting the TDM exception to scientific research could require these actors to renegotiate existing contracts or to identify and approach often unknown rightholders in relation to online

⁵ European Commission, *Towards a thriving data-driven economy*, COM 2014/0442 final, <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52014DC0442>.

⁶ European Commission, *Proposal for a Directive of the European Parliament and of the Council on copyright in the Digital Single Market*, COM(2016) 593 final, 14 September 2016, <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52016PC0593>.

⁷ Article 2(2) COM(2016) 593 final, 14 September 2016, <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52016PC0593>.

⁸ European Economic and Social Committee (EESC), *Opinion on the Proposal for a Directive on Copyright in the Digital Single Market (COM(2016) 593 final) and related proposals*, C 125/27, 21.4.2017, CELEX:52016AE5382, <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52016AE5382>, paragraph 1.7.

⁹ European Economic and Social Committee (EESC), *Opinion on the Proposal for a Directive on Copyright in the Digital Single Market (COM(2016) 593 final) and related proposals*, C 125/27, 21.4.2017, CELEX:52016AE5382, <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52016AE5382>, paragraph 5.3.

¹⁰ Council of the European Union, *Cover note — Opinion of the Chamber of Deputies of Romania on the proposal for a Directive on copyright in the Digital Single Market (COM(2016) 593)*, Council doc. ST 8995/17 INIT, 10 May 2017, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=consil%3AST_8995_2017_INIT.

content. This would constitute an onerous burden that startups, individuals, and public sector organisations are not equipped to handle. As a result, the Member States argue that the exception should be broadened.¹¹

In response to these proposals, an additional TDM exception was introduced. Although the revised proposal of 26 September 2017 did not yet include a broader TDM exception, the proposal published on 13 December 2017 introduced a new Article 3a, providing an optional TDM exception extending beyond scientific research.¹² In the attached explanatory note, the presidency of the council of the EU stated that a number of delegations supported the overall approach of the proposed new exception, subject to further textual amendments, whereas a small number of delegations expressed their opposition to the new optional exception.¹³ Subsequent discussions moved to the scope of this new optional exception for TDM, introduced as Article 3a to complement the mandatory exception for non-profit TDM in Article 3. The optional TDM exception covered temporary reproductions and extractions of works and other subject matter that are lawfully accessible, insofar as these acts form part of the TDM process.

On 23 April 2018, the Council recorded that, following the 17 December 2017 compromise draft, no further amendments were introduced to Article 3a. Although several delegations sought to narrow the optional TDM exception (e.g., by limiting it to temporary copies of content made freely available online), a majority of delegations opposed any further reduction of its scope.¹⁴

In stakeholder documents, published during the EU's final negotiations on the text of the CDSMD, stakeholders made clear that the (at the time) optional "commercial" TDM exception would lead to fragmentation, and therefore it should be made mandatory in order to create an EU-wide baseline.¹⁵ In addition, the signatories emphasised that a mandatory commercial TDM exception is necessary to ensure a thriving European research and innovation ecosystem. In the presidency's Council note of 16 November 2018, it is explicitly stated that the Commission requested, in the trilogues, that Article 3a be changed from an optional to a mandatory provision ("to turn Article 3a into a mandatory provision").¹⁶ In the proposal of 20 February 2019, the TDM exception of Article 3a—previously discussed as an optional provision—was for the first time presented as mandatory for Member States to implement.¹⁷ The same pattern is apparent in the recitals: the recital

¹¹ Council of the European Union, *Proposal of the Netherlands, Belgium, Germany, Finland and the Slovak Republic for an additional and optional exception for text and data mining* (Council doc. WK 12354/2017 INIT), https://www.knowledgerights21.org/wp-content/uploads/WK-12354-2017-INIT_en.pdf.

¹² Council of the European Union, *Proposal for a Directive of the European Parliament and of the Council on copyright in the Digital Single Market – Presidency compromise proposal (consolidated version) and state of play*, doc. ST 15651/17 INIT, 13 December 2017 (procedure 2016/0280 (COD)), https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=consil%3AST_15651_2017_INIT.

¹³ Council of the European Union, *Proposal for a Directive of the European Parliament and of the Council on copyright in the Digital Single Market – Presidency compromise proposal (consolidated version) and state of play*, doc. ST 15651/17 INIT, 13 December 2017 (procedure 2016/0280 (COD)), https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=consil%3AST_15651_2017_INIT, p. 4.

¹⁴ Council of the European Union, *Proposal for a Directive of the European Parliament and of the Council on Copyright in the Digital Single Market — Mandate for negotiations with the European Parliament*, doc. 8145/18, 23 April 2018, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=consil%3AST_8145_2018_INIT.

¹⁵ (Open Letter), Brussels, 13 November 2018, signed by AFNUM et al., chrome-extension://efaidnbmnnnibpccjpcglclefindmkaj/https://ia801803.us.archive.org/32/items/final-open-letter-mandatory-tdm-exception-copyright-trilogues-13november/final-open-letter-mandatory-tdm-exception-copyright-trilogues-13november.pdf?utm_source=chatgpt.com.

¹⁶ Council of the European Union (Presidency), *Note to the Permanent Representatives Committee: Proposal for a Directive of the European Parliament and of the Council on Copyright in the Digital Single Market – Preparation for the trilogue*, doc. 14334/18, Brussels, 16 November 2018 (LIMITE), Interinstitutional File 2016/0280(COD), chrome-extension://efaidnbmnnnibpccjpcglclefindmkaj/<https://data.consilium.europa.eu/doc/document/ST-14334-2018-INIT/en/pdf>.

¹⁷ Council of the European Union (General Secretariat), *Note to the Permanent Representatives Committee (Coreper) on the "Proposal for a Directive of the European Parliament and of the Council on Copyright in the Digital Single Market — Outcome of proceedings"*, doc.

accompanying the exception is also significantly broadened in the 20 February 2019 text compared with the 26 September 2018 version¹⁸.

t. In subsequent versions following the proposal of 20 February, the provision remained largely unchanged, save for minor drafting and structural adjustments (including renumbering and stylistic clarifications). It ended up as Article 4 of the final version.

2.1.2 - Scope

Under Article 4 CDSMD, Member States shall provide for an exception regarding copyright and related rights and database rights allowing anyone to reproduce protected works for the purposes of text and data mining, provided that the works are lawfully accessible. The definition of “text and data mining” remains unchanged compared to the original proposal, being defined as “any automated analytical technique aimed at analysing text and data in digital form in order to generate information which includes but is not limited to patterns, trends and correlations”. The exception is subject to reservations of rights (“opt-outs”) by rightsholders, who must express such reservations “in an appropriate manner”. For publicly available content, the provision mentions “machine-readable means” such as metadata and terms and conditions of a website or a service as an appropriate manner of opting out.¹⁹

Recital 18 of the CDSMD clarifies the rationale behind Article 4. It emphasises that TDM is not limited to scientific research, but is widely used by public and private actors for purposes such as public services, business decision-making and the development of new technologies. At the same time, it recognises legal uncertainty under the existing framework, particularly because the technical reproductions and extractions inherent in TDM may not always meet the conditions of the temporary copying exception in Article 5(1) of the InfoSoc Directive. Article 4 is therefore intended to provide legal certainty and to support innovation, including in the private sector.

2.1.3 - Generative AI training and TDM

Since the entry of LLMs onto the global stage, there has been considerable debate on whether the TDM exception covers the use of copyrighted material to train generative AI models. The training of generative AI involves ‘feeding’ it books, articles, images, and other creative works so it can learn patterns and produce new text or images based on those patterns.²⁰ This process happens on a larger scale and for a broader set of purposes than was typical for TDM when the CDSMD was adopted.

This has led some to argue that the TDM exception does not cover such training. Certain scholars argue that the required data collection – typically done using large-scale web crawling – is not an “analytical technique”

6637/19, Brussels, 20 February 2019 (Interinstitutional File 2016/0280(COD)), https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=consil%3AST_6637_2019_INIT.

¹⁸ Council of the European Union, *Proposal for a Directive of the European Parliament and of the Council on copyright in the Digital Single Market – Preparation for the trilogue – 4-column document*, ST 12513/18 INIT, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=consil%3AST_12513_2018_INIT.

¹⁹ Recital 18 CDSMD.

²⁰ Lucchi (2025), 26; Geiger (2024b), 1151; Sag (2023), 300; Gervais (2020), 2055-2059; Ginsburg/Budiardjo (2019), 401-402; Deltorn (2018), 173-174.

as defined in Article 2(2) CDSMD, but rather a preparatory step that precedes any extraction of information.²¹ Dornis argues that the data analysis part of TDM is not covered by copyright protection because of the idea/expression dichotomy and that the exception therefore only considers technical copying steps - such as scraping and downloading - that precede the data analysis. For this reason, he argues that uses going *beyond* data analysis by using the work's expressive features, such as AI training, are not covered by the TDM exception.

The Artificial Intelligence Act ("**AIA**") adopted in 2024 essentially refutes this line of argument, and confirms that the TDM exception extends to the training of GenAI models. Recital 105 AIA notes that "*the development and training of such models require access to vast amounts of text, images, videos and other data*" and that "*text and data mining techniques may be used extensively in this context for the retrieval and analysis of such content*". Article 53(1)(c) AIA further requires providers of general-purpose AI models to draw up a policy to comply with EU copyright law, which in particular should address the methods they use to identify and comply with opt-outs. The measures in the accompanying (non-binding) Code of Practice regarding copyright further require model providers to "reproduce and extract only lawfully accessible copyright-protected content" and "identify and comply with rights reservations" when crawling the web. Interpreting Article 4 CDSMD as not covering web crawling would be difficult to reconcile with this regime, which expressly presupposes that the text and data mining exception applies to the web crawling context.²²

2.2 - Critical views on the current TDM exception

In addition to the legal/doctrinal argument that the current TDM exception *does* not cover GenAI training, many rightholders, scholars and politicians have argued that the exception *should* not cover such training. Arguments include that typical generative AI use cases often circumvent traditional licensing mechanisms, increasing the risk of "systematic interference with exploitation markets" for certain categories of works, and especially in music, visual art, and journalism.²³ They further argue that the opt-out mechanism theoretically protects the legitimate interests of rightholders, but in practice fails to do so.²⁴

This criticism has been well-received in at least parts of the European Parliament. Axel Voss, a prominent MEP with a key role in drafting both the CDSMD and the AIA, argued in a 2025 Guardian interview that the CDSMD "was not conceived to deal with generative AI models" and that the AIA had opened up a "legal gap" by allowing general-purpose AI model providers to train their models on large amounts of copyrighted material.²⁵ He framed the issue as one where U.S. technology companies profit off the efforts of European creatives, stating that "what I do not understand is that we are supporting big tech instead of protecting European creative ideas and content". In mid-2025, Voss published an own-initiative draft report in which he expressed his "alarm" at generative AI products and called upon the Commission to "to immediately impose a

²¹ Dornis, Tim W., The Training of Generative AI Is Not Text and Data Mining (October 19, 2024). Available at SSRN: <https://ssrn.com/abstract=4993782> or <http://dx.doi.org/10.2139/ssrn.4993782>.

²² In a copyright infringement action against OpenAI, a German court has further confirmed that the TDM exception applies to the training of generative AI models. See: Landgericht München 11 November 2025, Case No. 42 O 14139/24 (GEMA/OpenAI), p. 41-42.

²³ M. Senftleben, 'Copyright Data Improvement for AI Licensing – The Role of Content Moderation and Text And Data Mining Rules' (2024), <https://ssrn.com/abstract=4817796>.

²⁴ Ibid & M. Manteghi, 'The insufficiency of the EU's text and data mining exceptions for using artificial intelligence' (2022) 44(11) EIPR 651, <https://www.utupub.fi/handle/10024/186469>.

²⁵ J. Rankin, 'EU accused of leaving "devastating" copyright loophole in AI Act', The Guardian, 19 February 2025, <https://www.theguardian.com/technology/2025/feb/19/eu-accused-of-leaving-devastating-copyright-loop-hole-in-ai-act>

remuneration obligation on providers of general-purpose AI models and systems in respect of the novel use of content protected by copyright”.²⁶ This report was adopted by the full JURI committee in early 2026, arguing that the TDM exception does not cover general-purpose AI model training and further proposing “clarification” of the exception’s “machine readable” and “lawful access” criteria, the establishment of a central opt-out registry and more comprehensive transparency requirements for AI model providers.²⁷

A similar argument was advanced by a JURI-commissioned study from July 2025, which argued that the TDM exception “was not designed to accommodate the expressive and synthetic nature of generative AI training, and its application to such systems risks distorting the purpose and limits of EU copyright exceptions”.²⁸

2.3 - The argument for an enduring TDM exception

The criticism outlined in section 2.2 should be assessed against the legislative history of the TDM exception. The scope of the exception was deliberately broadened over several years of negotiation. The original 2016 Commission Proposal limited it to research organisations acting for scientific purposes. It was the EESC, followed by a coalition of Member States including the Netherlands, Belgium, Germany, Finland and the Slovak Republic, that argued for extending TDM to commercial actors, on the basis that a narrow exception would create legal uncertainty, force parties to renegotiate contracts or identify unknown rightholders, and impose onerous burdens on startups, individuals, and public sector organisations. The broadening from Article 3a (optional) to Article 4 (mandatory) was a deliberate policy choice, not an oversight or a concession to large technology companies.

It might be true that U.S.-based general-purpose AI developers rely on the TDM exception in a way that was not foreseen by EU lawmakers. This however does not negate the benefit of the exception to EU-based companies. As chapter 3 illustrates, European companies across sectors ranging from pharmaceuticals to financial services and automotive benefit from the ability to process copyrighted content at scale without obtaining individual licences. In many of these cases, a comprehensive collective licensing mechanism does not exist. Without a TDM exception, European innovators would face the structural barriers identified in 2017: having to identify and approach often unknown rightholders for online content, with prohibitive costs for startups, individuals, and public sector organisations. Innovators would often be unable to (fully) overcome these barriers, leaving them with limited access to relevant data and putting them at a disadvantage to their non-EU counterparts.

Therefore, the concerns about generative AI training do not make it necessary or desirable to narrow or repeal the TDM exception. If the concern is that the opt-out is insufficiently effective in practice, the focus should be on improving its implementation, for instance by developing clearer technical standards for machine-readable reservations or by establishing monitoring mechanisms. The next chapter aims to show that restricting the

²⁶ European Parliament, JURI Committee, Draft Report 2025/2058(INI), Rapporteur: Axel Voss, 27 June 2025, https://www.europarl.europa.eu/meetdocs/2024_2029/plmrep/COMMITTEES/JURI/DV/2025/07-15/1323931EN.pdf

²⁷ See EUobserver, ‘Artists and songwriters need updated EU copyright law for AI, parliament committee says’, January 2026, <https://euobserver.com/200672/artists-and-songwriters-need-updated-eu-copyright-law-for-ai-parliament-committee-says/>

²⁸ N. Lucchi, Generative AI and Copyright: Training, Creation, Regulation, European Parliament, Policy Department for Justice, Civil Liberties and Institutional Affairs, Study PE 774.095 (July 2025), [https://www.europarl.europa.eu/thinktank/en/document/IUST_STU\(2025\)774095](https://www.europarl.europa.eu/thinktank/en/document/IUST_STU(2025)774095).

exception itself would deprive European innovators of a tool that the legislator determined was necessary for EU competitiveness.

3 - Sector-specific use cases for the TDM exception

This chapter outlines a number of use cases in which major European industry players benefit from the TDM exception by having flexible access to online content, in sectors ranging from pharmaceutical research to financial services and cyber security. Use case descriptions are based on online research and are not intended as authoritative or exhaustive accounts of practices in the relevant industry.

3.1 - Pharmaceutical literature mining

Pharmaceutical companies such as **Bayer** (Germany), **Sanofi** (France) and **Novo Nordisk** (Denmark) use AI and automated natural language processing (“NLP”) techniques to extract knowledge from biomedical literature, aiding them with drug discovery. As millions of biomedical publications from different publishers exist, great value lies in mining broad literature corporas to identify associations between genes and diseases, molecules in the body that drugs can interact with (“drug targets”) and other biological mechanisms.

Several key EU players use such approaches to advance innovation. Bayer’s public materials describe AI-supported target identification and NLP/LLM use in medical coding.²⁹ Researchers working for Novo Nordisk have published on how systematically mining scientific literature can help them identify new drug targets for type 2 diabetes, making them aware of breakthroughs without having to study all publications by hand.³⁰ Sanofi developed a similar approach by systematically mapping relationships between human genes, diseases, cell types and drugs in the full PubMed abstract corpus.³¹ Novo Nordisk further partnered with NVIDIA to build customized AI models and agents, including generative AI models enabling drug discovery.³²

A notable example illustrates the practical significance of this approach. In January 2020, BenevolentAI (UK) used AI-driven mining of biomedical literature to identify baricitinib as a potential COVID-19 treatment.³³ Baricitinib went on to become the first immunomodulatory treatment for COVID-19 to receive full FDA approval, and was subsequently recommended by the WHO. This demonstrates that NLP-based extraction from copyrighted scientific literature for drug target identification can produce meaningful therapeutic outcomes.

Why is the TDM exception needed? Biomedical literature is published by thousands of publishers, and no single licence or collective licensing mechanism covers the full corpus. The value of

²⁹ Bayer, “Artificial intelligence in pharma”, <https://www.bayer.com/en/pharma/artificial-intelligence>.

³⁰ L.K. Hansson, R.B. Hansen et al., “Semantic text mining in early drug discovery for type 2 diabetes,” PLOS ONE, 2020, <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0233956>.

³¹ Dongyu Liu, Cora Ames, Khader Shameer & Franck Rapaport, “SciLinker: a large-scale text mining framework for mapping associations among biological entities”, Frontiers in Artificial Intelligence, 2025, <https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2025.1528562/full>.

³² NVIDIA, “NVIDIA Partners With Novo Nordisk and DCAI to Advance Drug Discovery,” June 2025, <https://nvidianews.nvidia.com/news/nvidia-partners-with-novo-nordisk-and-dcai-to-advance-drug-discovery>.

³³ Peter J Richardson et al., “The AI-Assisted Identification and Clinical Efficacy of Baricitinib in the Treatment of COVID-19”, <https://pubmed.ncbi.nlm.nih.gov/35746559/>.

comprehensive literature mining lies precisely in its comprehensiveness: it enables the connection of insights scattered across papers from different publishers. If pharmaceutical companies had access to only a licensed subset of this data, this would leave gaps and lead to the risk that crucial relationships between findings go undetected, slowing the development of innovative medicines.

3.2 - Machine translation training

Neural machine translation service **DeepL** (Germany) allows individual users and professionals to quickly translate documents. The model used by DeepL was trained on approximately 1 billion bilingual sentence pairs collected from multilingual corporate websites, news sites, and other web content.³⁴

Why is the TDM exception needed? The web crawling that produced the training corpus scraped copyrighted content from millions of websites. Individual licensing from millions of website operators is impossible. No collective licensing mechanism exists for multilingual web content. Without the TDM exception, languages less-resourced language pairs and specialized domains would suffer from reduced training data diversity.

3.3 - Fashion trend prediction

Fashion data services provided by **Heuritech/Luxurysight** (France) allow brands such as LVMH, Prada, Moncler, and Adidas to use social media intelligence to forecast fashion trends. Major EU retailer **H&M** (Sweden) similarly uses social media intelligence to varying degrees. Predicting which colors, patterns, fabrics, and silhouettes will gain or lose popularity enables brands to adjust production volumes before committing to manufacturing, reducing overproduction and textile waste.

Heuritech monitors approximately 50,000 Instagram accounts per geography and approximately 400,000 Weibo accounts and maintains a database of roughly 500 million posts.³⁵ Its computer vision system analyses 3 million images per day and recognizes over 3000 fashion details, such as colors, textures, patterns and silhouettes, per image.³⁶ H&M reportedly analyzes keywords, hashtags, and image content on platforms including Instagram and TikTok as part of its AI-driven demand forecasting, alongside internal sales data and customer behavior.³⁷

Why is the TDM exception needed? Social media images are copyrighted by the individual users who post them. Licensing from millions of individual rightsholders is impossible, and no collective licensing body represents social media users' copyright in their photographs. For companies that scrape social media at scale (such as Heuritech), the TDM exception is the only plausible copyright basis for this

³⁴ DeepL, "How does DeepL work?," <https://www.deepl.com/en/blog/how-does-deepl-work>; TechCrunch, "DeepL schools other online translators with clever machine learning", <https://techcrunch.com/2017/08/29/deepl-schools-other-online-translators-with-clever-machine-learning/>

³⁵ Heuritech, "Trend Forecasting Based On Social Media: How Do We Do It?," <https://heuritech.com/articles/social-media-based-trend-forecasting/>.

³⁶ Heuritech, "About Heuritech," <https://heuritech.com/company-about-us/>.

³⁷ H&M Using AI: A Case Study," DigitalDefynd, 2025, <https://digitaldefynd.com/IQ/hm-using-ai-case-study/>; "H&M's AI: Conquering Trends & Inventory Chaos," Reruption, <https://reruption.com/en/knowledge/industry-cases/h-ms-ai-conquering-trends-inventory-chaos>.

activity. Without access to this data, EU companies are at a disadvantage- they would have less fine-grained insights in fashion trends, leading to less revenue, more overproduction and more waste.

3.4 - Alternative data in financial services

Large EU asset managers such as **Amundi** (France), **DWS** (Germany), and **Robeco** (Netherlands) use so-called “alternative data” – non-traditional data sources collected for financial analysis – to gain an information edge in financial markets. Global surveys indicate that three-quarters of buy-side investing firms use such data and most expect to increase their spending on alternative data in the next year.³⁸ **SEAMm** (France) is a key example of an alternative data provider.

Key types of alternative data include news and social media sentiment, web-scraped pricing data, and scraped job-posting data. Examples include Paris-based SESAMm, which analyses over 20 billion web articles from 4 million sources to deliver NLP-derived sentiment and ESG analytics and Causality Link (a U.S. company with French roots), which analyses over 100 million texts in 27 languages from 8,000+ newspaper sources to identify causal relationships between economic events.³⁹ On the asset management size, Robeco has integrated NLP signals into its quantitative strategies since 2019, using algorithms to “detect themes within a vast amount of alternative data, including company earnings calls, news articles and management interviews”.⁴⁰

ESG data mining is a specific form of alternative data which helps companies comply with the EU Sustainable Finance Disclosure Regulation (SFDR). For example, Dutch pension managers APG and PGGM jointly developed an AI platform that classifies approximately 10,000 listed investments against UN Sustainable Development Goals.⁴¹

Why is the TDM exception needed? In finance, every missing bit of information is a competitive disadvantage. Alternative data includes large amounts of unstructured, copyrighted text. Licensing markets for such text only cover a fraction of the total amount of available data. Without access to this data, EU asset managers are at a competitive disadvantage to their U.S. counterparts.

3.5 - Clinical AI and medical documentation in healthcare

European digital health companies such as **Corti** (Denmark), **Nabla** (France) and **Ada Health** (Germany) develop AI systems that support healthcare professionals and patients with clinical documentation, decision

³⁸ Grand View Research, “Alternative Data Market (2025 - 2030)”, <https://www.grandviewresearch.com/industry-analysis/alternative-data-market>
<https://www.hedgeweek.com/hedge-fund-alt-data-spending-set-to-surge-says-new-research/>

³⁹ Finovate, “Natural Language Analytics Innovator SESAMm Locks In \$37 Million in New Funding,” 2023, <https://finovate.com/natural-language-analytics-innovator-sesamm-locks-in-37-million-in-new-funding/>
 Causality Link, “Third-Generation Platform,” <https://causalitylink.com/causality-link-launches-third-generation-of-its-ai-driven-financial-research-platform>.

⁴⁰ “Robeco harnesses AI to elevate investment strategies,” SCMP/Robeco, 2025, <https://www.scmp.com/presented/business/banking-finance/topics/harnessing-ai-investing/article/3333049/robeco-harnesses-ai-elevate-investment-strategies-builds-25-years-quantitative-expertise>

⁴¹ PGGM, “APG and PGGM develop AI-powered platform for investing in the UN Sustainable Development Goals,” <https://www.pggm.nl/en/press/apg-and-pggm-develop-ai-powered-platform-for-investing-in-the-un-sustainable-development-goals>.

support and symptom assessment. At a time when EU health systems face structural staff shortages, these tools reduce the administrative burden on clinicians and help patients navigate to the right level of care. To perform reliably, however, they must master specialist medical language in each of the languages in which they operate, which requires training on large volumes of medical text.

Corti builds healthcare-specific foundation models that power medical transcription, ambient documentation and real-time clinical decision support. Its foundation model has been trained on more than 250,000 patient interactions,⁴² and the company states that its models learn from real-life patient interactions, peer-reviewed papers and regulatory-compliant medical data.⁴³ Nabla offers an ambient AI assistant that listens to consultations and automatically generates structured clinical notes across more than 55 specialties and 35 languages. It relies on speech recognition models fine-tuned on medical data (building on the open-source Whisper model) and has worked with fine-tuned versions of open foundation models such as Meta's Llama.⁴⁴ Berlin-based Ada Health provides an AI symptom assessment app used by tens of millions of people, combining probabilistic reasoning over a medical knowledge base with machine learning models that interpret users' free-text symptom descriptions.⁴⁵ Developing such systems requires processing large corpora of specialist text, including peer-reviewed biomedical literature, clinical guidelines published by professional societies, drug information, medical reference works, terminology resources and health-related web content.

Why is the TDM exception needed? The medical text needed to train clinical language and speech models is spread across thousands of publishers, professional societies and health organisations, and no single licence or collective licensing mechanism covers the full corpus. For smaller EU languages such as Danish or Dutch, suitable medical text is scarce to begin with; restricting the ability to mine it would make it significantly harder to build clinically reliable AI in those languages, favouring English-centric models from non-EU providers. Without the TDM exception, European healthcare AI companies would struggle to develop models tuned to European languages, clinical practice and regulatory requirements, at a direct cost to already overburdened EU health systems.

3.6 - AI tutoring and academic integrity in education

European educational technology companies such as **Knowunity** (Germany), **StudySmarter** (Germany) and **Compilatio** (France) rely on the large-scale processing of text to deliver personalised learning tools and to safeguard academic integrity. These applications depend on access to broad corpora of existing content: AI tutors must be able to explain any topic in the language and curriculum of the student, while plagiarism detection is only as good as the body of texts against which submissions are compared.

Berlin-based Knowunity operates an AI-powered study companion used by more than 20 million students across 15 countries; in Germany, one in three school students uses the app.⁴⁶ Knowunity partnered with

⁴² R. Zamani, "Corti launches real-time AI clinical assistant", Medical Device Network, December 2024, <https://www.medicaldevice-network.com/news/corti-launches-real-time-ai-clinical-assistant/>.

⁴³ Corti, "The AI that Healthcare deserves: Welcome to the next chapter of Healthcare AI", <https://www.corti.ai/stories/foundation-models-launch>.

⁴⁴ R. Dillet, "Nabla raises another \$24 million for its AI assistant for doctors that automatically writes clinical notes", TechCrunch, 5 January 2024, <https://techcrunch.com/2024/01/05/nabla-raises-another-24-million-for-its-ai-assistant-for-doctors/>.

⁴⁵ Ada Health, "Comparing AI for application in healthcare", <https://about.ada.com/editorial/what-llms-mean-for-healthcare/>.

⁴⁶ D. Cendón, "German EdTech startup Knowunity raises €27 million to bring AI tutor to 1 billion students", EU-Startups, 13 June 2025, <https://www.eu-startups.com/2025/06/german-edtech-startup-knowunity-raises-e27-million-to-bring-ai-tutor-to-1-billion-students/>.

French AI developer Mistral AI to fine-tune large language models for its AI tutor, which delivers curriculum-aligned explanations.⁴⁷ According to the company, the tutor grounds its answers in a corpus of over 3 million study materials localised to national curricula.⁴⁸ Munich-based StudySmarter, which reports over 40 million users worldwide, similarly uses AI to automatically generate summaries, flashcards, quizzes and personalised study plans from uploaded study materials.⁴⁹

On the academic integrity side, European providers such as Compilatio (France) and StrikePlagiarism (Poland) supply plagiarism detection software to universities. Compilatio, operating since 2005, is reported to be the most used plagiarism detection system among universities in France.⁵⁰ Its software compares submitted documents against billions of sources, including the global web, scientific publications and documents previously submitted by partner institutions.⁵¹ Poland-based StrikePlagiarism reports cooperating with more than 700 universities and compares submissions against scientific journals, open access repositories and documents submitted by participating institutions.⁵² Building and continuously updating the underlying reference indexes requires crawling, copying and indexing very large volumes of copyrighted web content and publications, and every similarity check involves comparing the submission against reproductions of those sources.

Why is the TDM exception needed? Plagiarism detection is only effective if the reference database approximates the full universe of texts a student might copy from; licensing that universe, which spans millions of websites and publications, is impossible, and no collective licensing mechanism covers it. AI tutoring tools must be adapted using educational content in every EU language and school system. Without the TDM exception, the tools that European universities rely on to protect academic integrity would rest on an uncertain legal footing, and affordable, personalised tutoring in smaller EU languages and national curricula would become significantly harder to build, to the benefit of English-centric offerings from non-EU providers.

3.7 - Cybersecurity and external threat intelligence

European cybersecurity companies such as CybelAngel and Sekoia.io (both France) protect organisations by continuously monitoring sources outside the corporate network: the open web, connected devices and cloud storage, code repositories, paste sites and dark web forums. There they detect leaked documents, compromised credentials and preparations for attacks before these can be exploited. The effectiveness of such services depends directly on the breadth of coverage: an exposure that is not scanned is an exposure that is not detected.

⁴⁷ Mistral AI, “Knowunity offers a daily AI companion powered by Mistral.”, <https://mistral.ai/customers/knowunity/>

⁴⁸ D. Cendón, “German EdTech startup Knowunity raises €27 million to bring AI tutor to 1 billion students”, EU-Startups, 13 June 2025, <https://www.eu-startups.com/2025/06/german-edtech-startup-knowunity-raises-e27-million-to-bring-ai-tutor-to-1-billion-students/>.

⁴⁹ StudySmarter, “For Companies”, <https://www.studysmarter.co.uk/for-companies/>; StudySmarter, “StudySmarter AI”, <https://www.studysmarter.co.uk/ai/>.

⁵⁰ M. Abdelhamid, S. Batata & F. Azouaou, “A Survey of Plagiarism Detection Systems: Case of Use with English, French and Arabic Languages”, arXiv:2201.03423, January 2022, <https://arxiv.org/abs/2201.03423>.

⁵¹ Compilatio, “Compilatio Magister: Academic Plagiarism Checker for teachers”, <https://www.compilatio.net/en/magister>.

⁵² StrikePlagiarism, “Anti-plagiarism system”, <https://strikeplagiarism.com/>; M. Abdelhamid, S. Batata & F. Azouaou, “A Survey of Plagiarism Detection Systems: Case of Use with English, French and Arabic Languages”, arXiv:2201.03423, January 2022, <https://arxiv.org/abs/2201.03423>.

Paris-based CybelAngel performs daily scans of the internet to map its clients' external attack surface. The company states that its systems process billions of data points every day and that its dark web monitoring analyses millions of posts in over 200 languages; its clients include Sanofi, Amadeus and Signify.⁵³ The public value of this scanning capability is illustrated by a 2020 CybelAngel investigation which, after scanning approximately 4.3 billion IP addresses, found more than 45 million medical images freely accessible on unprotected servers across 67 countries, allowing the exposures to be reported and remediated.⁵⁴

Sekoia.io operates an AI-driven security operations platform built on a proprietary cyber threat intelligence dataset covering both cybercriminal groups and state espionage; its clients include EDF, Vinci, SNCF and the French Ministry of Armed Forces, and the platform detected 4 million real threats in 2024.⁵⁵ Producing such threat intelligence requires continuously collecting and machine-analysing content from a vast range of external sources, including security blogs and vendor reports, hacker forum posts, marketplace listings, paste sites, and leaked documents and code, much of which is protected by copyright regardless of the context in which it appears. Demand for these monitoring capabilities is reinforced by EU law itself, as the NIS2 Directive requires a broad range of European organisations to manage their cybersecurity risks.⁵⁶

Why is the TDM exception needed? Threat intelligence depends on comprehensively copying and analysing third-party content from precisely those corners of the internet where consent will never be given. Cybercriminals do not license their forum posts, and stolen data must be detected wherever it surfaces. Even legitimate sources such as news sites, technical blogs and security advisories are far too numerous and dispersed to license individually, and no collective licensing mechanism covers them. Without the TDM exception, European security providers would face legal uncertainty over the web-scale scanning at the core of their services, even as EU legislation requires organisations to strengthen their cyber resilience. This would weaken the EU's cybersecurity posture and favour non-EU providers operating under more permissive copyright regimes.

⁵³ CybelAngel, "The All-in-One Solution Against External Threats", <https://cybelangel.com/>.

⁵⁴ CybelAngel, "45 Million Unprotected Medical Images Accessible Online", 15 December 2020, <https://cybelangel.com/blog/medical-data-leaks/>.

⁵⁵ Sekoia.io, "Sekoia.io Secures €26 Million in Series B to Democratize Cyber Operations with AI and Cyber Intelligence", 9 April 2025, <https://www.sekoia.io/en/presse/sekoia-io-secures-e26-million-in-series-b-to-democratize-cyber-operations-with-ai-and-cyber-intelligence/>.

⁵⁶ Directive (EU) 2022/2555 of the European Parliament and of the Council of 14 December 2022 on measures for a high common level of cybersecurity across the Union (NIS 2 Directive), <https://eur-lex.europa.eu/eli/dir/2022/2555/oj>.