



FACULTY OF LAW

LUND UNIVERSITY

Zhala Safaraliyeva

Copyright in the Age of Generative AI: Rethinking the EU Approach to Text and Data Mining

JAEM03 Master Thesis

European Business Law

30 higher education credits

Supervisor: Petra Gyöngyi

Term of graduation: Spring 2026

Table of Contents

Summary	3
Sammanfattning	4
Preface	5
Abbreviation	6
1 Introduction	8
1.1 Background	8
1.2 Research Aim and Research Questions.....	10
1.3 Methodology and Sources.....	11
1.4 Scope and Delimitations.....	12
1.5 Outline.....	13
2 The Role of TDM in AI Training and Its Copyright Implications under EU Law	15
2.1 Text and Data Mining: Definition and Procedure	15
2.2 The Role of TDM in the Training of GenAI Systems.....	16
2.3 TDM and Copyright Law.....	18
2.4 TDM under the EU Legal Framework before the CDSM Directive	19
2.5 The Regulation of TDM under the CDSM Directive.....	20
3 Critique of the TDM Framework under the CDSM Directive.....	24
3.1 Member States' National TDM Frameworks.....	24
3.2 National Implementation and the Limits of Harmonisation.....	25
3.3 A Critique of Article 3 of the CDSM Directive	27
3.3.1 The Limited Scope of Beneficiaries under Article 3.....	28
3.3.2 Restrictions on “Scientific Research Purposes”	29
3.3.3 The Impact of Technological Protection Measures on TDM...	29
3.3.4 Legal Uncertainty Surrounding “Lawful Access”	30
3.4 Critique of Article 4 and the Opt-Out Mechanism.....	32
3.4.1 Challenges Surrounding the Concept of “Machine Readability”	32
3.4.2 Practical Problems of Robots.txt as an Opt-Out Tool.....	34
3.4.3 Legal Issues Related to the Proof of Opt-Outs.....	35
3.4.4 <i>GEMA v OpenAI</i> and the Limits of the TDM Exception	36
3.5 The Three-Step Test and Text and Data Mining.....	37
3.5.1 The First step: Certain Special Cases.....	39

3.5.2	The Second Step: Normal Exploitation of the Work	40
3.5.3	The Third Step: Unreasonable Prejudice to Legitimate Interests 41	
4	Conflicting Perspectives on GenAI Training under the EU TDM Framework 44	
4.1	The Perspective of AI Developers and TDM Users.....	44
4.1.1	Lawful Access and Contractual Limitations	45
4.1.2	The Risk of Biased and Outdated AI Systems	45
4.1.3	“Take It or Leave It” Licensing Practices	46
4.1.4	Rivalry with Regulation & Global Innovation.....	46
4.2	The Rightholders’ Perspective	47
4.2.1	BurdenOpt-Out Mechanisms and the Burden of Monitoring...47	
4.2.2	The Economic Impact of AI Outputs on Rightholders	48
4.2.3	The Significance of Safeguarding Human Creativity	49
4.3	The Need for a New EU Approach to GenAI Training.....	49
5	Future Directions for TDM Regulation in the Era of GenAI.....	51
5.1	Comparative Approaches to TDM and GenAI Training	51
5.1.1	The US Fair Use Doctrine.....	51
5.1.2	The Japanese “Non-Enjoyment” Approach	52
5.2	Possible Improvements to the Existing EU Framework	53
5.2.1	Clarification of Ambiguous Legal Concepts.....	53
5.2.2	Enhancing the Functionality of Opt-Out Systems.....	53
5.2.3	Measures Concerning TPMs and Access Restrictions	54
5.2.4	Possible Remuneration-Based Solutions.....	54
5.3	Final Remarks	55
6	Conclusion.....	57
7	Bibliography	58
7.1	Case Law.....	58
7.2	Treaties, Conventions, Guidelines and Legislation	59
7.3	Books, Journal Articles, Research Papers and Studies.....	60
7.4	Master Theses and Dissertations	65
7.5	Online Sources	65
7.6	Technical, Scientific and Non-Legal Sources Related to TDM and AI Training	65

Summary

Generative Artificial Intelligence (“GenAI”) has raised major problems to copyright law, especially with regard to the use of copyrighted works for AI training via Text and Data Mining (“TDM”). In the European Union, this question is mostly addressed by Articles 3 and 4 of the CDSM Directive, with Article 4 emerging as the main legal basis for commercial AI training activities. However, the rapid growth of GenAI systems has generated fundamental considerations as to whether the current EU framework is still able to strike the correct balance between technical innovation and protection of rightholders.

This thesis analyses Articles 3 and 4 of the CDSM Directive in the context of GenAI training and discusses the legal and practical problems emerging out of the current EU framework. Particular attention is paid to concerns such as lawful access, machine-readable opt-outs, transparency obligations, fragmented reservation systems and the wider uncertainty around the applicability of TDM exceptions to modern AI systems. The thesis also looks at whether the current system manages to find a balance between innovation and copyright protection, particularly in the context of the international three-step test and the broader aims of EU copyright law.

Moreover, the thesis addresses broader regulatory debates surrounding AI training and examines alternative approaches to balancing the interests of AI developers and rightholders. The paper analyses legislation, case law, policy developments and academic commentary to identify a number of legal and practical deficiencies in the current framework and discusses possible improvements to improve the practical functioning of the EU TDM regime.

Ultimately, the thesis argues that although the EU framework seeks to support innovation while maintaining strong copyright protection, the current regulation increasingly struggles to accommodate the technological realities of contemporary GenAI systems. The existing framework thus needs to be further developed and more technologically responsive regulation that can meet the special issues posed by the training of GenAI.

Sammanfattning

Inom upphovsrättslagstiftningen har användningen av upphovsrättsskyddade verk för AI-utbildning via text- och datautvinning (TDM) skapat betydande problem. Denna fråga behandlas i första hand av artiklarna 3 och 4 i CDSM-direktivet i Europeiska unionen. Artikel 4 utgör den rättsliga grunden för kommersiell AI-utbildning. Ändå har den snabba expansionen av GenAI-system lett till betydande tvivel om huruvida det nuvarande EU-systemet kan hålla rätt balans mellan teknisk innovation och skydd av rättighetsinnehavare.

Denna avhandling undersöker artiklarna 3 och 4 i CDSM-direktivet i samband med GenAI-utbildning. Den undersöker också de juridiska och praktiska frågor som uppstår från den nuvarande EU-lagstiftningen. Frågor som laglig åtkomst, maskinläsbara undantag, skyldigheter för transparens, fragmenterade reservationssystem och den ökande osäkerheten om tillämpligheten av TDM-undantag på moderna artificiella intelligenssystem har fått särskild uppmärksamhet. Dessutom undersöker avhandlingen om det nuvarande systemet lyckas balansera framsteg och upphovsrättsskydd, särskilt med tanke på det internationella trestegstestet och EU:s upphovsrättslag.

Dessutom undersöker avhandlingen alternativa sätt att balansera AI-utvecklares och rättighetsinnehavares intressen och tar upp större regulatoriska frågor om AI-utbildning. För att identifiera ett antal rättsliga och praktiska brister i det nuvarande ramverket undersöker artikeln lagstiftning, rättspraxis, policyutveckling och akademiska kommentarer. Den diskuterar också potentiella förbättringar för att förbättra hur EU:s TDM-system fungerar i praktiken.

Slutligen hävdar avhandlingen att även om EU:s regler upprätthåller starkt upphovsrättsskydd och främjar innovation, har de nuvarande reglerna svårt att anpassa sig till de tekniska kraven i moderna GenAI-system. Som ett resultat behöver det nuvarande ramverket förbättras ytterligare, och en mer tekniskt responsiv reglering krävs för att hantera de unika utmaningar som utbildningen av GenAI medför.

Preface

This thesis marks the final milestone of a two-year journey that became one of the most meaningful adventures of my life. And so my Lund story comes to an end too. It was a journey that not only broadened my academic knowledge and developed me intellectually, but also gave me amazing memories, valuable people and experiences that I will always take with me.

First of all, I would like to thank my supervisor, **Petra Gyöngyi**, for her guidance, support, valuable feedback and encouragement that helped me to finish this thesis. I am endlessly grateful for her patience and thoughtful guidance throughout the way. I would also like to thank all the teachers and academic staff that taught, motivated and challenged me during these two years and helped me grow in so many different ways.

I also want to thank the Swedish Institute for giving me the opportunity to experience this and to be a part of such an inspirational environment.

And to lastly, my family, miles distant but always making me feel one heartbeat away. Your love, belief and support supported me through all the difficult moments. To my friends, thank you for being there for me: for inspiring me when I felt exhausted, for calming me when I became stressed, and for always reminding me to believe in myself.

Every chapter ends so that a new one may begin. But I know that this chapter will always remain one of my favorite parts of the book.

Lund, May 2026

Zhala Safaraliyeva

Abbreviation

AI Artificial Intelligence

AI Act Regulation (EU) 2024/1689 laying down harmonised rules on Artificial Intelligence

CDSM Directive (EU) 2019/790 on copyright and related rights in the Digital Single Market

CFR Charter of Fundamental Rights of the European Union

CJEU Court of Justice of the European Union

EU European Union

GDPR Regulation (EU) 2016/679 General Data Protection Regulation

GenAI Generative Artificial Intelligence

InfoSoc Directive 2001/29/EC on the harmonisation of certain aspects of copyright and related rights in the information society

JURI Committee on Legal Affairs

SMEs Small and Medium-sized Enterprises

TDM Text and Data Mining

TFEU Treaty on the Functioning of the European Union

TPM Technological Protection Measures

TRIPS Agreement on Trade-Related Aspects of Intellectual Property Rights

WCT WIPO Copyright Treaty (1996)

WIPO World Intellectual Property Organization

WTO World Trade Organization

1 Introduction

1.1 Background

In recent years, GenAI has become one of the fastest developing and most discussed technologies in the world. AI systems are no longer limited to analytical or repetitive functions, but are now capable of generating creative outputs such as texts, images, music, software code and videos, as demonstrated by systems like DALL-E 2, Midjourney, Stable Diffusion and ChatGPT. GenAI technologies are increasingly used in education, healthcare, software development, scientific research, marketing and entertainment. At the same time, the capabilities of these systems continue to develop rapidly, making GenAI one of the defining technologies of the digital era.

GenAI systems are heavily dependent on the analysis of large amounts of data. To produce human-like text, realistic visuals or other creative outputs, AI models must “learn” from pre-existing material. TDM supports this learning process, as it enables automated systems to analyze data sets and discover the patterns, trends and relationships in the data. Classical TDM and machine learning are not the same thing but both deploy automated analytical techniques to find patterns within information.¹ These techniques can help AI systems learn representations of language, images, sounds and styles that then allow them to generate new outputs.

The increasing creative capabilities of AI systems have generated legal and ethical questions regarding copyright law. AI systems are now capable of generating music, artworks, poems and highly realistic digital content, while also imitating artistic styles and creative expression. As a result, discussions concerning originality, authorship, ownership and protection of AI-generated outputs have become increasingly important. However, copyright concerns arise long before the final output is generated. In practice, GenAI systems function only because they are trained on massive quantities of pre-existing material, much of which may be protected by copyright. For example, an AI system cannot generate lyrics in the style of The Beatles² or create artworks such as the Belamy portrait³ without first analysing existing works and identifying stylistic patterns within them. Copyright law is therefore relevant not only at the stage of AI-generated outputs, but also during the training

¹ Eleonora Rosati, ‘Copyright as an Obstacle or an Enabler? A European Perspective on Text and Data Mining and its Role in the Development of AI Creativity’ (2020) 27(2) *Asia Pacific Law Review* 198–217 <https://ssrn.com/abstract=3452376> accessed 22 April 2026.

² Ibid, In 2017, SONY CSL Research Lab introduced *Daddy’s Car*, one of the first songs composed by AI through the FlowMachines system, which was trained on the repertoire of The Beatles.

³ Ibid, The “Belamy portrait” refers to *Edmond de Belamy*, an AI-generated artwork created by the French collective Obscure Art using a generative adversarial network (GAN), which was auctioned at Christie’s in 2018.

phase itself.

This issue has become increasingly important in legal and public discussions, particularly following disputes concerning the legality of AI training on copyrighted material. One of the most important recent developments in Europe is the German case *GEMA v OpenAI*.⁴ In the present case, the German collecting society GEMA brought copyright infringement claims against OpenAI for the use of German song lyrics protected by copyright in the training of GPT language models. GEMA argued that the AI models memorised and reproduced protected lyrics without authorisation, while OpenAI maintained that the systems merely learned abstract statistical relationships and patterns from training data rather than storing copyrighted works themselves. The dispute therefore raised fundamental questions as to whether GenAI training should be understood as a form of analytical TDM or as an act of copyright-relevant reproduction.

The significance of *GEMA v OpenAI* goes beyond the specific case. The case further intensified discussions concerning the applicability of the EU TDM framework to GenAI systems and exposed several uncertainties within the current legal regime. In particular, it highlighted unresolved questions concerning the distinction between analytical processing and memorisation of protected works, the limits of lawful access, the scope of the Article 4 TDM exception and whether existing copyright exceptions are capable of addressing the technical realities of modern GenAI systems. As GenAI technologies continue to evolve rapidly, these concerns have become increasingly central to discussions on the future balance between copyright protection and technological innovation within the EU.

The legal status of TDM itself lies at the centre of these debates. One argument suggests that TDM should not be restricted by copyright because it focuses on analysing information and identifying patterns rather than exploiting the expressive qualities of a work. This idea is commonly reflected in the phrase “the right to mine is the right to read”.⁵ From this perspective, if a person has lawful access to a work, no additional authorisation should be necessary to analyse it through automated methods. Some scholars further argue that excessive restrictions on TDM may interfere with freedom of expression and access to information, as protected under the CFR.⁶

At the same time, important concerns remain regarding the rights of authors and other rightholders. TDM activities often involve copying, extracting,

⁴ *GEMA v OpenAI* (Landgericht München I, 11 November 2025) Case No 42 O 14139/24

⁵ Christophe Geiger, Giancarlo Frosio and Oleksandr Bulayenko, ‘Text and Data Mining: Articles 3 and 4 of the Directive 2019/790/EU’ in Concepción Saiz García and Raquel Evangelio Llorca (eds), *Propiedad intelectual y mercado único digital europeo* (Tirant lo Blanch 2019) 27–71 <https://ssrn.com/abstract=3470653> accessed 5 May 2026.

⁶ Ibid, *Charter of Fundamental Rights of the European Union* [2012] OJ C326/391.

storing and processing protected material on a massive scale. Even where these acts are carried out for analytical rather than expressive purposes, the technical processes involved may still fall within the scope of copyright and related rights. This view argues that unrestricted TDM may pose a threat to the protection of intellectual property rights, and in particular the right to property protected by Article 17 of the CFR.⁷ The debate around TDM thus reflects a broader tension between technological innovation and copyright protection. Within the EU, these issues are mainly regulated by the CDSM Directive, especially Articles 3 and 4, which introduced specific exceptions for TDM.⁸ These provisions were adopted with the aim of modernising EU copyright law and adapting it to the realities of the digital environment. However, the Directive was drafted before the rapid emergence of large-scale GenAI systems such as ChatGPT and image-generation models. At the time, the TDM exceptions were mainly designed with analytical uses of data in mind rather than systems capable of generating new expressive outputs based on processed material.

Although the wording of the Directive is broad enough to potentially include GenAI training, the scale, complexity and practical challenges of modern AI systems were not fully anticipated during the legislative process. The application of the CDSM framework to GenAI therefore raises significant legal and practical questions. There is still uncertainty regarding how Article 4 opt-out mechanisms work, the breadth of lawful access, the distinction between analysis and memorisation in training AI, and whether the current legal framework is adequate to regulate contemporary GenAI technologies.

This thesis aims to analyse the extent to which the current EU legal framework achieves the appropriate balance between copyright protection and the need to promote technological innovation in the context of GenAI. In particular, the thesis takes into account the role of TDM in the training of GenAI systems and assesses whether the current legal framework under the CDSM Directive is capable of addressing the legal and practical challenges raised by modern AI technologies.

1.2 Research Aim and Research Questions

The aim of this thesis is to analyse the role of TDM in the training of GenAI systems and to examine how the current EU legal framework regulates this process from a copyright perspective. In particular, the thesis focuses on the TDM exceptions introduced under Articles 3 and 4 of the CDSM Directive and evaluates whether these provisions are capable of addressing the practical and legal realities of modern GenAI training. The research further aims to assess whether the current framework successfully balances two competing

⁷ Ibid, 7.

⁸ *Directive (EU) 2019/790 of the European Parliament and of the Council on Copyright and Related Rights in the Digital Single Market* [2019] OJ L130/92.

interests: on the one hand, the protection of authors' and rightholders' copyright interests, and on the other hand, the need to support technological development and innovation in the field of artificial intelligence.

Against this background, the main research question of this thesis is:

To what extent does the current EU legal framework on Text and Data Mining, particularly under the CDSM Directive, effectively balance the protection of copyright with the promotion of technological innovation in the context of GenAI training?

In order to answer this question, the thesis addresses the following sub-questions:

- To what extent are the TDM exceptions under Articles 3 and 4 of the CDSM Directive suitable for the practical realities and technical processes of modern GenAI training?
- How far does the current EU TDM framework practically support and encourage technological innovation and the development of GenAI systems?
- To what extent can the rights of authors and other rightholders be effectively protected in practice under the current framework, and how effective is the opt-out mechanism under Article 4(3) of the CDSM Directive?
- What lessons can the EU framework draw from alternative approaches to TDM regulation, particularly Japan's flexible TDM model and the U.S. fair use system?

1.3 Methodology and Sources

This thesis mainly employs a doctrinal legal research method, also called legal dogmatic method. This method emphasises the analysis of legal rules, principles, concepts and existing legal frameworks in order to improve the understanding of the functioning of the law and of the interpretation of particular legal difficulties.⁹ The doctrinal method is particularly suitable for this research because the thesis aims to investigate how the EU legal framework regulates the TDM in the context of training GenAI and whether the existing rules can address the legal issues arising from the modern AI technologies. The study is primarily undertaken through the analysis of legislation, case law and academic legal opinion as the thesis is primarily concerned with the interpretation, application and effectiveness of current legal provisions.

⁹ J M Smits, 'What Is Legal Doctrine? On the Aims and Methods of Legal-Dogmatic Research' in R van Gestel, H Micklitz and E L Rubin (eds), *Rethinking Legal Scholarship: A Transatlantic Dialogue* (Cambridge University Press 2017) 207–228.

The research conducted in this thesis is primarily based on both primary and secondary legal sources. The primary legal materials include key EU legal instruments concerning copyright, TDM, and AI regulation, particularly the *CDSM Directive*, the *InfoSoc Directive*, the *AI Act* and relevant provisions of the TFEU. International legal instruments, including the *Berne Convention*, the *TRIPS Agreement* and the *WIPO Copyright Treaty*, are also examined where relevant to the discussion of copyright exceptions and the three-step test. In addition, the thesis refers to comparative legislative approaches, including the US fair use doctrine, the Japanese TDM framework, and certain Member State approaches to TDM regulation.

The thesis also refers to judicial decisions, in particular the case law of the CJEU on reproduction rights, copyright exceptions and digital uses of protected works. Decisions such as *Infopaq*, *ACI Adam*, *GS Media* and etc. are considered in relation to reproduction rights. The thesis also discusses relevant national decisions, such as *Robert Kneschke v LAION*, to better understand the practical application of opt-out mechanisms and disputes regarding AI training.

In addition to primary materials, the thesis makes use of a broad range of secondary academic literature on copyright law, TDM and GenAI. Of particular importance are the works of *Martin Senftleben* and *Eleonora Rosati* on the three-step test, TDM exceptions and the interpretation of the EU copyright framework. The writings of *Tatiana Margoni* are referred to especially in relation to lawful access and opt-out mechanisms, while *Stepanka Havlikova* is used on machine readability and the technical difficulties around opt-out systems. The thesis also makes use of the scholarship of *Christophe Geiger*, especially with regard to the interpretation and balancing objectives of the CDSM Directive. Alongside these authors, various journal articles, policy papers, official reports, and academic blog discussions are used throughout the research.

1.4 Scope and Delimitations

The thesis discusses the tension between copyright protection and technical innovation in the context of training GenAI, especially the regulation of TDM under the CDSM Directive. In particular, the research analyses how far the existing EU legal framework governs the use of protected content in the training of AI models, and whether the TDM exceptions established by the CDSM Directive can meet the legal and practical problems posed by contemporary GenAI systems. Hence, the thesis analyses the relationship between copyright law, AI development and innovation in the EU legal environment.

The main focus of this thesis is the legal framework of the EU. Some national cases and examples may be referenced when appropriate, but the thesis does not aim to provide a full study of the domestic copyright laws of specific

Member States. Instead, the legislative focus is still at the level of the EU, in particular the CDSM Directive and its TDM exceptions.

It also is specifically about GenAI systems, not simpler predictive or discriminative machine learning models. This restriction is because GenAI is at the center of legal, technological and public discussions today. GenAI raises more complex and broad copyright issues than more traditional AI models, especially regarding the use of copyrighted works for training.

Furthermore, the thesis only addresses copyright problems that arise during the input or training phase of AI systems. The study in question is thus on the use of data and copyrighted material for TDM and AI training purposes. Questions of legal status of AI generated results such as authorship, ownership or copyright protection of works produced by AI systems are out of scope of this thesis and hence will not be examined in detail.

Lastly, the thesis studies the main features of TDM to be regulated under the CDSM Directive, particularly copyright issues related to acts of reproduction and acts of extraction happening in the course of TDM activities. Though, there might be a whole range of other legal issues associated with AI training, which are out of the scope of this research. This may include, for example, concerns relating to the moral rights, privacy and personal data protection and other rights which may arise depending on the type of data involved, e.g. images, text, audio or personal information. Although these difficulties are closely related to the development of AI, they are beyond the scope of this thesis.

1.5 Outline

The aim of this thesis is to evaluate the EU legal framework that allows TDM for the training of GenAI, in particular Articles 3 and 4 of the CDSM Directive, and to find out to what extent these provisions achieve a fair balance between the protection of copyright and the promotion of technical innovation.

Chapter 2 covers the importance of TDM for GenAI systems, and its relation to copyright law. It also outlines the historical and legal reasons that culminated in the approval of special TDM exceptions in EU legislation.

In Chapter 3, the legal and practical problems emerging from the implementation of Articles 3 and 4 of the CDSM Directive are examined. In specifically, the chapter investigates notions such as “lawful access”, “opt-out mechanisms”, “machine-readable reservations” and the compatibility of TDM exceptions with the three-step test in the context of GenAI training.

In chapter 4 the topic is discussed from the viewpoint of AI developers and rightholders. The paper addresses the real-world tensions between

technological innovation and copyright protection, stressing the challenges experienced by both parties under the current structure and the ongoing need for a more balanced and sustainable strategy.

Chapter 5 finally discusses possible future directions for the improvement of the current framework. It considers comparative approaches in jurisdictions such as the United States and Japan, prospective improvements within the current EU framework, and strategies that may lead to a more effective and balanced regulation of AI training and copyright.

2 The Role of TDM in AI Training and Its Copyright Implications under EU Law

2.1 Text and Data Mining: Definition and Procedure

TDM is the process of examining enormous volumes of text and data to extract relevant information, detect patterns and discover correlations between different pieces of information. These activities are impossible or at least very hard to complete manually, and hence they are performed via automated computing methods.¹⁰ TDM has been described as “a term that is commonly used to describe the automated processing (‘machine reading’) of large volumes of text and data to discover new knowledge or insights” in this context.¹¹

In practice, TDM often - though not always - entails copying enormous volumes of material, removing relevant data and recombining it to detect patterns and relationships.¹² But, because of the great variety of TDM approaches, aims and technologies, it is impossible to present a single, all-encompassing description of how it works in all cases. Nevertheless, the process can be described in three main steps.¹³

That’s step one: access to the content. In order for TDM to occur, the relevant data or text must be available, either because it is in the public domain or because access has been gained under a licence. This stage is very essential because the outcomes of the analysis will depend on the type and quality of the data available. The materials can be raw data, processed data sets or collections such as databases.

The second phase is the extraction and/or copying of content, and it is here that most of the legal issues arise. Even if access to the content is lawful, this does not indicate that TDM activities are inherently free to be carried out. In many circumstances, data copying or extraction for analysis purposes is necessary and these operations may be subject to copyright, related rights or database protection. At the same time, not all TDM techniques involve

¹⁰ Eleonora Rosati, ‘The Exception for Text and Data Mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market: Technical Aspects’ (Briefing requested by the JURI Committee, Policy Department for Citizens’ Rights and Constitutional Affairs, European Parliament 2018) 2 [https://www.europarl.europa.eu/Reg-Data/etudes/BRIE/2018/604942/IPOL_BRI\(2018\)604942_EN.pdf](https://www.europarl.europa.eu/Reg-Data/etudes/BRIE/2018/604942/IPOL_BRI(2018)604942_EN.pdf) accessed 20 April 2026.

¹¹ European Commission, ‘Commission Staff Working Document – Impact Assessment on the Modernisation of EU Copyright Rules Accompanying the Proposal for a Directive on Copyright in the Digital Single Market’ SWD(2016) 301 final, pt 1/3, para 4.3.1.

¹² Rosati (n10) 2.

¹³ Ibid, 4.

copying and not all copying requires authorisation.¹⁴ For example, temporary copies might be covered under the exception under Article 5(1) of the InfoSoc Directive and no authorisation is required.¹⁵ Some actions, such as certain types of web mining, may also be performed without consent in some circumstances;¹⁶ But the legal limits can apply if copying or extraction is outside of these exclusions. In particular, both copyright and sui generis database rights may be applicable when dealing with structured collections of data, and rightholders may control acts such as copying, extraction or re-use. In addition to the absence of intellectual property rights, other legal regimes (such as contract law or data protection and privacy laws under the General Data Protection Regulation (GDPR)) may still impose restrictions.¹⁷

The third phase is the extraction of data and text and the finding of knowledge. This phase involves processing the obtained data through phases of cleaning and preprocessing where low quality or unneeded data is removed and transformation where data is prepared for analysis.¹⁸ The algorithms are then used to search for patterns, trends and relationships within the data set. Finally, the results are reviewed and presented in a systematic way, to extract relevant insights from the data.

2.2 The Role of TDM in the Training of GenAI Systems

TDM is a basic technique used in the training of GenAI systems. In a nutshell, these systems learn by looking at huge amounts of data – books, websites, photographs or audio – to find patterns and connections within the material. Instead of memorising a single piece of work, the model learns how things relate to one another via TDM, such as how words combine to make phrases, or how visual aspects combine to form objects. This approach allows the system to learn common knowledge about data and then utilise that to generate new stuff.

This is where TDM comes in during training, turning raw data into structured representations usually called embeddings. These are mathematical representations that capture the most essential material qualities, such as meaning, style, or visual aspects. This learned representations allows the model to generalise from the existing data, rather than simply memorising it.

¹⁴ Ibid, 4-7

¹⁵ *Directive 2001/29/EC of the European Parliament and of the Council on the harmonisation of certain aspects of copyright and related rights in the information society* [2001] OJ L167/10 ('InfoSoc Directive')

¹⁶ Ibid.

¹⁷ *Regulation (EU) 2016/679 of the European Parliament and of the Council on the Protection of Natural Persons with regard to the Processing of Personal Data and on the Free Movement of Such Data (General Data Protection Regulation)* [2016] OJ L119/1.

¹⁸ Artha Dermawan, 'Text and Data Mining Exceptions in the Development of Generative AI Models: What the EU Member States Could Learn from the Japanese "Non-enjoyment" Purposes?' (2024) 27(1) *The Journal of World Intellectual Property* 45.

This is the reason why GenAI systems are able to produce outputs that seem unique and original, but are in reality built from data that has already been looked at.

There are several examples of TDM use in AI training for different types of generative models. For instance, large language models such as GPT-4 are trained on vast quantities of text data including books, articles and online content. In training TDM is used to evaluate language patterns, grammar and context so that the model can predict the next word in a phrase and produce coherent text.¹⁹

Practical examples of TDM usage in the image generating industry are DALL-E 2 and Stable Diffusion. These systems create visuals from text cues, such as “an astronaut on Mars” or “a teddy bear playing basketball.”²⁰ They are trained on very huge datasets of image-text pairs to do this. TDM allows the model to learn the mapping of some words to visual elements such as forms, colours and styles.

Central to this technique is Contrastive Language–Image Pre-training (CLIP), which trains on big datasets to learn the link between text and images.²¹ TDM facilitates CLIP to learn correlations between verbal and visual features represented as embeddings. Diffusion modelling is another key approach. It begins with random noise and slowly converts it into a coherent image based on learned patterns.²² Without TDM in beforehand, the system would lack the information to steer this transformation.

Similarly, models like Stable Diffusion are trained on datasets like LAION-5B (which contains billions of image-text pairs).²³ These datasets are not used to replicate individual photos, but to enable the system to recognise general patterns across massive data sets. Through iterative analysis, the model learns to correlate ideas (such “dog,” “sunset,” or “painting style”) with visual features, and then mixes them in novel ways to create fresh outputs.²⁴

What is obvious from these instances is that TDM is not only a supporting tool, but a required prerequisite for AI training. Without TDM the models

¹⁹ OpenAI, ‘GPT-4 Technical Report’ (2023) arXiv preprint arXiv:2303.08774 <https://arxiv.org/abs/2303.08774> accessed 24 April 2026.

²⁰ Dermawan (n18) 48.

²¹ A Ramesh, P Dhariwal, A Nichol, C Chu and M Chen, ‘Hierarchical Text-Conditional Image Generation with CLIP Latents’ (2022) arXiv:2204.06125 <https://arxiv.org/pdf/2204.06125.pdf> accessed 21 April 2026.

²² Robin Rombach and others, ‘High-Resolution Image Synthesis with Latent Diffusion Models’ (2022) *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 10684 <https://arxiv.org/abs/2112.10752> accessed 20 April 2026.

²³ C Schuhmann, R Vencu, R Beaumont, T Coombes, C Gordon, A Katta, R Kaczmarczyk and J Jitsev, ‘LAION-5B: A New Era of Open Large-Scale Multi-Modal Datasets’ (2022) <https://laion.ai/blog/laion-5b/> accessed 20 April 2026.

²⁴ Dermawan (n18) 48.

would not be able to process big datasets, detect patterns or construct the representations necessary for generation. In other words, without TDM, the system would just perceive bits and pieces of data and wouldn't be able to learn from it. Thus, GenAI would not be able to generate coherent text, realistic visuals or organised outputs, as the underlying learning process is completely dependent on the ability to assess and extract patterns from data at scale.

2.3 TDM and Copyright Law

Since data becomes valuable only when meaning and patterns are extracted from it, an important question arises: why should researchers and developers of GenAI systems be concerned about copyright when using data for TDM purposes? Given that TDM focuses on analysing information rather than using the expressive form of a work, it is often considered that such activities may fall outside the scope of copyright protection. This follows the basic rule that copyright does not protect the data itself but rather the original expression in which that data is presented.²⁵ In this sense, the use of data for analytical purposes may not seem problematic and TDM may look as an activity which does not need prior authorisation.²⁶ It has also been argued that imposing restrictions on TDM could go against the objectives of copyright law and may even affect fundamental rights.²⁷

However, this stance is not so evident when we look at how TDM works in practice. Usually, TDM does not work on raw data alone, and the data itself may not be protected. Instead, it depends on massive datasets that typically include words, photographs, sounds and other protected works.²⁸ In the world of big data it is becoming more difficult to distinguish unprotected data from protected content. Consequently, TDM activities may involve the usage of content that are protected by copyright law.

Also, some specific technological processes within the TDM process can lead to legal protection. Thus, if TDM entails the extraction and/or copying of content, especially in the second step of the process outlined above, this may be under the right of reproduction or in some situations the sui generis database right.²⁹ At the same time, it should be noted that not all TDM techniques require copying, and not all acts of copying require prior authorization, particularly where they fall within existing exceptions and

²⁵ Geiger (n5) 6.

²⁶ Ibid.

²⁷ Ibid.

²⁸ C Gerrish and AM Skavlan, 'European Copyright Law and Text and Data Mining Exceptions and Limitations: In Light of the Recent DSM Directive, is the EU Approach a Hindrance or Facilitator to Innovation in the Region?' (2019) 2(2) *SIPLR* 58.

²⁹ *The British Horseracing Board Ltd and Others v William Hill Organization Ltd* (C-203/02) EU:C:2004:695.

limitations under EU law.

2.4 TDM under the EU Legal Framework before the CDSM Directive

Before the implementation of the CDSM Directive, EU law did not have a clear exception for TDM. Instead, it was sometimes argued that certain TDM activities could fall under existing restrictions, and in particular under the exception for temporary reproductions under Article 5(1) of the InfoSoc Directive.³⁰ This rule permits certain restricted actions of copying which are transient or incidental and are necessary for a technological procedure, for example, to enable the legitimate use of a work, and which have no independent economic significance.

On the face of it, this exemption may seem relevant to TDM, as the analysis of data sometimes involves copying as part of a technical process. But the conditions are quite strict. The reproduction must be transient, incidental and technically necessary and it must be an integrated part of an automated process. These requirements are already challenging for TDM in AI development, since data are typically saved, reused and analysed over longer time periods instead of short and incidental handling.

The case law of the CJEU has also narrowed the scope of this exception. In the *Infopaq* judgement³¹, the Court concluded that even very short extracts from a work, such as eleven words, may be protected reproduction provided that they indicate the originality of the work. The Court also noted that the exemption is relevant only if copying is strictly necessary for the effective functioning of a technological process and is carried out as part of an automated process. This interpretation considerably limits the scope of Article 5(1), since many TDM activities, in particular those requiring storage and further use of data, can go beyond what is strictly essential for the technical process.

A similar restrictive approach can be found in national case law. In the German case *Kneschke v LAION*³², the court considered whether the copying involved in using photos in a dataset may be covered by the temporary reproduction exception. It found that it could not. The court said the photos were actively downloaded and analysed, meaning the reproductions were not incidental. Moreover, the removal of some images was not an automatic and simply technological procedure, but rather a question of deliberate programming and user choice. These copies were not only temporary, nor limited to what was strictly necessary for a technological

³⁰ InfoSoc Directive (n15) Art 5(1).

³¹ *Infopaq International A/S v Danske Dagblades Forening* (C-5/08) EU:C:2009:465.

³² *Robert Kneschke v LAION eV* (Regional Court of Hamburg, Case No 310 O 227/23, 27 September 2024).

procedure, but the data was kept and used for subsequent research. The activity therefore did not meet the requirements of Article 5(1).

This example points up a bigger issue. In the current data-driven setting, TDM is characterised by the repeated use, storage and processing of massive amounts of data across time. Such features are not easily compatible with the idea of transient, accidental and purely technical reproductions defined by the InfoSoc Directive. Even if some limited features of TDM could be covered by Article 5(1), the provision as a whole was not able to comprehensively cover TDM operations in AI training.

In short, while there are existing regulations that arguably may be used to TDM, the pre-CDSM framework did not provide a clear or solid legal basis for such actions. The temporary reproduction exception was difficult to implement in practice because of its severe restrictions, as explained in the above-discussed cases. This shows that the previous copyright framework was not entirely well matched to the needs of the digital age. For this reason, the adoption of the CDSM Directive aimed to introduce rules that are better adapted to modern technologies and the realities of TDM.

2.5 The Regulation of TDM under the CDSM Directive

The CDSM Directive represented a major shift in the EU's attitude towards TDM. As stated in the previous section, the legal environment prior to the CDSM Directive did not provide a clear basis for TDM activities, particularly in the context of modern technology and AI systems. One of the key reasons for the adoption of the CDSM Directive was the need to update copyright law and adapt it to the reality of the digital age. This intention is reflected in Recital 1 of the Directive which states that the legislation aims to modernise copyright and related rights in the digital environment and "contribute to the functioning of the internal market, provide for a high level of protection for rightholders, facilitate the clearance of rights and create a framework in which the exploitation of works and other protected subject matter can take place".³³

The discussions on TDM formally began in 2016 when the European Commission presented the Proposal for a Directive on Copyright in the Digital Single Market. One of the most noteworthy points of the plan was the inclusion of a new mandatory exception particularly related to TDM activities. However, the idea was somewhat limited in its scope at the initial stage. The exception was primarily intended for scientific research reasons, allowing research institutions to produce reproductions and extractions of works they had lawful access to for the purpose of conducting TDM activities.³⁴ In other words, the original idea was more about supporting for

³³ CDSM Directive (n8) recital 1.

³⁴ Rosati (n1) 18-20.

academic and scientific research, rather than focusing on wider commercial or technological uses of TDM.

The scope of the proposal gradually extended throughout the legislative process. It was then recommended that the exception should be extended to cultural heritage institutions as well as research bodies. More notably, the discussions within the Committee on Legal Affairs (JURI), in particular through the report of Axel Voss, led to the possibility of a wider TDM exception.³⁵ The proposal was to allow TDM operations by any person who had legitimate access to the material in question, provided that the rightholder had not expressly reserved the use of the work, including by machine-readable mechanisms. This discovery was the basis for what subsequently became Article 4 of the Directive.

Thus, the final text of the CDSM Directive contained two different TDM exceptions in Articles 3 and 4. The Directive was officially adopted and published in 2019, bringing for the first time a more thorough legal framework for TDM within EU copyright law. In particular, Article 4 permitted TDM activities for a far broader category of users, not only scientific research institutes. This larger approach was particularly relevant for technical development and AI systems since it may allow the use of protected materials for GenAI training where lawful access was available and the usage was not explicitly reserved by the rightholder.

At the same time, questions rapidly emerged as to whether the provisions were indeed meant to apply to GenAI systems. It has been suggested that the EU legislator did not have modern GenAI technology in mind when drafting the TDM exclusions.³⁶ The primary aim at that time was to allow large-scale analytical applications of protected works, such as looking for patterns, trends and correlations through data analysis.³⁷ The original discussion wasn't about whether AI systems might use these patterns to develop whole new outputs down the line.³⁸

The wording of the Directive, however, allows a broader interpretation. Article 2(2) of the CDSM Directive defines TDM as any automated analytical technique used to analyse text and data in digital form in order to generate information, including patterns, trends and correlations.³⁹ This concept is wide and technology-neutral and also explicitly refers to the generation of information through automated analysis and therefore might

³⁵ Ibid.

³⁶ Martin Senftleben, 'Text and Data Mining, Generative AI, and the Copyright Three-Step Test' (2026) 57(1) *International Review of Intellectual Property and Competition Law* 67–107 https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5373903 accessed 05 May 2026.

³⁷ CDSM Directive (n8) recital 8.

³⁸ Senftleben (n36) 68.

³⁹ Ibid.

also cover the training procedures of GenAI systems, which too make use of the automated analysis of big data sets to find patterns and generate representations.

This larger meaning has gradually won recognition in later EU legislation and case law. In particular, the AI Act appears to clarify that the exemption for TDM under the CDSM Directive may be applicable to the training of GenAI. Recital 105 of the AI Act refers directly to the TDM exceptions provided for in the CDSM Directive, and Article 53(1)(c) of the AI Act obliges providers of certain AI systems to respect opt-outs stated by rightholders under Article 4(3) of the CDSM Directive.⁴⁰ This means that the EU legislature considers Article 4 potentially applicable to training GenAI.

In the final framework TDM is regulated in different ways and for different beneficiaries under Article 3 and Article 4 of the CDSM Directive.

Article 3 provides for a mandatory exception for TDM carried out for the purposes of scientific research. This article is applicable to research bodies and institutions of cultural heritage, if they have lawful access to the relevant works or subject matter. Under this provision, certain institutions are able to reproduce and extract protected materials for TDM purposes without prior authorisation from the rightholder. Furthermore, Article 3 also provides that contractual provisions opposed to the exception are not to be enforceable, so that rightholders cannot circumvent the exception through private agreements. At the same time, however, the scope of Article 3 is still rather limited, as it applies only to certain beneficiaries and only for scientific research reasons.

Article 4 takes a more general approach. It is not limited to institutions of cultural heritage or research organisations, as in Article 3. Instead, it applies broadly to anyone who has lawful access to the content. This is why Article 4 is of special importance to commercial actors and AI developers. The exception under Article 4, however, is subject to an essential limitation: the rightholders may explicitly reserve their rights and decide not to use the exception. The Directive also permits for such reservation to be made by machine-readable means, especially in the online context. This means that Article 4 provides for access to TDM and at the same time allows for some control by rightholders over the use of their works.

The difference between Articles 3 and 4 is the result of the EU legislator's attempt to balance competing interests. On the one hand, the Directive pursues the purpose of promoting innovation, research and technological progress through the possibility of certain forms of automated data analysis. On the other hand, it also intends to protect the interests of rightholders by

⁴⁰ Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (*Artificial Intelligence Act*) [2024] OJ L 2024/1689, art 53(1)(c).

preserving methods of control, in particular by way of the opt-out system under Article 4.

Overall, the adoption of Articles 3 and 4 was a significant step to make the eu copyright law fit for the digital age and the increasing relevance of TDM technologies. The CDSM Directive developed a more modern framework that might potentially deal with activities that the previous legal regime could not effectively support. Nevertheless, these advances leave substantial legal problems and uncertainties as to the practical execution of these regulations, especially in the context of GenAI training. These problems and challenges will be addressed in the next chapter.

3 Critique of the TDM Framework under the CDSM Directive

3.1 Member States' National TDM Frameworks

Before the CDSM Directive was adopted, several Member States had begun to implement TDM exclusions in their national copyright laws. But the approaches taken by the different countries differed substantially as to scope, conditions and beneficiaries. The differences resulted in a certain level of fragmentation in the EU legal framework and led to an inconsistent regulation of TDM among the Member States.

For example, the United Kingdom amended the Copyright, Designs and Patents Act 1988 in 2014 to provide a TDM exception for non-commercial research purposes.⁴¹ Under the UK approach, a person who has lawful access to a work might carry out computational analysis for non-commercial research purposes without infringing copyright. But that exception does not apply to database rights.

France went one step further and introduced exceptions to TDM for both copyrighted works and databases.⁴² French law permitted reproductions from lawful sources for TDM purposes and also permitted the storage and distribution of files created during TDM research activities. At the same time, the French system was based on the concept of “lawful sources,” which meant that the relevant material must be made available lawfully with the consent of the rightholders.

Germany took a slightly different route. German law, unlike the French approach, did not need a “lawfully accessed source” and did not limit mining only to text and data related with scientific literature.⁴³ With regard to databases, German law treated reproductions carried out for TDM purposes as part of the “normal use” of the database. This created a somewhat wider framework for mining activity.⁴⁴

In 2017, Estonia likewise adopted a TDM exception.⁴⁵ Under Estonian law, copyrighted works might be processed for TDM purposes as long as the use was non-commercial. Unlike the French framework, the Estonian framework

⁴¹ Copyright and Rights in Performances (Research, Education, Libraries and Archives) Regulations 2014, SI 2014/1372, reg 3, inserting s 29A into the Copyright, Designs and Patents Act 1988.

⁴² Loi n° 2016-1321 du 7 octobre 2016 pour une République numérique, art 38, inserting para 10 into art L122-5 and para 5 into art L342-3 of the Code de la propriété intellectuelle (Intellectual Property Code)

⁴³ Geiger (n5)

⁴⁴ Ibid.

⁴⁵ Estonian Copyright Act, art 19(3).

did not explicitly request lawful access. However, it required users to mention the name of the author, the title of the work and the original publication where possible.

These instances show that the Member States have taken varied legislative measures as far as TDM is concerned during the pre-CDSM Directive period. The restrictions related to lawful access, database use and beneficiaries varied from one country to another, creating legal uncertainty for cross-border research and innovation collaborations inside the EU.

In this perspective, the inclusion of mandatory TDM exceptions binding across Member States was one of the beneficial features of the CDSM Directive, which helped to a more unified EU approach towards TDM operations. However, this harmonising impact has not yet eliminated all fragmentation, as will be explained in the next section.

Another beneficial aspect of the Directive is the contractual override. Before the CDSM framework, licensing agreements could often be used by publishers and rightholders to prevent or limit TDM activity. In fact, the transaction costs of getting permission from various publications could be considerable and make some research programs economically impractical.⁴⁶ The Directive so merits credit for the protection of scientific research under Article 3 against contractual override. This helps ensure that rightholders cannot simply exclude lawful TDM research by private contracts.

However, the framework provided by the Directive is not without restrictions. While activities of scientific research are protected from contractual restrictions by Article 3, the larger exception under Article 4 is subject to opt-outs by rightholders. Moreover, these provisions face several practical and legal issues that remain to be addressed with respect to their extent and effectiveness. Therefore, the CDSM framework, although being a very important addition to the regulation of TDM in the EU, has also been subject to considerable criticism, which will be analysed in the following sections.

3.2 National Implementation and the Limits of Harmonisation

As explained above, the CDSM Directive introduced mandatory TDM exceptions under Articles 3 and 4 and therefore established a more harmonised EU approach towards TDM activities. Still, fragmentation has not been eliminated even with this harmonising influence.

This is mainly due to the legal nature of the directives themselves. Article 288 of the TFEU lays forth that directives shall be binding upon each Member State to whom they are addressed as to the result to be achieved, but shall

⁴⁶ Geiger (n5)

leave to the national authorities the choice of form and means of implementation.⁴⁷ Unlike regulations, directives are not immediately applicable in national legal systems, but must be transposed into domestic law by means of national legislative procedures.

In the context of the CDSM Directive, Member States had to alter their national copyright frameworks to integrate the TDM exceptions under Articles 3 and 4. However, while these rules are mandatory, Member States have some freedom as to the wording, legislative structure, interpretation and practical application. Differences between national legal systems may therefore still arise in the presence of common EU rules.

This issue is abundantly demonstrated by the discussions on national implementation. For instance, Pereira argues that the Portuguese implementation was largely faithful to the wording and objectives of the Directive, while also trying to balance the interests of researchers, users and right holders.⁴⁸ This reflects an implementation model that is closely aligned with the original framework adopted at EU level.

At the same time, other researchers have urged that Member States should adopt a more comprehensive and innovation-oriented strategy in the implementation of the Directive. For example, Micke argued that Finland should apply the TDM provisions more widely in order to better safeguard freedom to conduct mining activities.⁴⁹ This approach implies that Member States should read the Directive in a way that encourages research and technological development as far as possible within the constraints of EU law.

A similar discussion took place regarding the Italian implementation of the Directive. Calabrese noted that the EU legislature did not expressly exclude TDM users from legal liability in relation to the right of communication to the public.⁵⁰ Calabrese viewed the Italian implementation positively in this respect as Italy took a rather broader approach in the transposition process.⁵¹ This argument suggests that the Italian model could be seen as an attempt to create a more innovation-friendly framework and react to the growing international competition of legal systems outside the EU offering more

⁴⁷ *Consolidated Version of the Treaty on the Functioning of the European Union* [2016] OJ C202/47, art 288.

⁴⁸ Giorgos Dimitris Vrakas, 'A Literature Review of "Lawful" Text and Data Mining' (2024) 4 *Open Research Europe* 153 <https://doi.org/10.12688/openreseurope.18013.2> accessed 10 May 2026, Alexandre L Dias Pereira, 'TDM Copyright for AI in Europe: A View from Portugal' (2023) 18 *Journal of Intellectual Property Law & Practice* 900.

⁴⁹ Vrakas (n48), Micke Lindholm, 'Text and Data Mining under Finnish Copyright Law before and after the DSM Directive' (2020)

⁵⁰ *Ibid.*, Bernardo Calabrese, 'Scientific TDM Exception and Communication to the Public: Did Italians Do It Better ... or at Least Not Worse?' (2022) 17 *Journal of Intellectual Property Law & Practice* 399

⁵¹ *Ibid.*

flexible conditions for TDM and AI development.⁵²

These instances show that even when adopting the same mandatory provisions, Member States may nonetheless interpret and apply the Directive differently to some extent. Some governments might stick more closely to the wording of the Directive while others might aim for broader interpretations that are more supportive of innovation and technological progress.

However, the freedom of Member States in the implementation is not boundless. Rosati argues that the requirements of Articles 3 and 4 necessitate a relatively narrow and careful approach to transposition, given that the essential normative content of the provisions cannot be fundamentally altered at national level.⁵³ Under this interpretation, Member States cannot simply introduce significantly broader TDM exceptions than those envisaged by the Directive itself, as this could conflict with EU law and undermine the objective of creating a more coordinated framework.

Even a little flexibility in implementation can nevertheless lead to differences in practice. National legislators and courts may still interpret questions differently relating to legitimate access, TPMs, contractual override or the practical implementation of opt-out systems.

However, although the CDSM Directive has diminished fragmentation in comparison to the prior legal landscape, a considerable degree of divergence still exists across Member States in the practical execution of the framework.

3.3 A Critique of Article 3 of the CDSM Directive

Article 3 of the CDSM Directive provided for a mandatory exception for TDM for scientific research objectives. This rule permits research organisations and cultural heritage institutions to copy and extract protected works for TDM purposes, provided that they have lawful access to the necessary materials. The article was introduced to encourage scientific research and reduce legal confusion in the research sector in relation to data analysis operations.

There are certain crucial requirements to the scope of Article 3. Firstly, the exception is limited to research organisations and cultural heritage institutions. Second, the TDM activity must be conducted solely for scientific research purposes. Thirdly, the entities relying on the exception must have lawful access to the relevant works or other protected subject matter. Finally,

⁵² Ibid.

⁵³ Eleonora Rosati, 'No Step-Free Copyright Exceptions: The Role of the Three-Step in Defining Permitted Uses of Protected Content (Including TDM for AI-Training Purposes)' (2024) 46(5) *European Intellectual Property Review* 262 <https://ssrn.com/abstract=4629528> accessed 02 May 2026

Article 3(3) authorises rightholders to take measures that are required for the security and integrity of their networks and databases, but such measures shall not go beyond what is necessary to attain that objective.⁵⁴

3.3.1 The Limited Scope of Beneficiaries under Article 3

One of the primary features of the article 3 is that the exemption is applicable only to certain beneficiaries. This provision may be used only by research organisations and cultural heritage institutions. The Directive thus assumes a somewhat modest institutional approach to TDM.

The term of “research organisation” in the Directive is likewise quite specific. To fall within the exception, organisations have to be non-profit and reinvest their profits into scientific research or be engaged in activities related to a public interest mission.⁵⁵ As set out in Recital 12, this public interest mission can be evidenced by means of public funding, national laws or public contracts.⁵⁶

Organisations controlled by commercial undertakings, which may be the case where a private company influence through shareholder status or membership rights, are excluded from the Directive. It also does not apply where commercial entities are given preferential access to the results of the research.⁵⁷

But the Directive also acknowledges that scientific research today often relies on collaboration between public institutions and commercial entities. For this reason, recital 11 states that the exception may also benefit public-private partnerships.⁵⁸ This means that research organisations can benefit from private technological infrastructure, expertise or technical help when carrying out TDM activities. At the same time the cooperation must be designed in such a way that the commercial partner cannot take control of the research organization or have privileged access to the research results.⁵⁹

Another key point is that independent researchers and individuals conducting research outside qualifying organisations are not covered within the scope of Article 3.⁶⁰ This is a narrower framework than in some other jurisdictions. The UK TDM exception applies not just to institutions but also to individual researchers and persons with lawful access to a work.⁶¹ As a result, Article 3 narrows the pool of beneficiaries to institutional actors mainly within the conditions provided by the Directive.

⁵⁴ CDSM Directive (n8), Article 3(3).

⁵⁵ CDSM Directive (n8), Art. 2(1) (a-b).

⁵⁶ Ibid, Recital 12

⁵⁷ Ibid, Art. 2(1) and Recital 12.

⁵⁸ Ibid, Recital 11

⁵⁹ Ibid.

⁶⁰ Geiger (n5).

⁶¹ Ibid, Copyright, Designs and Patents Act 1988, § 29A (UK)

3.3.2 Restrictions on “Scientific Research Purposes”

According to Article 3 TDM activities shall be conducted solely for scientific research reasons. Some experts suggest that limiting the exception to only “scientific research” could diminish the practical usefulness of the provision and generate confusion in practice.

For example, a university may already be entitled to lawfully access a database under a permission acquired for educational reasons. It is not apparent whether, in this case, the university would still require an additional licence to use the same database for scientific research via TDM activities.⁶² Since an exception is not one which can be contracted out of under Article 3 the answer would probably be no.⁶³ But such uncertainty might nevertheless cause practical issues for researchers and institutions.

This can make research organisations hesitant to engage in TDM projects due to potential legal issues, ambiguous license borders and concerns about potential liability. For this reason, some scholars argue that when the exception is already limited to certain beneficiaries, such as research organisations and institutions dedicated to cultural heritage, further limiting the exception to “scientific research purposes” is unnecessary and would make the provision narrower.

3.3.3 The Impact of Technological Protection Measures on TDM

Another key problem under Article 3 is TPMs. Article 3(3) allows rightholders to take measures necessary to maintain the security and integrity of their networks and databases.⁶⁴ The reasoning for this clause is reasonable, since rightholders may wish to avoid dangers such as unauthorised access, excessive automatic downloading or harm to their digital systems.

However, these measures may also pose practical obstacles for researchers engaged out lawful TDM activity. Even where a research organisation is able to comply with all the standards set out in Article 3, and has lawful access to the materials in question, technical limitations may nonetheless make it impossible to acquire data or limit the ability to carry out TDM activities efficiently.

Recital 16 tries to solve this problem by saying that such measures must be proportionate, must not go further than necessary and must not put at risk the effective application of the exception.⁶⁵ In theory, this means that rightholders should not use TPMs in such a way that the TDM exception is useless in practice.

⁶² Geiger (n5).

⁶³ Ibid

⁶⁴ CDSM Directive (n8), Art. 3(3)

⁶⁵ Ibid, Recital 16.

But in practice issues still are there. Many databases and online platforms have technical barriers such as access control, anti-bot systems, download limits or automated blocking technologies. These systems are designed to protect, but can potentially stop researchers from accessing and analysing data even if the activity itself is allowed under Article 3.

The anti-circumvention provisions make the matter much more problematic. Academic literature has recognised that TPMs may block not only infringing uses but also uses that are actually allowed by exceptions and limitations to copyright.⁶⁶ As Guibault et al. explain, Article 6(4) obliges rightholders to provide the means to benefit from an exception but does not give users the right to circumvent technological measures themselves.⁶⁷ In other words, researchers may have a legal right to rely on the TDM exception, but still face technical obstacles that they cannot legally overcome by themselves.

This means that in some cases the practical success of Article 3 may depend not only on the content of the exception, but also on the way in which TPMs are applied by rightholders in practice.

3.3.4 Legal Uncertainty Surrounding “Lawful Access”

A further major concern in relation to Articles 3 and 4 of the CDSM Directive is the interpretation of the “lawful access” requirement. The Directive provides that TDM may only be carried out where users have lawful access to works or other protected subject matter, but does not define the concept. Recital 14 provides several examples of lawful access, but the precise scope of the requirement is unclear.⁶⁸

The challenge has been particularly apparent in recent litigation involving GenAI systems. In *Authors Guild v. OpenAI*, one of the principal complaints was the alleged use of copyrighted works from unauthorized online sources to train the AI.⁶⁹ This raises a key question for GenAI systems: if copyrighted content is publicly accessible online, does accessing and using that material for TDM automatically satisfy the lawful access requirement? The problem becomes more complicated because the same work may simultaneously appear on different lawful and unlawful online sources. In such situations, it remains unclear whether lawful access should depend on the legality of the source itself or on the conduct and knowledge of the user accessing the material.

⁶⁶ Geiger (n5).

⁶⁷ Ibid, L Guibault and others, *Study on the Implementation and Effect in Member States’ Laws of Directive 2001/29/EC on the Harmonisation of Certain Aspects of Copyright and Related Rights in the Information Society* (Report prepared for the European Commission, DG Internal Market, ETD/2005/IM/D1/91, 2007).

⁶⁸ CDSM Directive (n8), Recital 14.

⁶⁹ *Authors Guild v OpenAI Inc* (SDNY, Case No 1:23-cv-08292, 19 September 2023).

This uncertainty creates serious practical problems for AI developers and research organisations. Modern AI systems are trained on enormous datasets containing millions or billions of online content items collected automatically through web scraping and data crawling techniques. Verifying the legal status of every individual source used during training is therefore technically and practically unrealistic. The issue becomes even more complicated after training has already taken place, since AI systems learn from data through patterns and statistical associations rather than storing simple removable copies. If unlawful content is later found, it may be technically impossible or prohibitively expensive to remove its effects from the model.⁷⁰ AI systems don't just store copies of works in the typical way (like deleting a file or removing a hyperlink).⁷¹

Some scholars therefore argue that interpreting lawful access too strictly could undermine the practical effectiveness of the TDM exceptions. Margoni, in particular, argues that lawful access should not necessarily depend on whether every individual source itself is fully legitimate.⁷² Rather, the focus should be on the user's conduct and knowledge. In this light, it would be impossible to impose a burden on TDM activities for the scale and automated nature of the AI training systems, requiring AI developers to check the legality of every online source.

The debate is linked also to existing CJEU case law. *ACI Adam*⁷³ and *Copydan*⁷⁴ took a restrictive approach to unlawful sources in the area of reproduction rights, while *GS Media*⁷⁵ and *Stichting Brein*⁷⁶ focused more on user knowledge and financial purpose in assessing online access to unauthorized content. These cases suggest two possible approaches: a strict position excluding unlawful sources entirely, or a more flexible approach based on whether the user knew or should have known that the source was unlawful.⁷⁷

Margoni further argues that Recital 14 supports a broader understanding of lawful access because it refers not only to subscriptions and contractual arrangements, but also to content that is “freely available online.”⁷⁸ Under this interpretation, publicly accessible online material could potentially fall within

⁷⁰ Thomas Margoni, ‘TDM and Generative AI: Lawful Access and Opt-Outs’ (draft v0.2, August 2024, forthcoming in *Auteurs & Media* 2024/2) <https://ssrn.com/abstract=5036164> accessed 19 May 2026.

⁷¹ *Land Nordrhein-Westfalen v Dirk Renckhoff* (C-161/17) EU:C:2018:634.

⁷² Margoni, (n70)

⁷³ *ACI Adam BV and Others v Stichting de ThuisKopie and Stichting Onderhandeligen ThuisKopie vergoeding* (C-435/12) EU:C:2014:254.

⁷⁴ *Copydan Båndkopi v Nokia Danmark A/S* (C-463/12) EU:C:2015:144.

⁷⁵ *GS Media BV v Sanoma Media Netherlands BV and Others* (C-160/15) EU:C:2016:644.

⁷⁶ *Stichting Brein v Ziggo BV and XS4ALL Internet BV* (C-610/15) EU:C:2017:456.

⁷⁷ Margoni, (n70).

⁷⁸ CDSM Directive (n8), Recital 14.

lawful access even where the original upload itself was unauthorized, provided that users did not bypass paywalls or TPMs. This approach aims to take account of the practicalities of AI training and large-scale online data collection.

However, this broader interpretation is controversial. Copyright holders may contend that the legality of TDM of material from illegal online sources undermines the protection of copyright and reduces control over the use of protected works, especially in the context of GenAI systems trained on huge amounts of copyright content without clear authorisation. The debate on lawful access thus reflects a broader tension in EU copyright law between effective copyright protection and the practicalities of TDM and AI development in the digital environment.

3.4 Critique of Article 4 and the Opt-Out Mechanism

While Article 3 is of primary relevance for research organisations and cultural heritage institutions and scientific research purposes, Article 4 of the CDSM Directive created a more general TDM exception available to anyone with lawful access to the relevant content. This increased scope was a big step forward for innovation as it meant that commercial organisations and AI developers may use the TDM exception for AI training and other mining activities.

However, at the same time, Article 4 also created an opt-out provision under Article 4(3), which allowed rightholders to retain their rights and preclude the use of their works for TDM purposes.⁷⁹ This approach seeks to balance innovation with copyright protection, but has also produced a number of legal and practical problems. In practice, uncertainty remains high due to a lack of clear and standardised opt-out system, no precise definition of “machine-readable” reservations, different reservation methods used by rightholders, and practical difficulties regarding proof and timing of opt-outs during AI training.

3.4.1 Challenges Surrounding the Concept of “Machine Readability”

Under Article 4(3) of the CDSM Directive, rightholders are entitled to reserve their rights in relation to TDM. Recital 18 clearly states that for content freely available online “it should *only* be deemed appropriate to reserve those rights by the use of machine-readable means.”⁸⁰ Thus, machine readability was effectively made an mandatory requirement for a valid opt-out with relation to

⁷⁹ CDSM Directive (n8).

⁸⁰ Ibid, recital 18.

publicly available online content under Article 4.

The primary problem, however, is that neither Article 4 nor the Directive itself provides a precise definition of what exactly “machine-readable” is. Recital 18 simply notes that this might include metadata and terms and conditions of a website or service. These examples are not exhaustive. This means that the exact scope of machine readability is still unknown.

In *Robert Kneschke v. LAION* the German court established a somewhat broad view of this notion. In this decision, the German court clarified that machine readability should not be viewed narrowly in a technical sense, but in a broader meaning as “machine understandability.”⁸¹ In other words, the court reviewed whether reservations formulated in ordinary natural language could be considered machine-readable if AI systems could understand them. Since modern AI technologies increasingly enable machines to understand human language and even a very mundane textual statement on a website could be a valid machine-readable opt-out.⁸²

The court further stated that the Directive does not require that reservations be made in the “simplest” technical way possible, but in a ‘appropriate’ manner.⁸³ This interpretation may boost the standing of rightholders, but also gives rise to serious practical problems for AI developers and TDM users. It is a fact that AI systems are getting better and better at understanding natural language. But this does not mean that all AI systems grasp all languages or wordings or contexts equally well. AI models can understand reservations differently and vary in trustworthiness according on the language or technology system they are deployed in.

This creates legal confusion for users relying on the TDM exception. And even if AI developers are sincerely trying to follow the opt-out rules, they are still at danger of infringement if their algorithms don’t do a good enough job of understanding or detecting natural language objections.⁸⁴ The situation is more complicated in multilingual online contexts as reservations can be communicated in many different formats and languages.

More generally, the CDSM Directive aimed to provide a balanced framework which would give copyright protection, but also encourage innovation and technological advancement in a digital world. Both parties have a responsibility to get that balance right. On the one hand, AI developers and TDM users should exercise reasonable measures to detect, interpret and comply with opt-

⁸¹ *Kneschke v LAION* (n32).

⁸² *Ibid.*

⁸³ *Ibid.*

⁸⁴ Stepanka Havlíková, ‘Technical Challenges of Rightsholders’ Opt-out From Gen AI Training after *Robert Kneschke v LAION*’ (2025) 16 *JIPITEC* 93-95 <https://www.jipitec.eu/jipitec/article/view/422> accessed 14 May 2026.

out reservations in the course of data collection and training activities. On the other hand, rightholders should be able to voice their reservations by adequately explicit, reliable and technologically accessible means.

3.4.2 Practical Problems of Robots.txt as an Opt-Out Tool

One of the most widespread ways of reserving rights against TDM is the usage of Robots.txt files.⁸⁵ Robots.txt is a voluntary standard that provides advice to web crawlers as to whether parts of a website may or may not be viewed. The file often includes rules for the proper user agent, including folders or files that are accessible and those that should not be accessed.⁸⁶

However, in fact Robots.txt poses numerous key restrictions as an opt-out mechanism under Article 4(3) of the CDSM Directive. The key problems include the fact that the system relies mostly on identifying unique user agents and it is the scrapers themselves who set their own “user-agent” name.⁸⁷ The “user-agent” option allows website owners to set different criteria for different crawlers. For example, a website may allow some users to index its material but ban other AI crawlers or scraping programs. In this way, rightholders can select between a whitelist approach, allowing just certain crawlers, or a blacklist approach, excluding certain crawlers.⁸⁸

This could appear to be a practical solution, but it also places substantial burdens on the rightholders. Rightholders need to know at all times which crawlers exist, how they identify themselves and which user-agent names to restrict or allow, to use the system efficiently.

Another key disadvantage is the lack of precision of Robots.txt describing the purpose of use.⁸⁹ The system cannot easily forbid only certain sorts of TDM actions and allow others. In fact, banning a user agent may prevent all access by that crawler, regardless of the intended usage of the crawler.

Also, crawlers independently decide and reveal their own user-agent names, creating another transparency concern. Major AI businesses generally publicly announce their crawler names and provide advice for how websites can opt out of AI training. But not all developers behave the same way, some crawlers may fail to identify themselves clearly.⁹⁰ This generates inequity

⁸⁵ P Bernt Hugenholtz, ‘The New Copyright Directive: Text and Data Mining (Articles 3 and 4)’ (*Kluwer Copyright Blog*, 24 July 2019) <<https://copyrightblog.kluweriplaw.com/2019/07/24/the-new-copyright-directive-text-and-data-mining-articles-3-and-4/>> accessed 14 May 2026.

⁸⁶ Havlikova (n84) 97.

⁸⁷ Martijn Koster, Illyés Zoltán, Henrik Zeller and Lizzi Sassman, ‘RFC 9309: Robots Exclusion Protocol’ (September 2022) <https://www.rfc-editor.org/rfc/rfc9309.html> accessed 14 May 2026.

⁸⁸ Havlikova (n84) 98.

⁸⁹ Ibid.

⁹⁰ Ibid.

between developers who cooperate voluntarily and those who do not.

Google has developed Google-Extended, which allows websites to opt out of some AI training usage. While this has been framed as an effort to comply with Article 4(3) of the CDSM Directive, several commentators suggest that the intent of Article 4 was to give rightholders the option of reserving rights against TDM generally, rather than requiring separate opt-outs for every single AI model or company.⁹¹ The obligation on creators to be continually aware of emerging AI systems and to repeatedly submit separate reservations may become very difficult and practically ineffective, especially for smaller rightholders with little technological resources.

This highlights a major weakness of the current opt-out system. Reliance on Robots.txt and similar tools makes Article 4(3) difficult to enforce in practice, because there are no clear and uniform technical standards for AI training and crawler registration.

3.4.3 Legal Issues Related to the Proof of Opt-Outs

Another important issue concerning Article 4 is the timing of opt-outs and the burden of proof as to whether a valid reservation existed at the time of the TDM activity. It is generally accepted that the opt-out mechanism under Article 4(3) operates mainly on an “*ex ante*” basis, i.e., rightholders are expected to reserve their rights prior to the TDM process.

At the same time, some academics argue that Article 4(3) does not explicitly exclude the possibility of “*ex post*” reservations.⁹² However, such an approach creates serious practical difficulties in the context of GenAI training. Once an AI model has been trained on data, it may become technically impossible or extremely costly to remove the influence of that information from the model.⁹³ During AI training, works are not merely copied but transformed into patterns, relationships and representations embedded within the system. Effective “*ex post*” opt-outs therefore appear highly impractical once training has already occurred.

As a result, rightholders are effectively required to implement reservations before TDM activities begin. In practice, however, this creates significant burdens, particularly because rightholders may need to continuously monitor technological developments, new AI systems and emerging scraping methods in order to ensure effective reservations. Such ongoing monitoring may be

⁹¹ Paul Keller, ‘Generative AI and Copyright: Convergence of Opt-outs?’ (Kluwer Copyright Blog, 23 November 2023) <https://legalblogs.wolterskluwer.com/copyright-blog/generative-ai-and-copyright-convergence-of-opt-outs/> accessed 14 May 2026.

⁹² Péter Mezei, ‘A Saviour or a Dead End? Reservation of Rights in the Age of Generative AI’ (2024) 46(7) European Intellectual Property Review 461, 469 <https://ssrn.com/abstract=4695119> accessed 10 May 2026.

⁹³ Margoni, (n70).

especially difficult for smaller creators and individual rightholders.

There are also additional challenges with regard to proof and timing. Websites and metadata and terms of use may change over time, making it difficult to establish whether a valid opt-out existed at the time of data collection. This is of particular relevance in copyright disputes where rightholders claim unauthorized TDM, while AI developers argue that no valid reservation existed at the time of the mining.

The question of the burden of proof therefore becomes highly significant. German explanatory materials suggest that the burden may fall on the user invoking the TDM exception to demonstrate the absence of a reservation.⁹⁴ Similarly, some scholars argue that because Article 4 permits TDM only where rights “have not been reserved,” users relying on the exception may need to prove that no valid opt-out existed at the relevant time.⁹⁵ At the same time, however, rightholders themselves remain responsible for communicating reservations clearly and effectively. Thus, in practice, the large-scale and automated nature of AI training creates substantial evidential difficulties.

3.4.4 *GEMA v OpenAI*⁹⁶ and the Limits of the TDM Exception

One of the key issues addressed by the *GEMA v OpenAI* judgement is whether GenAI training can still be classified as basic text and data mining within the meaning of Art 4 CDSM Directive. This case is particularly important because it challenges one of the basic assumptions underlying the current EU TDM framework: that AI systems primarily extract abstract information, patterns and statistical relationships from training data, rather than reproducing protected expression itself.

OpenAI argued that large language models do not directly store copyrighted works, but instead learn statistical relationships and patterns from training data. Under this understanding, AI training resembles traditional TDM, where the purpose is to analyse information rather than exploit the expressive content of works themselves. But the German court rejected this interpretation, holding that the protected song lyrics had been sufficiently fixed and memorized within the parameters of the model. The German court emphasised that it was sufficient that the model could reproduce the lyrics through user prompts, even if the lyrics were not stored in a directly readable form.

⁹⁴ Ibid, Explanatory memorandum (Gesetzesbegründung) of the German Government (Bundesregierung) to its legislative proposal implementing the CDSM Directive: Entwurf eines Gesetzes zur Anpassung des Urheberrechts an die Erfordernisse des digitalen Binnenmarktes, Gesetzesbegründung: Besonderer Teil. No. 19/27426, p. 88

⁹⁵ L Löbbling, Ch Handschigl, K Hofmann and J Schwedhelm, ‘Navigating the Legal Landscape: Technical Implementation of Copyright Reservations for Text and Data Mining in the Era of AI Language Models’ (2023) 14 JIPITEC 499.

⁹⁶ *GEMA v OpenAI* (Landgericht München I, 11 November 2025) Case No 42 O 14139/24 (n4)

Memorisation and later reproduction of protected expression were therefore treated as copyright-relevant reproductions rather than as purely analytical acts.

Particularly important is the German court's division of AI training into different phases. According to the judgment, the first phase consists of pre-training activities, namely the extraction and conversion of content into machine-readable form. The German court accepted that these preparatory acts may potentially fall within the scope of the TDM exception. However, the second phase — the actual analysis and training of the AI model — was treated differently.⁹⁷ According to the German court, once analysed information becomes encoded within the model in a way that allows later reproduction of protected expression, the activity can no longer be regarded as mere analytical mining. The training process therefore falls outside the protection of the TDM exception.

This interpretation creates some practical difficulties for Article 4 of the CDSM Directive. In practice, the conversion of information into machine-readable form is valuable for AI developers primarily because it enables subsequent computational analysis and model training. Merely extracting and converting data into machine-readable datasets has little independent value if the actual learning process is excluded from the scope of the exception. Consequently, if Article 4 protects only the preparatory stage of dataset creation while excluding the actual training phase, the exception risks becoming largely ineffective for modern GenAI systems.

The judgment therefore narrows the practical certainty surrounding the application of Article 4 to GenAI training. Although the wording of Articles 2(2) and 4 CDSM Directive appears broad enough to cover automated analytical processes used in AI training, the reasoning in *GEMA v OpenAI* introduces increasing uncertainty regarding the extent to which contemporary GenAI systems can rely on the TDM exception in practice.

While the judgment is not yet a final EU-wide position and is still likely subject to appeal, it does raise important questions about the future direction of the EU's approach to GenAI training under the Article 4 TDM framework. The case highlights a rising degree of ambiguity concerning the extent of the Article 4 in relation to contemporary GenAI systems.

3.5 The Three-Step Test and Text and Data Mining

The three-step test is one of the fundamental principles of international copyright law and serves as a framework for assessing whether copyright

⁹⁷ Karim, Mohamed Ataul, 'GEMA versus OpenAI: The Key Takeaways' (4iP Council, 2025) <https://www.4ipcouncil.com/application/files/1017/6545/5349/Paper-GEMA_v_OpenAI_02-12-25.pdf> accessed 25 May 2026.

exceptions and limitations are permissible. The test was first introduced in Article 9(2) of the Berne Convention, which allows limitations and exceptions to the reproduction right only in “certain special cases” that do not conflict with the normal exploitation of the work and do not unreasonably prejudice the legitimate interests of the author.⁹⁸ Similar formulations were later incorporated into other international instruments, including the TRIPS Agreement⁹⁹ and the WIPO Copyright Treaty.¹⁰⁰

In EU copyright law, the three-step test is reflected in Article 5(5) of the InfoSoc Directive¹⁰¹ and also applies to the TDM framework introduced by the CDSM Directive through Article 7(2)¹⁰². Within the EU legal order, the three-step test has gained not only legislative importance but also significant interpretative value through the case law of the “CJEU”. The CJEU has consistently treated the test not merely as a drafting principle for legislators, but also as an interpretative standard that national courts and Member States must consider when applying copyright exceptions and limitations.¹⁰³

The test consists of three cumulative conditions. First, the exception must apply only in certain special cases. Second, it must not conflict with the normal exploitation of the work. Third, it must not unreasonably prejudice the legitimate interests of the rightholder.

In the context of the CDSM Directive, Article 3 generally appears to satisfy the requirements of the three-step test more easily. The exception applies only to a limited category of beneficiaries, namely research organisations and cultural heritage institutions, and is restricted to scientific research purposes. In addition, the lawful access requirement means that the use of works may still depend on subscriptions, licences, or other lawful contractual arrangements. These limitations help preserve the normal exploitation of works while also reducing the risk of disproportionate prejudice to rightholders’ interests.

For this reason, most debates surrounding the compatibility of TDM with the three-step test arise primarily in relation to Article 4 and its broader

⁹⁸ *Berne Convention for the Protection of Literary and Artistic Works* (adopted 9 September 1886, revised at Paris 24 July 1971, amended 28 September 1979) 828 UNTS 221 (‘Berne Convention’) art 9(2).

⁹⁹ *Agreement on Trade-Related Aspects of Intellectual Property Rights* (adopted 15 April 1994, entered into force 1 January 1995) 1869 UNTS 299 (‘TRIPS Agreement’) art 13.

¹⁰⁰ *WIPO Copyright Treaty* (adopted 20 December 1996, entered into force 6 March 2002) 2186 UNTS 121 (‘WCT’) art 10.

¹⁰¹ InfoSoc Directive (n15) Art 5(5).

¹⁰² CDSM Directive (n9), Article 7(2)

¹⁰³ *Football Association Premier League Ltd and Others v QC Leisure and Others* (C-403/08 and C-429/08) EU:C:2011:631 [187]–[188]; *Luksan v van der Let* (C-277/10) EU:C:2012:65 [85]; *Mc Fadden v Sony Music Entertainment Germany GmbH* (C-484/14) EU:C:2016:689 [36]; *Poland v Parliament and Council* (C-401/19) EU:C:2022:297 [26]–[30].

application to commercial actors and GenAI training. The following sections therefore focus mainly on the challenges raised by Article 4 within the framework of the three-step test.

3.5.1 The First step: Certain Special Cases

The first step of the three-step test requires that a copyright exception applies only in “certain special cases,” meaning that the exception must have sufficiently clear and identifiable limits. Compared to Article 3 of the CDSM Directive, Article 4 appears considerably broader. Its scope includes commercial actors such as AI developers, while GenAI training may involve the large-scale use of copyrighted works and the creation of massive numbers of reproductions. This raises questions regarding whether Article 4 can still be regarded as a sufficiently “special” case within the meaning of the three-step test.¹⁰⁴

At the same time, Article 4 still contains several limiting elements. The exception applies only where users have lawful access to the relevant content, meaning that access may still depend on subscriptions, licences or other lawful arrangements. In addition, the opt-out mechanism under Article 4(3) allows rightholders to reserve their rights against TDM activities. Recital 18 further emphasizes that licensing opportunities should not be unduly affected by the exception.¹⁰⁵ These safeguards demonstrate that the scope of permitted uses is not entirely unlimited.

The large quantity of works involved in AI training should not alone determine compatibility with the first step of the three-step test. Copyright law already contains several exceptions involving broad use of protected works that have nevertheless been considered compatible with the three-step test.¹⁰⁶ Private copying permits large numbers of reproductions, while temporary copying under Article 5(1) of the InfoSoc Directive also concerns extensive temporary reproductions created through browsing, caching and streaming activities.¹⁰⁷ In *Infopaq*, the CJEU considered that even in the case of a large number of works and reproductions, temporary copying could still satisfy the three-step test.¹⁰⁸

The WTO Panel also noted that an exception does not necessarily need to pursue a narrowly defined purpose in order to satisfy the “certain special

¹⁰⁴ Senftleben (n 36); Nicolo Lucchi, *Generative AI and Copyright – Training, Creation, Regulation* (European Parliament, Policy Department for Justice, Civil Liberties and Institutional Affairs 2025); Christophe Geiger, ‘Elaborating a Human Rights-Friendly Copyright Framework for Generative AI’ (2024) 55 *International Review of Intellectual Property and Competition Law* 1129.

¹⁰⁵ CDSM Directive (n8), Recital 18.

¹⁰⁶ Senftleben (n36).

¹⁰⁷ *Ibid.*

¹⁰⁸ *Infopaq* (n31) [56].

case” requirement.¹⁰⁹ Nevertheless, where a clear policy objective exists, this objective may still be considered when assessing the legitimacy of the exception.¹¹⁰ In the context of GenAI, AI systems may provide important technological, economic, social, and cultural benefits by supporting new creative processes, improving products and services, reducing production costs, and broadening access to creative and informational resources.¹¹¹

For these reasons, there are good reasons to conclude that TDM exceptions for AI training are capable of satisfying the first step of the three-step test.

3.5.2 The Second Step: Normal Exploitation of the Work

The second step of the three-step test requires that an exception to copyright shall not conflict with the normal exploitation of the work. The opt-out mechanism provided for in Article 4(3) of the CDSM Directive provides one of the strongest arguments for compatibility with this requirement within the framework of Article 4 of the CDSM Directive. Since rightholders may reserve their rights against TDM and require authorization through licensing agreements, the possibility of commercializing works for AI training purposes remains protected. In this sense, Article 4 may be viewed as preserving potential licensing markets for TDM and AI training activities. Because rightholders retain the ability to prohibit unauthorized mining and negotiate licences where they wish to monetize such use, Article 4 may be viewed as avoiding a direct conflict with the normal exploitation of works.¹¹²

Senftleben further argues that even without an opt-out mechanism, a conflict with normal exploitation would not necessarily arise automatically. According to this view, the normal exploitation test protects exploitation modes that constitute a major source of income for a work rather than every possible commercial opportunity.¹¹³ During AI training, works are primarily processed as data inputs for machine learning rather than consumed directly for their expressive value. Since AI systems are trained on extremely large datasets, the inclusion of a single work will rarely constitute an economically significant exploitation method comparable to traditional markets such as book sales, streaming or licensing.¹¹⁴ In this respect, TDM for AI training differs from traditional exploitation “as a work”¹¹⁵ and may not necessarily interfere

¹⁰⁹ WTO (2000) 15 June 2000, Document WT/DS160/R, United States–110 (5) of the US Copyright Act, Report of the Panel. World Trade Organization, Geneva; Senftleben (n36)

¹¹⁰ Senftleben (n36). See also Martin R F Senftleben, *Copyright, Limitations and the Three-Step Test: An Analysis of the Three-Step Test in International and EC Copyright Law* (Kluwer Law International 2004) 138–152.

¹¹¹ Ibid.

¹¹² Ibid.

¹¹³ Ibid.

¹¹⁴ Margoni (n70).

¹¹⁵ R Ducato and A Strowel, ‘Ensuring Text and Data Mining: Remaining Issues with the EU Copyright Exceptions and Possible Ways Out’ (2021) 43 *European Intellectual Property Review* 322.

with the core commercialization interests protected by copyright law.

At the same time, more difficult questions arise regarding AI-generated outputs. Certain CJEU decisions demonstrate that copyright exceptions may conflict with normal exploitation where they contribute to the circulation of substitute content. In *ACI Adam*, the CJEU held that the private copying exception could not extend to copies made from unlawful sources because this would reduce lawful transactions relating to protected works.¹¹⁶ In *Stichting Brein*, the Court also held that temporary reproductions of the works in the context of illegal streaming services interfered with the normal exploitation of the works as they reduced the number of transactions on the legal market.¹¹⁷

The situation in the case of GenAI outputs is more complicated, since the AI-generated content does not usually reproduce protected works exactly, but produces an output that is based on more general ideas, styles or patterns derived from the training data.¹¹⁸ As a result, the economic harm caused to original works may often be more indirect than in cases involving piracy or direct unlawful copying. Not every AI-generated output will therefore automatically conflict with the normal exploitation of the original work. Nevertheless, cases may arise where AI outputs become sufficiently similar to protected works and start competing directly with them.

In this respect, the reasoning in *GEMA v OpenAI* becomes particularly relevant. Although the Court did not explicitly refer to “normal exploitation,” its reasoning reflects concerns that AI outputs capable of reproducing protected lyrics may interfere with the economic exploitation interests of rightholders.¹¹⁹ The Court’s emphasis on memorisation and reproducible outputs suggests that the closer AI-generated outputs move towards functioning as substitutes for original works, the more difficult it becomes to justify broad TDM exceptions under the second step of the three-step test.

However, such conflicts cannot be assessed in abstract terms for AI systems generally and instead require evaluation of the specific output itself. For these reasons, although AI-generated outputs may create case-by-case conflicts with normal exploitation, this does not necessarily mean that TDM exceptions themselves are inherently incompatible with the second step of the three-step test.

3.5.3 The Third Step: Unreasonable Prejudice to Legitimate Interests

The third step of the three-step test requires that a copyright exception must

¹¹⁶ *ACI Adam* (n73) [39].

¹¹⁷ *Stichting Brein* (n76) [70].

¹¹⁸ *Senftleben* (n36)

¹¹⁹ *GEMA v OpenAI* (n4)

not unreasonably prejudice the legitimate interests of right holders. This requirement is broader and more flexible than the previous step concerning normal exploitation. Whereas the normal exploitation analysis focuses primarily on whether a copyright exception interferes with the primary commercialization of an individual work, the unreasonable prejudice test allows a wider assessment of economic and market harm. As the WTO Panel explained, prejudice becomes unreasonable at the level where a copyright exception causes or has the potential to cause an unreasonable loss of income to the copyright owner.¹²⁰

In the context of GenAI, this stage of the analysis may create greater difficulties for TDM exceptions. Unlike the normal exploitation test, the assessment is not limited to the commercialization of one specific work. Instead, broader market effects may also become relevant. For example, AI-generated outputs could potentially add to dilution of the market for certain types of human-created works. If AI systems are capable of generating content that increasingly replaces demand for particular artistic styles, genres, or creative formats, questions may arise regarding whether the interests of authors and rightholders are being prejudiced at an unreasonable level.¹²¹

Similarly, the issue may also be examined from the perspective of licensing revenue. While the use of one individual work within a massive AI training dataset may not itself represent a major source of income, broader repertoires of works could still possess collective economic value for licensing markets. In this sense, the prejudice analysis is more elastic than the normal exploitation test because it allows consideration of wider economic consequences for groups of works or creative sectors rather than focusing exclusively on a single work in isolation.

At the same time, the existence of some level of harm does not automatically mean that the third step is violated. Every copyright exception inevitably limits exclusive rights to some extent.¹²² The decisive issue is therefore whether the prejudice exceeds what can still be regarded as reasonable within the balance between copyright protection and broader societal objectives. In this respect, the third step is essentially a proportionality test. In the context of TDM, this balancing exercise also implies taking into account the benefits linked to the development of AI with respect to innovation, scientific progress and technological advancement.

Moreover, the InfoSoc Directive, in particular Article 5(5) and Recital 31 as interpreted by the CJEU, requires that copyright exceptions strike a fair

¹²⁰ WTO (n109)

¹²¹ Senftleben (n36)

¹²² WTO (n109)

balance between the interests of rightholders and users.¹²³ As a result, the assessment under the third step is not limited to the protection of rightholders' economic interests only, but also includes a consideration of broader public interests and user freedoms that are balanced against the possible prejudice suffered by copyright holders.

Senftleben argues that mechanisms such as opt-outs and remuneration systems may mitigate potential prejudice to reasonable levels. In theory, if rightholders are able either to reserve their rights against TDM, or to receive compensation through licensing or remuneration systems, the balance required under the third step may still be maintained.¹²⁴ In this sense, the prejudice caused by AI training may potentially be mitigated through properly functioning regulatory and market mechanisms.

Still, as discussed throughout this chapter, practical implementation of these safeguards is problematic. Issues around lawful access, machine-readable opt-outs, transparency of AI training practices, identification of crawlers and enforcement of reservations all impact the practical efficacy of the current framework. As a result, even if the TDM exceptions may theoretically satisfy the three-step test, significant uncertainties continue to arise in practice.

Overall, the CDSM Directive demonstrates that the EU legislature intended to support the use of TDM for AI training and broader technological innovation. Nevertheless, because the Directive was adopted before the rapid development of modern GenAI systems, many of today's practical and technological challenges could not be fully anticipated during the drafting process. Consequently, although the CDSM framework provides an important legal basis for TDM activities, it does not fully align with the realities of contemporary AI training in practice, and several legal and technical uncertainties continue to remain unresolved.

¹²³ *VG Bild-Kunst v Stiftung Preußischer Kulturbesitz* (C-392/19) EU:C:2021:181 [54]; *Spiegel Online GmbH v Volker Beck* (C-516/17) EU:C:2019:625 [36]; *Pelham GmbH and Others v Hütter and Schneider-Esleben* (C-476/17) EU:C:2019:624 [32]; *Funke Medien NRW GmbH v Bundesrepublik Deutschland* (C-469/17) EU:C:2019:623 [51].

¹²⁴ Senftleben (n36)

4 Conflicting Perspectives on GenAI Training under the EU TDM Framework

4.1 The Perspective of AI Developers and TDM Users

Both AI developers and TDM users see the necessity of large-scale datasets for the advancement of the technology. GenAI systems depend on vast amounts of text, images, audio, and other data to learn patterns, structures, and generate output. Modern AI systems would be far less effective without widespread access to data. That's why many researchers and technology companies think that heavy copyright regulations can be a huge impediment to innovation and scientific progress.¹²⁵

Meanwhile, the existing EU framework displays a certain conflict in its own approach to TDM. The CDSM Directive, on the one hand, seeks to encourage the development of AI and digital innovation through the provision of statutory exceptions to TDM. However, Article 4 also has a structure that leaves much control in the hands of rightholders through opt-outs and licensing alternatives. Hugenholtz observes that the TDM framework generates a potential derivative market for TDM activity, where rightholders have the ability to control, license, or forbid the mining of their works.¹²⁶ Seen from the users' point of view, this could risk making TDM a very fragmented licensing environment rather than a truly innovation-friendly exception.

Some experts contend that such an approach may ultimately limit users' ability to access and use information effectively in the digital environment.¹²⁷ In practice, this means that AI developers may have significant confusion about whether datasets can be legally used, if opt-outs were applied appropriately, and whether licensing agreements are required. These uncertainties are especially relevant in the context of modern AI training, when millions or billions of individual data points might be analysed in parallel.

Similarly, research into TDM licensing procedures demonstrates the practical issues that users could have. For instance, McCracken and Raub talk about the licensing landscape around TDM as being fragmented and inconsistent with different publishers using different rules, permits, and contractual

¹²⁵ Michelle Brook, Peter Murray-Rust and Charles Oppenheim, 'The Social, Political and Legal Aspects of Text and Data Mining (TDM)' (2014) 20 *D-Lib Magazine* <http://www.dlib.org/dlib/november14/brook/11brook.html> accessed 14 May 2026.

¹²⁶ Hugenholtz (n85).

¹²⁷ Maryna Manteghi, 'The Insufficiency of the EU's Text and Data Mining Exceptions for Using Artificial Intelligence' (2022) 44(11) *European Intellectual Property Review* 651.

conditions.¹²⁸ Their research found that even universities and academic researchers faced significant challenges in obtaining access for TDM projects from different content suppliers.¹²⁹ This indicates that even if TDM is officially allowed, real access may be difficult, costly and legally unclear.

4.1.1 Lawful Access and Contractual Limitations

Lawful access is one of the most important issues to AI developers. As noted in earlier chapters, before TDM may take place, lawful access to the necessary content is required, both under Article 3 and Article 4 of the CDSM Directive. In theory, this need seems quite sensible, as it maintains the licensing markets and restricts access to copyrighted materials without permission. In practice, however, the approach could pose major obstacles to the training of AI.

Many databases, scientific publishers and online platforms use contracts, subscription systems or website conditions of use to limit TDM.¹³⁰ Some publishers forbid TDM altogether in licensing agreements, while others restrict it through platform constraints and access regulations.¹³¹

This generates a real contradiction in the system. In theory, developers and researchers of AI may be able to benefit from the TDM exception, but only if they have authorised access, for example through a subscription, licence or open access arrangement. Actually, getting access to millions of protected works through individual memberships or license agreements can be prohibitively expensive. Large AI systems are trained on very vast and diverse datasets collected from innumerable sources throughout the internet. This could, therefore, become economically unfeasible to require legitimate subscription access to each of these sources.

4.1.2 The Risk of Biased and Outdated AI Systems

Restrictive licensing arrangements may potentially have wider technology repercussions. If it becomes too expensive, or impractical, to legally acquire high-quality, up-to-date datasets, AI developers may increasingly use older, partial, publicly available, or lower-quality datasets. The European Copyright Society has raised this problem, warning that publishers would end up passing on the price of TDM license to their subscriptions, making it unaffordable for

¹²⁸ Peter McCracken and Emma Raub, 'Licensing Challenges Associated With Text and Data Mining: How Do We Get Our Patrons What They Need?' (2023) 11 *Journal of Librarianship and Scholarly Communication* 13 <https://www.iastatedigitalpress.com/jlsc/article/id/15530/> accessed 14 May 2026.

¹²⁹ Ibid.

¹³⁰ Maarten Truysens and Patrick Van Eecke, 'Legal Aspects of Text Mining' in Nicoletta Calzolari and others (eds), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)* (European Language Resources Association (ELRA) 2014) http://www.lrec-conf.org/proceedings/lrec2014/pdf/452_Paper.pdf accessed 15 May 2026.

¹³¹ Ibid.

many research institutes and developers.¹³²

Such a circumstance can lead to the development of biased or obsolete AI systems. Developers lacking access to broad and up-to-date data sets may resort to older materials, unverifiable data, or pre-trained systems with ambiguous underlying training data.¹³³ This may, in turn, have a negative impact on the quality, reliability and fairness of AI outputs. Biased training data might reinforce existing mistakes, discriminatory behaviours, or obsolete knowledge, producing larger societal risk that reach beyond copyright law itself.

4.1.3 “Take It or Leave It” Licensing Practices

Another key objection is that huge rightholders are in a better bargaining position than TDM consumers. Often access to datasets or platforms is based on acceptance of non-negotiable contract conditions imposed by publishers or database providers. Many TDM agreements are, as Geiger notes, “take it or leave it” arrangements, requiring users to accept the provider’s rules or be cut off totally.¹³⁴

Obtaining permissions from numerous publishers simultaneously may therefore become extremely complex and administratively burdensome, particularly for smaller AI developers and research teams. Large firms can afford to negotiate license agreements and legal compliance measures, but startups and SMEs generally do not have the resources to do so. This might indirectly strengthen the market position of the major tech corporations while reducing competition and innovation from smaller actors.

Manteghi consequently suggests that Member States should interpret and execute the TDM framework in a way that is more friendly to startups and SMEs participating in the development of AI.¹³⁵ Otherwise, the practical implementation of the current structure could favour huge firms who can afford to pay the expenses of licensing and the legal risks, while reducing the ability of smaller innovators to enter the market.

4.1.4 Rivalry with Regulation & Global Innovation

Another major worry for AI developers is that the EU’s relatively restrictive

¹³² Jonathan Griffiths, Tatiana Synodinou and Raquel Xalabarder, ‘Comment of the European Copyright Society Addressing Selected Aspects of the Implementation of Articles 3 to 7 of Directive (EU) 2019/790 on Copyright in the Digital Single Market’ (2023) 72 GRUR International 22, 11-12

¹³³ Thomas Margoni and Martin Kretschmer, ‘A Deeper Look into the EU Text and Data Mining Exceptions: Harmonisation, Data Ownership, and the Future of Technology’ (2022) 71 GRUR International 685

¹³⁴ Christophe Geiger, Giancarlo Frosio and Oleksandr Bulayenko, ‘Text and Data Mining in the Proposed Copyright Reform: Making the EU Ready for an Age of Big Data?’ (2018) 49 *International Review of Intellectual Property and Competition Law* 814, 826.

¹³⁵ Manteghi (n127).

and regulation-heavy TDM framework may reduce the competitiveness of the European AI companies compared to jurisdictions that are more flexible in their approach to AI training. The CDSM Directive seeks to strike a balance between innovation and copyright protection by providing mechanisms such as lawful access requirements, opt-outs, and licensing options. However, these safeguards may also lead to legal complexities and compliance challenges for AI developers operating in the EU market. As a result, many experts think that the EU is increasingly confronted with regulatory competition by countries whose copyright systems are more open to innovation in this field.¹³⁶

Here, the worry is particularly significant in the context of global AI development, as technology tends to evolve swiftly and across national borders. If AI developers find the EU framework too restricted, fragmented, or legally ambiguous, they may gradually redirect their research and investment efforts to nations with more permissive TDM freedoms or more predictable legal environments.

4.2 The Rightholders' Perspective

From the rightholders' standpoint, GenAI poses significant questions regarding copyright protection and the economic worth of creative works. In the EU copyright is recognised as a fundamental intellectual property right. Authors, artists, publishers, photographers and other creators have legitimate rights in regulating and benefiting from the use of their works. Given the extensive usage of copyright-protected material to train GenAI systems, many rightholders consider that this use must not take place without safeguards or economic payback.

The CDSM Directive aims to safeguard rightholders via measures such as lawful access restrictions and opt-out reservations. But the practical application of these precautions is tricky, as described in earlier chapters. The framework, in theory, does retain some protections, but the actual implementation of these rights may be technically complex and legally ambiguous.

4.2.1 BurdenOpt-Out Mechanisms and the Burden of Monitoring

One of the key safeguards afforded to rightholders under Article 4 of the CDSM Directive is the opportunity of opting out from TDM operations. This approach theoretically enables rightholders to reserve their rights and prohibit the use of their works for AI training purposes. In practice, however, the efficiency of this system is strongly dependent on ongoing monitoring, technological awareness and proper execution.

¹³⁶ Geiger (n5).

As already established, opt-outs pursuant to Article 4 are predominantly “*ex-ante*”. In other words, the reservation must be created prior to the TDM activity in most cases. This implies that rightholders need to keep an eye on the operation of AI developers and web crawlers to ensure that reservations are realised in time.

That puts a heavy burden on creators and copyright owners. At currently, there is no unified opt-out approach that applies to all AI developers and platforms. Instead, rightholders generally rely on fragmented technical methods such as metadata, website conditions or Robots.txt protocols. As we discussed earlier Robots.txt is one of the most often used techniques, but its operation is mostly based on the user-agent identification systems defined by the developers themselves. As a result, rightholders are required to keep up with technological changes, identify new crawlers, to watch announcements by AI businesses and periodically update their reservation systems.

The practical burden of this process should not be underestimated. When reservations are not a standard opt-out for TDM activities but have to be adapted to each company or crawler, they become more difficult to implement. Such fragmentation may make the exercise of copyright protection less practicable in practice.

In addition, many creators do not have the technical knowledge to properly implement machine-readable reservations. Smaller creators, independent artists, journalists, or photographers may not know that opt-out options exist or how such systems work technically. Still, failure to apply a technically correct reservation may still result in the unrestricted use of their works for AI training purposes. This means that, from the standpoint of rightholders, it is problematic that the successful protection of copyright increasingly rely not just on legal rights but also on technological literacy and ongoing monitoring capacities.

4.2.2 The Economic Impact of AI Outputs on Rightholders

Another major rightholders’ worry is related to the outputs of AI systems. As stated before in regard to the three-step test, the economic value of human-created works may be affected by GenAI outputs, so prejudicing the legitimate interests of rightholders.

Even where the AI systems do not directly reproduce copyrighted works, they may produce outputs that can compete with human authors in the market place. AI-generated images, words, music and other content may lower demand for some forms of creative labour by providing faster, cheaper alternatives. This is of particular significance where AI systems replicate styles of art, ideas or modes of expression learned from copyrighted training data.

In theory, the opt-out process gives rightholders the chance to stop such uses,

but in reality, its shortcomings severely reduce this protection. Many creators could not have adequate reservations in place or might not even know their works are being used for AI training. Thus, AI generated outputs can still lead to market dilution and income loss for certain classes of creators.

Thus, although TDM exceptions may in theory qualify for the three-step test, the practical effect of AI-generated outputs remains a cause for concern for the economic interests of rightholders and the long-term sustainability of creative markets.

4.2.3 The Significance of Safeguarding Human Creativity

Besides the short-term economic interests, the protection of rightholders is also crucial for the long-term sustainability of creativity itself. GenAI systems are inherently reliant on human-made works as training material. AI systems learn patterns, styles and structures from literature, music, journalism, photography, cinema and other creative works. Without an ongoing supply of new human creativity, the quality and variety of AI systems would progressively deteriorate.¹³⁷

Human producers generate new artistic directions, question social norms, and build novel forms of expression affected by cultural and social change. AI systems still depend on current datasets and cannot go beyond the knowledge given to them on their own. If the creative industries are damaged because creators lack the economic incentive to generate new works, AI systems might progressively reuse older materials and repetitive styles, rather than benefiting from new cultural input. Therefore, the protection of human creativity is good not only for right holders, but also for the long-term interests of AI advancement.

4.3 The Need for a New EU Approach to GenAI Training

As we have seen in the previous sections, the current EU framework on TDM faces problems in effectively protecting in practice either AI developers or rightholders. While the EU has tried to adopt a more cautious and balanced approach than jurisdictions such as the United States or Japan, the current framework seems increasingly incapable of providing either effective protection for rightholders or sufficient legal certainty for AI innovation. The framework formally seeks to balance both interests, yet in practice it often leaves both sides dissatisfied.

This concern has become even more significant following *GEMA v OpenAI*, as the approach taken in the case indicates that the legal position of AI

¹³⁷ Martin Senftleben, ‘Generative AI and Author Remuneration’ (2023) 54 *International Review of Intellectual Property and Competition Law* 1535–1560 <https://ssrn.com/abstract=4478370> accessed 04 May 2026.

developers under the EU TDM framework may become considerably more restrictive in the future.

The judgment poses wider questions about the future scope of the Article 4 TDM exception, especially by intimating a more restrictive approach to GenAI training. By distinguishing between preparatory extraction activities and the actual analytical training process, the decision creates substantial uncertainty regarding whether contemporary AI training can continue to rely on the EU TDM framework at all. If this approach continues to develop, the practical value of Article 4 for GenAI systems may become significantly weakened, especially for commercial AI developers relying on large-scale model training.

The judgment also points out that the current framework was not conceived with the technological realities of today's GenAI systems in mind. While the CDSM Directive sought to modernise EU copyright law for the digital age, the rapid pace of development of large scale generative AI has exposed shortcomings not fully anticipated at the time of the legislative process. This means that the current framework risks growing increasingly divorced from the practical and technological realities of modern AI development.

The EU attempted to reduce risk through a more protective regulatory model, but the growing deficiencies of the framework expose the need for a more technologically realistic and AI-centric approach to TDM regulation. Absent further clarification and evolution of regulation, the current system may be legally uncertain and practically ineffective in the age of GenAI.

5 Future Directions for TDM Regulation in the Era of GenAI

5.1 Comparative Approaches to TDM and GenAI Training

As noted in previous sections, some scholars argue that the EU may face difficulties competing with jurisdictions that adopt more flexible approaches toward AI training and TDM.¹³⁸ In terms of the US and Japan, more flexible copyright limitations, which offer wider possibilities for the development of AI, are often mentioned. The US system is mostly based on the doctrine of “fair use”, and the Japanese system on the concept of “non-enjoyment” use. Thus, observing these systems can provide useful comparative perspectives for evaluating the strengths and limitations of the current EU framework.

5.1.1 The US Fair Use Doctrine

The United States approaches TDM primarily through the fair use doctrine. Fair use is not really an exception or limitation to copyright, but instead a legal defence that might justify the unauthorized use of copyrighted works in specific situations. US courts do not provide a closed list of permitted uses, but instead evaluate each case on an individual basis by weighing factors such as the purpose of the use, the amount used, and the effect on the market for the original work.¹³⁹

Many scholars, including many EU scholars and authors argue that TDM and AI training may fall within the scope of fair use.¹⁴⁰ This view was strongly influenced by the case *Authors Guild v. Google, Inc.*, where US courts accepted the large-scale digitisation of copyrighted books for searchable databases as fair use.¹⁴¹ The courts stressed that the use was transformative because the works were used for computational analysis and search functions, not for ordinary reading purposes.

In this context, the US system is often seen as more flexible and innovation-friendly. But this flexibility also leads to a lot of legal uncertainty, since fair use is highly dependent on judicial interpretation and case-by-case analysis.

¹³⁸ Martin R F Senftleben, Tiziano Margoni and others, ‘Ensuring the Visibility and Accessibility of European Creative Content on the World Market—the Need for Copyright Data Improvement in the Light of New Technologies and the Opportunity Arising from Article 17 of the CDSM Directive’ (2022) 13 *JIPITEC* 67, 72–73.

¹³⁹ Charlotte Gerrish, *European Copyright Law and the Text and Data Mining Exceptions and Limitations: With a Focus on the DSM Directive, Is the EU Approach a Hindrance or Facilitator to Innovation in the Region?* (dissertation, Uppsala University 2019).

¹⁴⁰ Vrakas (n48); Edward Lee, ‘Technological Fair Use’ (2010) 83 *Southern California Law Review* 797; Matthew Sag, ‘Copyright and Copy-Reliant Technology.’ [2009] *Faculty Publications & Other Works*.

¹⁴¹ *Authors Guild v Google Inc* 804 F 3d 202 (2d Cir 2015).

As Larry Lessig said, fair use might become “the right to hire a lawyer”,¹⁴² as legal certainty often only arises at the end of an expensive litigation process. This may be particularly problematic for start-ups and SMEs.

The CDSM Directive, however, takes a more restrictive approach to TDM but also provides more predictability with its set rules and conditions. As Rosie Burbidge states, while more stringent regulation may seem less flexible, it could also encourage companies to resolve copyright issues sooner and offer more legal certainty in the longer term.¹⁴³

5.1.2 The Japanese “Non-Enjoyment” Approach

Japan has adopted one of the broadest approaches toward TDM in the world through Article 30-4 of the Japanese Copyright Act.¹⁴⁴ Unlike the EU system, the Japanese approach generally permits the use of copyrighted works for TDM purposes regardless of whether the activity is commercial or non-commercial, whether rightsholders apply reservations, or whether lawful access exists.¹⁴⁵ The provision allows for exploitation of works for activities such as extraction, comparison, classification and statistical analysis, as long as the use is not intended to “enjoy” the work itself and does not unreasonably prejudice the interests of the copyright holder.¹⁴⁶

The Japanese model is founded on the concept of the difference between “enjoyment” and “non-enjoyment” use. In this case, copyright protection is primarily associated with the human appreciation of a work in terms of intellectual, emotional or aesthetic enjoyment. By contrast, TDM and AI training are viewed as computational processes that use works as data rather than as creative expressions intended for human consumption. For this reason, the Japanese legislature considered that such activities should generally fall outside the core economic function of copyright protection.¹⁴⁷

Importantly, this concept is not entirely foreign to European legal thinking. Similar reasoning can already be found in earlier German case law through the concept of *Freier Werkgenuss*, where courts distinguished between copyright-relevant exploitation and mere technical or functional use of works.¹⁴⁸ German courts emphasized that copyright primarily protects the intellectual

¹⁴² Eric Goldman, ‘Why Google’s Fair Use Victory in Google Books Suit Is a Big Deal—And Why It Isn’t’ *Forbes* (14 November 2013) <https://www.forbes.com/sites/eric-goldman/2013/11/14/why-googles-fair-use-victory-in-google-books-suit-is-a-big-deal-and-why-it-isnt/> accessed 19 May 2026.

¹⁴³ Gerrish (n140); Interview with Rosie Burbidge, IP Partner at Gunner Cooke, London, 27 April 2019.

¹⁴⁴ *Copyright Act of Japan* (Act No 48 of 1970, as amended) art 30-4.

¹⁴⁵ Dermawan (n18) 54.

¹⁴⁶ *Ibid.*

¹⁴⁷ *Ibid.*

¹⁴⁸ *Ibid.*, BGH: BGH 04.10.1990 I ZR 139/89 “Operating System” (GRUR 1991, 449) 453.

or aesthetic enjoyment of works, such as reading a book, listening to music, or watching a film.¹⁴⁹ In this sense, the Japanese “non-enjoyment” approach reflects ideas that have also appeared, to some extent, within European copy-right theory.

Thus, the Japanese approach appears to provide one of the most technologically adaptable and innovation-friendly models currently available for regulating TDM in the context of GenAI training. Compared to the current EU framework, the Japanese model offers considerably broader flexibility for computational uses of protected works.

At the same time, such an approach may not be entirely incompatible with European copyright thinking. As noted above, concepts similar to “non-enjoyment” use have already appeared within European legal theory. Furthermore, the Japanese approach still contains an important safeguard by requiring that the use must not unreasonably prejudice the interests of the rightholder. This logic closely resembles the balancing rationale underlying the international three-step test, which is also deeply embedded within EU copyright law.

In this respect, the similarities between the Japanese model and existing European copyright principles may facilitate the future development of a more flexible and AI-oriented approach within the EU TDM framework.

5.2 Possible Improvements to the Existing EU Framework

Although this thesis argues that the current EU framework may require deeper structural and conceptual reform in the era of GenAI, it is nevertheless useful to consider several practical measures that could help alleviate some of the existing legal and technical difficulties surrounding TDM and AI training under the current system.

5.2.1 Clarification of Ambiguous Legal Concepts

One important issue concerns the lack of clarity surrounding concepts such as “lawful access” and “machine-readable” opt-outs. As discussed throughout previous chapters, the uncertain scope of these concepts continues to create legal unpredictability for both rightholders and AI developers. Clearer guidance and more harmonised technical standards could therefore improve legal certainty and reduce inconsistencies between Member States when applying the CDSM Directive.

5.2.2 Enhancing the Functionality of Opt-Out Systems

¹⁴⁹ Ibid.

One of the key challenges to overcome is the fragmented nature of current opt-out systems. As mentioned above, many creators may not have the necessary technical expertise to correctly implement reservations of rights, and AI developers often depend on unreliable systems for crawler identification. More standardised and user-friendly reservation mechanisms could therefore improve the practical application of Article 4(3).

One possible improvement could be the establishment of a centralised system for TDM reservations at the EU level. The European Commission intended to investigate the possibility of a centralised registry system for TDM reservations.¹⁵⁰ Such a system could reduce monitoring burdens for rightholders and provide a more harmonised framework for reservations of rights.

Similarly, Robots.txt systems could potentially become more effective through standardized vocabulary and more generic user-agent categories such as “-aiuser” or “-tdm.”¹⁵¹ More detailed standardized categories could even allow creators to prohibit specific uses, such as GenAI training, while still permitting other forms of AI-related analysis.¹⁵² These could simplify the identification of AI crawlers and reduce fragmentation within current reservation systems.

5.2.3 Measures Concerning TPMs and Access Restrictions

Additional concerns arise where TPMs unintentionally block lawful TDM activities. In some situations, researchers or developers may already possess lawful access to content, yet TPMs may still prevent the automated access necessary for TDM purposes. Mechanisms restoring access within a reasonable time where TPMs unlawfully interfere with lawful TDM activities could therefore improve the practical effectiveness of the framework.¹⁵³

Similar supervisory approaches already exist in certain Member States. For example, France established an authority responsible for overseeing TPMs and ensuring that copyright exceptions are not undermined in practice. A similar EU-wide supervisory mechanism, specifically dealing with TDM-related disputes, could possibly improve enforcement and reduce conflicts over lawful access.¹⁵⁴

5.2.4 Possible Remuneration-Based Solutions

Another possible direction could involve remuneration-based systems such

¹⁵⁰ European Commission, *Study to Assess the Feasibility of a Central Registry of Text and Data Mining Opt-Out Expressed by Rightsholders* (File No EC-CNECT/2025/OP/0002, 2025)

<https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/opportunities/tender-details/8726813a-bd9b-4f58-8679-01c80f7a1abf-CN> accessed 12 May 2026.

¹⁵¹ Havlikova (n84) 100.

¹⁵² *Ibid*, *Recommendation of Czech Association for Internet Development*.

¹⁵³ Dermawan (n18) 58.

¹⁵⁴ Gerrish (n139)

as “*levy models*”.¹⁵⁵ Instead of relying entirely on fragmented opt-outs and individual licensing during the training stage, AI companies could contribute financially to collective compensation mechanisms based on the commercial scale of their AI services. The collected funds could then potentially be distributed to creators through collecting societies or authors’ organisations.¹⁵⁶

Compared to the current input-oriented architecture, such a model could provide several practical advantages. First, it could reduce the significant transaction costs associated with negotiating individual licences for massive AI training datasets. Second, it may alleviate some of the transparency problems currently surrounding AI training because compensation would no longer depend on proving the use of every individual work within training datasets.¹⁵⁷ Instead, remuneration could be linked more generally to the commercial scale and economic success of AI systems themselves.

Although these proposals would not fully resolve the broader structural tensions discussed throughout this thesis, they may nevertheless contribute to improving the practical functioning of the current EU framework in the short term.

5.3 Final Remarks

Although the EU framework attempts to present itself as a balanced model between copyright protection and technological innovation, this thesis has demonstrated that such a balance remains increasingly difficult to achieve in practice. Compared to more innovation-oriented approaches such as the US fair use doctrine or the Japanese “non-enjoyment” model, the EU system adopts a considerably narrower and more restrictive approach toward AI training. However, despite this stricter approach, the current framework still struggles to provide sufficient legal certainty, practical enforceability and coherent doctrinal guidance for either AI developers or rightholders.

In particular, the uncertainty surrounding the recent *GEMA v OpenAI* decision demonstrates the growing complexity of applying the existing TDM framework to contemporary GenAI systems. Although the judgment cannot yet be regarded as representing the final position of EU law, the approach taken by the Court raises important concerns regarding the future approaches for interpretation of Article 4 of the CDSM Directive. In this respect, the case illustrates whether the position of AI developers under the EU framework may become even more restrictive in the future.

More fundamentally, the difficulties examined throughout this thesis suggest that the current problems are not merely issues of interpretation, but reflect a

¹⁵⁵ Senftleben (n138).

¹⁵⁶ Ibid.

¹⁵⁷ Ibid.

deeper structural mismatch between the architecture of the CDSM Directive and the technological realities of modern GenAI systems. Although Article 4 has increasingly become the primary legal basis relied upon for AI training activities, the Directive was not originally designed to regulate large-scale generative AI. Instead, the EU framework attempts to fit modern AI training into legal concepts originally developed for more traditional forms of TDM and analytical data processing.

As demonstrated throughout this thesis, issues surrounding lawful access, machine-readable opt-outs, fragmented reservation systems, transparency obligations increasingly expose the limitations of the current framework. The existing system therefore risks becoming progressively disconnected from the practical realities of contemporary AI development.

For this reason, future EU regulation may eventually need to move beyond incremental clarification of the current TDM provisions. While the proposals discussed above may improve the practical functioning of the existing framework to some extent, they are unlikely to fully resolve the broader structural tensions created by GenAI training. Instead, the rapid development of AI technologies may necessitate a more autonomous, technologically feasible and AI-focused regulatory approach that is intended to govern the relationship between copyright law and GenAI training.

Future reforms need to aim at not only protecting copyright, but also at building a framework that can realistically accommodate the technological functioning of contemporary AI systems. Instead of continuing to apply traditional TDM concepts to technologies operating on a completely different scale and purpose, the EU might eventually need to rethink its copyright framework in a more fundamental way in order to remain both technologically relevant and legally effective in the GenAI era.

6 Conclusion

This thesis examined the following research question: *To what extent does the current EU legal framework on Text and Data Mining, particularly under the CDSM Directive, effectively balance the protection of copyright with the promotion of technological innovation in the context of GenAI training?*

Chapter 2 addressed the position of TDM in the training of GenAI, its relation to copyright law and the legal developments that happened prior to the enactment of the CDSM Directive. *Chapter 3* looked at Articles 3 and 4 of the CDSM Directive and demonstrated that although the Directive seeks to achieve a balance between the protection of copyright and the encouragement of technical innovation, major legal and practical problems remain. In particular, serious uncertainties persist for both rightholders and AI developers due to difficulties with lawful access, machine-readable opt-outs, fragmented reservation systems and the application of the three-step test. This uncertainty has been further compounded by the recent *GEMA v OpenAI* ruling, which raises broader doubts regarding the future approaches about Article 4 in the context of GenAI training, as described in the chapter.

Chapter 4 explores these challenges from the perspective of right holders and AI developers and shows that the current EU framework is failing to adequately safeguard either side in practice. The EU has attempted to adopt a more cautious and balanced approach than jurisdictions like the United States or Japan. But the analysis showed that the current framework frequently does not offer appropriate protection for rightholders or adequate legal certainty for AI innovation. Finally, in *Chapter 5*, some future options for regulation of TDM and AI training were considered. It compared the EU framework with more flexible systems in the U.S. and Japan, in particular the Japanese “non-enjoyment” model, and looked at prospective improvements to the current EU framework. To conclude, the chapter claimed that future approaches may need to be ever more technology adaptable and AI centred.

This thesis concludes that while the EU framework represents a major attempt to combine copyright protection and technological innovation, the current TDM approach is becoming insufficient to the realities of modern GenAI systems. The fast growth of AI technology has revealed structural constraints and legal concerns that could not have been foreseen at the time of the implementation of the CDSM Directive. In particular, the increasing difficulties of lawful access, opt-out systems, AI training and the interpretation of Article 4 indicate that the current framework may no longer be able to effectively control the modern GenAI training. Consequently, future EU regulation may eventually need to move beyond the current TDM structure toward a more technologically realistic, AI-oriented and autonomous regulatory approach specifically designed for the relationship between copyright law and GenAI systems.

7 Bibliography

7.1 Case Law

ACI Adam BV and Others v Stichting de ThuisKopie and Stichting Onderhandeligen ThuisKopie vergoeding (C-435/12) EU:C:2014:254.

Authors Guild v Google Inc 804 F 3d 202 (2d Cir 2015).

Authors Guild v OpenAI Inc (SDNY, Case No 1:23-cv-08292, 19 September 2023).

BGH, Operating System (4 October 1990) I ZR 139/89, *GRUR* 1991, 449.

Copydan Båndkopi v Nokia Danmark A/S (C-463/12) EU:C:2015:144.

Football Association Premier League Ltd and Others v QC Leisure and Others (C-403/08 and C-429/08) EU:C:2011:631.

Funke Medien NRW GmbH v Bundesrepublik Deutschland (C-469/17) EU:C:2019:623.

GEMA v OpenAI (Landgericht München I, 11 November 2025) Case No 42 O 14139/24.

GS Media BV v Sanoma Media Netherlands BV and Others (C-160/15) EU:C:2016:644.

Infopaq International A/S v Danske Dagblades Forening (C-5/08) EU:C:2009:465.

Land Nordrhein-Westfalen v Dirk Renckhoff (C-161/17) EU:C:2018:634.

Luksan v van der Let (C-277/10) EU:C:2012:65.

Mc Fadden v Sony Music Entertainment Germany GmbH (C-484/14) EU:C:2016:689.

Pelham GmbH and Others v Hütter and Schneider-Esleben (C-476/17) EU:C:2019:624.

Poland v Parliament and Council (C-401/19) EU:C:2022:297.

Robert Kneschke v LAION eV (Regional Court of Hamburg, Case No 310 O 227/23, 27 September 2024).

Spiegel Online GmbH v Volker Beck (C-516/17) EU:C:2019:625.

Stichting Brein v Ziggo BV and XS4ALL Internet BV (C-610/15) EU:C:2017:456.

The British Horseracing Board Ltd and Others v William Hill Organization Ltd (C-203/02) EU:C:2004:695.

VG Bild-Kunst v Stiftung Preußischer Kulturbesitz (C-392/19) EU:C:2021:181.

7.2 Treaties, Conventions, Guidelines and Legislation

A. EU

Charter of Fundamental Rights of the European Union [2012] OJ C326/391.

Consolidated Version of the Treaty on the Functioning of the European Union [2016] OJ C202/47.

Directive 2001/29/EC of the European Parliament and of the Council on the harmonisation of certain aspects of copyright and related rights in the information society [2001] OJ L167/10.

Directive (EU) 2019/790 of the European Parliament and of the Council on Copyright and Related Rights in the Digital Single Market [2019] OJ L130/92.

Regulation (EU) 2016/679 of the European Parliament and of the Council on the Protection of Natural Persons with regard to the Processing of Personal

Data and on the Free Movement of Such Data (General Data Protection Regulation) [2016] OJ L119/1.

Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) [2024] OJ L 2024/1689.

B. International Sources

Agreement on Trade-Related Aspects of Intellectual Property Rights (adopted 15 April 1994, entered into force 1 January 1995) 1869 UNTS 299.

Berne Convention for the Protection of Literary and Artistic Works (adopted 9 September 1886, revised at Paris 24 July 1971, amended 28 September 1979) 828 UNTS 221.

WIPO Copyright Treaty (adopted 20 December 1996, entered into force 6 March 2002) 2186 UNTS 121.

World Trade Organization, *United States – Section 110(5) of the US Copyright Act* (Report of the Panel) WT/DS160/R (15 June 2000).

C. National Legislation

Copyright Act of Japan (Act No 48 of 1970, as amended).

Copyright and Rights in Performances (Research, Education, Libraries and Archives) Regulations (UK SI 2014/1372).

Copyright, Designs and Patents Act 1988 (UK), s 29A.

Estonian Copyright Act, 11 November 1992

Loi n° 2016-1321 du 7 octobre 2016 pour une République numérique (France).

7.3 Books, Journal Articles, Research Papers and Studies

Brook M, Murray-Rust P and Oppenheim C, ‘The Social, Political and Legal Aspects of Text and Data Mining (TDM)’ (2014) 20 *D-Lib Magazine*

<http://www.dlib.org/dlib/november14/brook/11brook.html> accessed 14 May 2026.

Calabrese B, ‘Scientific TDM Exception and Communication to the Public: Did Italians Do It Better ... or at Least Not Worse?’ (2022) 17 *Journal of Intellectual Property Law & Practice* 399.

Carroll MW, ‘Copyright and the Progress of Science: Why Text and Data Mining Is Lawful’ (2019) 53 *UC Davis Law Review* 893.

Dermawan A, ‘Text and Data Mining Exceptions in the Development of Generative AI Models: What the EU Member States Could Learn from the Japanese “Non-enjoyment” Purposes?’ (2024) 27(1) *The Journal of World Intellectual Property* 45.

Ducato R and Strowel A, ‘Ensuring Text and Data Mining: Remaining Issues with the EU Copyright Exceptions and Possible Ways Out’ (2021) 43 *European Intellectual Property Review* 322.

European Commission, ‘Commission Staff Working Document – Impact Assessment on the Modernisation of EU Copyright Rules Accompanying the Proposal for a Directive on Copyright in the Digital Single Market’ SWD(2016) 301 final.

European Commission, *Study to Assess the Feasibility of a Central Registry of Text and Data Mining Opt-Out Expressed by Rightsholders* (File No EC-CNECT/2025/OP/0002, 2025) <https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/opportunities/tender-details/8726813a-bd9b-4f58-8679-01c80f7a1abf-CN> accessed 24 May 2026.

Geiger C, ‘Elaborating a Human Rights-Friendly Copyright Framework for Generative AI’ (2024) 55 *International Review of Intellectual Property and Competition Law* 1129.

Geiger C, Frosio G and Bulayenko O, ‘Text and Data Mining in the Proposed Copyright Reform: Making the EU Ready for an Age of Big Data?’ (2018) 49 *International Review of Intellectual Property and Competition Law* 814.

Geiger C, Frosio G and Bulayenko O, ‘Text and Data Mining: Articles 3 and 4 of the Directive 2019/790/EU’ in Concepción Saiz García and Raquel Evangelio Llorca (eds), *Propiedad intelectual y mercado único digital europeo*

(Tirant lo Blanch 2019) 27–71 <https://ssrn.com/abstract=3470653> accessed 5 May 2026.

Gerrish C and Skavlan AM, ‘European Copyright Law and Text and Data Mining Exceptions and Limitations: In Light of the Recent DSM Directive, is the EU Approach a Hindrance or Facilitator to Innovation in the Region?’ (2019) 2(2) *SIPLR* 58.

Griffiths J, Synodinou T and Xalabarder R, ‘Comment of the European Copyright Society Addressing Selected Aspects of the Implementation of Articles 3 to 7 of Directive (EU) 2019/790 on Copyright in the Digital Single Market’ (2023) 72 *GRUR International* 22.

Guibault L and others, *Study on the Implementation and Effect in Member States’ Laws of Directive 2001/29/EC on the Harmonisation of Certain Aspects of Copyright and Related Rights in the Information Society* (Report prepared for the European Commission, DG Internal Market, ETD/2005/IM/D1/91, 2007).

Havlíková S, ‘Technical Challenges of Rightsholders’ Opt-out From Gen AI Training after *Robert Kneschke v LAION*’ (2025) 16 *JIPITEC* 93 <https://www.jipitec.eu/jipitec/article/view/422> accessed 14 May 2026.

Karim, Mohamed Ataul, ‘GEMA versus OpenAI: The Key Takeaways’ (4iP Council, 2025) https://www.4ipcouncil.com/application/files/1017/6545/5349/Paper-GEMA_v_OpenAI_02-12-25.pdf accessed 22 May 2026.

Lee E, ‘Technological Fair Use’ (2010) 83 *Southern California Law Review* 797.

Lindholm M, ‘Text and Data Mining under Finnish Copyright Law before and after the DSM Directive’ (2020).

Löbbling L, Handschigl Ch, Hofmann K and Schwedhelm J, ‘Navigating the Legal Landscape: Technical Implementation of Copyright Reservations for Text and Data Mining in the Era of AI Language Models’ (2023) 14 *JIPITEC* 499.

Lucchi N, *Generative AI and Copyright – Training, Creation, Regulation* (European Parliament, Policy Department for Justice, Civil Liberties and Institutional Affairs 2025).

Manteghi M, ‘The Insufficiency of the EU’s Text and Data Mining Exceptions for Using Artificial Intelligence’ (2022) 44(11) *European Intellectual Property Review* 651.

Margoni T, ‘TDM and Generative AI: Lawful Access and Opt-outs’ (2024) *Auteurs & Media* 172 <https://ssrn.com/abstract=5036164> accessed 19 May 2026.

Margoni T and Kretschmer M, ‘A Deeper Look into the EU Text and Data Mining Exceptions: Harmonisation, Data Ownership, and the Future of Technology’ (2022) 71 *GRUR International* 685.

McCracken P and Raub E, ‘Licensing Challenges Associated With Text and Data Mining: How Do We Get Our Patrons What They Need?’ (2023) 11 *Journal of Librarianship and Scholarly Communication* 13 <https://www.iastatedigitalpress.com/jlsc/article/id/15530/> accessed 14 May 2026.

Mezei P, ‘A Saviour or a Dead End? Reservation of Rights in the Age of Generative AI’ (2024) 46(7) *European Intellectual Property Review* 461 <https://ssrn.com/abstract=4695119> accessed 10 May 2026.

Pereira ALD, ‘TDM Copyright for AI in Europe: A View from Portugal’ (2023) 18 *Journal of Intellectual Property Law & Practice* 900.

Rosati E, ‘Copyright as an Obstacle or an Enabler? A European Perspective on Text and Data Mining and its Role in the Development of AI Creativity’ (2020) 27(2) *Asia Pacific Law Review* 198–217 <https://ssrn.com/abstract=3452376> accessed 22 April 2026.

Rosati E, ‘No Step-Free Copyright Exceptions: The Role of the Three-Step in Defining Permitted Uses of Protected Content (Including TDM for AI-Training Purposes)’ (2024) 46(5) *European Intellectual Property Review* 262 <https://ssrn.com/abstract=4629528> accessed 2 May 2026.

Rosati E, ‘The Exception for Text and Data Mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market: Technical Aspects’ (Briefing requested by the JURI Committee, Policy Department for Citizens’ Rights and Constitutional Affairs, European Parliament 2018)

[https://www.europarl.europa.eu/Reg-DATA/etudes/BRIE/2018/604942/IPOL_BRI\(2018\)604942_EN.pdf](https://www.europarl.europa.eu/Reg-DATA/etudes/BRIE/2018/604942/IPOL_BRI(2018)604942_EN.pdf) accessed 20 April 2026.

Sag M, ‘Copyright and Copy-Reliant Technology’ [2009] Faculty Publications & Other Works.

Senftleben M, ‘Generative AI and Author Remuneration’ (2023) 54 *International Review of Intellectual Property and Competition Law* 1535–1560 <https://ssrn.com/abstract=4478370> accessed 4 May 2026.

Senftleben M, ‘Text and Data Mining, Generative AI, and the Copyright Three-Step Test’ (2026) 57(1) *International Review of Intellectual Property and Competition Law* 67–107 https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5373903 accessed 5 May 2026.

Senftleben MRF, *Copyright, Limitations and the Three-Step Test: An Analysis of the Three-Step Test in International and EC Copyright Law* (Kluwer Law International 2004).

Senftleben MRF, Margoni T and others, ‘Ensuring the Visibility and Accessibility of European Creative Content on the World Market—the Need for Copyright Data Improvement in the Light of New Technologies and the Opportunity Arising from Article 17 of the CDSM Directive’ (2022) 13 *JIPITEC* 67.

Smits JM, ‘What Is Legal Doctrine? On the Aims and Methods of Legal-Dogmatic Research’ in R van Gestel, H Micklitz and E L Rubin (eds), *Rethinking Legal Scholarship: A Transatlantic Dialogue* (Cambridge University Press 2017) 207–228.

Truyens M and Van Eecke P, ‘Legal Aspects of Text Mining’ in Nicoletta Calzolari and others (eds), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14)* (European Language Resources Association (ELRA) 2014) http://www.lrec-conf.org/proceedings/lrec2014/pdf/452_Paper.pdf accessed 15 May 2026.

Ueno T, ‘The Flexible Copyright Exception for “Non-Enjoyment” Purposes: Recent Amendment in Japan and Its Implication’ (2021) 70 *GRUR International* 145.

Vrakas GD, ‘A Literature Review of “Lawful” Text and Data Mining’ (2024) 4 *Open Research Europe* 153 <https://doi.org/10.12688/open-reseurope.18013.2> accessed 10 May 2026.

7.4 Master Theses and Dissertations

Gerrish C, *European Copyright Law and the Text and Data Mining Exceptions and Limitations: With a Focus on the DSM Directive, Is the EU Approach a Hindrance or Facilitator to Innovation in the Region?* (dissertation, Uppsala University 2019).

7.5 Online Sources

Goldman E, ‘Why Google’s Fair Use Victory in Google Books Suit Is a Big Deal—And Why It Isn’t’ *Forbes* (14 November 2013) <https://www.forbes.com/sites/ericgoldman/2013/11/14/why-googles-fair-use-victory-in-google-books-suit-is-a-big-deal-and-why-it-isnt/> accessed 19 May 2026.

Hugenholtz PB, ‘The New Copyright Directive: Text and Data Mining (Articles 3 and 4)’ (Kluwer Copyright Blog, 24 July 2019) <https://copyrightblog.kluweriplaw.com/2019/07/24/the-new-copyright-directive-text-and-data-mining-articles-3-and-4/> accessed 14 May 2026.

Keller P, ‘Generative AI and Copyright: Convergence of Opt-outs?’ (Kluwer Copyright Blog, 23 November 2023) <https://legalblogs.wolterskluwer.com/copyright-blog/generative-ai-and-copyright-convergence-of-opt-outs/> accessed 14 May 2026.

Interview with Rosie Burbidge, IP Partner at Gunner Cooke, London, 27 April 2019.

7.6 Technical, Scientific and Non-Legal Sources Related to TDM and AI Training

Koster M, Zoltán I, Zeller H and Sassman L, ‘RFC 9309: Robots Exclusion Protocol’ (September 2022) <https://www.rfc-editor.org/rfc/rfc9309.html> accessed 14 May 2026.

OpenAI, ‘GPT-4 Technical Report’ (2023) arXiv:2303.08774 <https://arxiv.org/abs/2303.08774> accessed 24 April 2026.

Ramesh A, Dhariwal P, Nichol A, Chu C and Chen M, ‘Hierarchical Text-Conditional Image Generation with CLIP Latents’ (2022) arXiv:2204.06125 <https://arxiv.org/pdf/2204.06125.pdf> accessed 21 April 2026.

Rombach R and others, ‘High-Resolution Image Synthesis with Latent Diffusion Models’ (2022) *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 10684 <https://arxiv.org/abs/2112.10752> accessed 20 April 2026.

Schuhmann C, Vencu R, Beaumont R, Coombes T, Gordon C, Katta A, Kaczmarczyk R and Jitsev J, ‘LAION-5B: A New Era of Open Large-Scale Multi-Modal Datasets’ (2022) <https://laion.ai/blog/laion-5b/> accessed 20 April 2026.