TURUN
YLIOPISTO
UNIVERSITY
OF TURKU

Readjusting the EU's Regulation of Text and Data Mining in the Age of Artificial Intelligence

Maryna Manteghi

TURUN YLIOPISTON JULKAISUJA – ANNALES UNIVERSITATIS TURKUENSIS
SARJA – SER. B OSA – TOM. 750 | HUMANIORA | TURKU 2025

TURUN
YLIOPISTO
UNIVERSITY
OF TURKU

READJUSTING THE EU'S REGULATION OF TEXT AND DATA MINING IN THE AGE OF ARTIFICIAL INTELLIGENCE

Maryna Manteghi

TURUN YLIOPISTON JULKAISUJA – ANNALES UNIVERSITATIS TURKUENSIS
SARJA – SER. B OSA – TOM. 750 | HUMANIORA | TURKU 2025

University of Turku

Faculty of Law
Intellectual Property Law
Doctoral Programme in Law

Supervised by

Professor, Tuomas Mylly
University of Turku
Turku, Finland

Docent, Juha Vesala
University of Turku
Turku, Finland

Reviewed by

Professor, Christophe Geiger
Luiss Guido Carli University
Rome, Italy

Professor, Rosa Ballardini
University of Lapland
Rovaniemi, Finland

Opponent

Professor, Rosa Ballardini
University of Lapland
Rovaniemi, Finland

ISBN 978-952-02-0424-2 (PRINT)
ISBN 978-952-02-0425-9 (PDF)
ISSN 0082-6987 (Print)
ISSN 2343-3191 (Online)
Painosalama, Turku, Finland 2025

UNIVERSITY OF TURKU
Faculty of Law
Intellectual Property Law
MARYNA MANTEGHI: Readjusting the EU's Regulation of Text and Data Mining in the Age of Artificial Intelligence
Doctoral Dissertation, 102 [74] pp.
Doctoral Programme in Law
November 2025

ABSTRACT

This doctoral dissertation examines whether the EU's current regulation of text and data mining (TDM) adequately considers the interests and implications related to technological developments and the fundamental freedoms of research and information. The study establishes three key aspects for discussion: first, it analyses how copyright and the *sui generis* database right may restrict the use of TDM in the EU; second, it explores the reasons for adopting specific TDM exceptions, and how the situation has evolved following the introduction of Articles 3 and 4 of Directive (EU) 2019/790 on Copyright and Related Rights in the Digital Single Market; and third, it identifies the most effective ways to improve the current EU regulation of TDM.

The dissertation consists of this summary and four substudies. The first article provides background on the topic and analyses potential improvements to the scope of TDM exceptions during their national transpositions. The second substudy examines the legal framework governing TDM from the perspective of fundamental rights and freedoms. The third publication reviews Article 35 of the proposed Data Act as an alternative mechanism for researchers. Lastly, the fourth article addresses the growing copyright-related concerns arising from the training of generative AI systems and explores potential solutions.

The findings of this dissertation emphasise the need to readjust the current EU regulation of TDM to ensure that it keeps pace with the rapid development of AI while remaining aligned with fundamental constitutional values. Accordingly, the research proposes changes to TDM exceptions aimed at establishing a long-lasting balance between the rights and interests of rightholders and those of TDM users.

KEYWORDS: copyright, text and data mining, CDSM Directive's TDM exceptions, digital and AI technology, generative AI models, fundamental rights, freedom to information and research, EU law

TURUN YLIOPISTO
Oikeustieteellinen tiedekunta
Kansainvälinen oikeus
MARYNA MANTEGHI: Euroopan unionin tekstin- ja tiedonlouhintaa koskevan sääntelyn uudelleentarkastelu tekoälyn aikakaudella
Väitöskirja, 102 [74] s.
Oikeustieteen tohtoriohjelma
Kuukausi 2025

TIIVISTELMÄ

Tämä väitöskirja tarkastelee, ottaako Euroopan unionin nykyinen tekstin- ja tiedonlouhintaa (TDM) koskeva sääntely riittävästi huomioon teknologiseen kehitykseen sekä tutkimuksen ja tiedonsaannin perusvapauksiin liittyvät intressit ja vaikutukset. Tutkimus rakentuu kolmeen keskeiseen näkökulmaan: ensinnäkin se analysoi, miten tekijänoikeus ja *sui generis* -tietokantaoikeus voivat rajoittaa TDM:n käyttöä Euroopan unionissa; toiseksi se tarkastelee syitä TDM-poikkeusten säätämiselle sekä sitä, miten tilanne on kehittynyt Euroopan parlamentin ja neuvoston direktiivin (EU) 2019/790 tekijänoikeudesta ja lähioikeuksista digitaalisilla sisämarkkinoilla 3 ja 4 artiklan käyttöönoton jälkeen; ja kolmanneksi se tunnistaa tehokkaimmat keinot nykyisen EU:n TDM-sääntelyn parantamiseksi.

Väitöskirja koostuu tästä yhteenvedosta ja neljästä osatutkimuksesta. Ensimmäinen artikkeli tarjoaa aiheeseen liittyvän taustan ja analysoi mahdollisia parannuksia TDM-poikkeusten soveltamisalaan niiden kansallisen täytäntöönpanon yhteydessä. Toinen osatutkimus tarkastelee TDM:ää koskevaa oikeudellista kehystä perusoikeuksien ja -vapauksien näkökulmasta. Kolmas julkaisu arvioi ehdotetun dataasetuksen 35 artiklaa vaihtoehtoisena mekanismina tutkijoille. Viimeinen artikkeli käsittelee generatiivisten tekoälyjärjestelmien kouluttamiseen liittyviä tekijänoikeudellisia huolia ja tarkastelee mahdollisia ratkaisuja.

Väitöskirjan tulokset korostavat tarvetta tarkistaa ja ajanmukaistaa EU:n nykyistä TDM-sääntelyä siten, että se pysyy tekoälyn nopean kehityksen tahdissa ja säilyttää samalla yhteyden keskeisiin perustuslaillisiin arvoihin. Tutkimus ehdottaa muutoksia TDM-poikkeuksiin tavoitteenaan saavuttaa kestävä tasapaino oikeudenhaltijoiden ja TDM:n käyttäjien oikeuksien ja etujen välillä.

AVAINSANAT: tekijänoikeus, tekstin- ja tiedonlouhinta, DSM-direktiivin TDM-poikkeukset, digitaalinen ja tekoälyteknologia, generatiiviset tekoälymallit, perusoikeudet, tiedon ja tutkimuksen vapaus, EU-oikeus

Acknowledgements

My journey with the University of Turku began many years ago when I came to Finland to study in the Master's Degree Programme in Law and Information Society. At that time, I understood that I was passionate about issues related to the relationship between copyright and digital technologies. After my graduation, I decided to continue my career in academia because I felt that I could offer a lot as a researcher. I worked on the research project *Profit4 Digital Futures*, funded by the Academy of Finland, before I applied for the Doctoral Programme in Law. The choice of topic was deliberate because I already had two publications related to the legal aspects of text and data mining technology. I was keen on continuing my research in this area. My prospective supervisor, Professor Tuomas Mylly, kindly agreed to support me in my endeavours.

However, life does not always bring only positive moments. Sometimes challenges hit hard. I lost my beloved grandmother. Moreover, the beginning of my PhD studies coincided with the start of the war in Ukraine, my motherland. I remember those days with tears in my eyes. I was heartbroken and utterly devastated because my family, my home, and my country were under constant threat. Despite these feelings, I decided to be as strong as my people were. The brave Ukrainian people who, at the beginning of the war, came to the streets and tried to stop the cruel russian troops and their tanks with empty hands have inspired me. Our President Volodymyr Zelenskyy, a man of dignity and a courageous leader, has inspired me. I am proud to be Ukrainian and to represent my country in the scientific field.

Hard work and proper time management helped me manage my feelings and emotions, and led me to this stage where I am now writing this introduction. It has been an emotionally challenging period in my life, but I have really enjoyed doing research, developing new perspectives, and analysing complex issues. I am truly happy because, during this journey, I have not been alone.

This work would not have been possible without the guidance and support of my supervisor, Professor Tuomas Mylly. I am very proud to be able to work under the supervision of such a prominent and brilliant scholar. He has offered constructive criticism while always believing in my abilities as a researcher. It is an honour for me to work with someone who is not only a true professional but also a kind and supportive person. Professor, your support has shaped my academic journey more than words can express! I am also thankful to my second supervisor, Docent Juha Vesala, who joined us at the final stage of my PhD and provided valuable feedback on my work.

I would like to express my special gratitude to Professor Mika Viljanen, Dean of the Faculty of Law. His meaningful guidance and insightful discussions have greatly

stimulated my research thinking and influenced my future direction. I also wish to thank our research coordinator, Kirsi Tuohela, for her invaluable advice and guidance.

I extend my sincere gratitude to Professor Rosa Ballardini for her invaluable contributions as both my pre-examiner and opponent. Thank you for your insightful comments, your dedication, and for honouring me with your presence here in Turku on such a special day in my life. I also wish to express my appreciation to Professor Christophe Geiger for examining my dissertation and offering valuable feedback.

I would like to thank the Faculty of Law of the University of Turku for providing me with the means to conduct my research over the years. I would also like to thank the Finnish Society of Sciences and Letters, the Turku University Foundation, and the Otto A. Malm Foundation for financially supporting my research.

Last but not least, I would like to express my gratitude to my family, the most important people in my life. I am thankful to my husband, who has always supported me throughout our more than 20 years of marriage. He has always had confidence in me and encouraged me to continue even when I was ready to give up. Thank you for your unwavering love, support, patience, and belief in me. You are my rock and my greatest supporter! Everything I have achieved in my life is thanks to you. I will love you always and forever! I am also grateful to my beloved sister for her continuous encouragement and unwavering belief in me throughout this journey and in life, which means the world to me. I am lucky to have such a strong, supportive, and reliable sister. Having you in my life is a true gift. You are the best sister in the world! I am so proud of you!

I thank my mother, who has always supported me, kept her worries in her heart, and never let them burden me. I became a lawyer only because of you. Many years ago, you stayed by my side when I was applying to law school despite financial and other challenges. Thank you for believing in me! Thank you for the hard work you put in to support me and my sister. I would like to thank my mother-in-law for her constant encouragement and support. I am really grateful for your kindness and warm attitude towards me. You believed in me even when I doubted myself. Your presence in my life is incredibly precious and valuable, and I truly cherish it.

I also want to express my deep gratitude to my treasured grandmother, who passed away several years ago. It was the most heartbreaking moment of my life! Thank you for your endless love and gentle wisdom. Thank you for your kindness, unwavering support, and care. I keep memories of you close to my heart, and I am forever grateful for everything you have done for me. I would like to thank other important people who are no longer with us, such as my adored grandmother-in-law, my esteemed grandfather, and my dear father. Finally, I would like to say that I am truly grateful for all of you, whether still by my side or living on in my heart. I feel incredibly proud to have such a wonderful Family.

Turku 2025,
Maryna Manteghi

Table of Contents

List of Original Publications

This dissertation is based on the following original publications:

I Maryna Manteghi, 'The Insufficiency of the EU's Text and Data Mining Exceptions for Using Artificial Intelligence' (2022) 44 (11) European Intellectual Property Review 652.

II Maryna Manteghi, 'In Search of Balance: Text, Data Mining and Copyright in the Digital Single Market Directive from a Fundamental Rights Perspective' (2023) 48 (4) European Law Review 443.

III Maryna Manteghi, 'Overcoming Barriers to Text and Data Mining in the Era of ChatGPT: The Proposed Data Act as a Game-Changer' (2024) 73 (1) GRUR International, 34.

IV Maryna Manteghi, 'Can Text and Data Mining Exceptions and Synthetic Data Training Mitigate Copyright-Related Concerns in Generative AI?' (2024) 16 (2) Law, Innovation and Technology 663.

1 Introduction

1.1 Background

Data-driven innovation has become a primary engine of progress and growth in the twenty-first century.[1] Advanced forms of data usage have triggered the emergence of new actors and the development of novel business strategies in the digital economy.[2] However, the rapid increase in digital interactions (social media, digital publishing, e-commerce, etc.) has resulted in a massive growth of digital data. Advanced analytical techniques have become indispensable for managing and processing Big Data stored within the global information network. In this sense, text and data mining (TDM) is a highly valued technology. It aims to analyse structured and unstructured raw digital data in order to generate new knowledge, such as patterns, correlations, and insights.[3]

Nowadays, many private and public actors across various fields leverage TDM to improve their operations and meet higher standards.[4] For instance, TDM is at the core of rapidly emerging generative AI (GenAI) models such as ChatGPT and Midjourney. Despite the benefits, the use of TDM may also raise certain legal issues. To mine, a computer must access, collect, and copy large volumes of digital data.[5] The obtained materials may consist of data protected by copyright or the *sui generis* database right. The use of protected works for mining without prior authorisation could lead to potential infringement of copyright or database owners' exclusive rights.

1 Maryna Manteghi, 'In Search of Balance: Text, Data Mining and Copyright in the Digital Single Market Directive from a Fundamental Rights Perspective' (2023) 48 European Law Review 422, 443.

2 Jose Ramon Saura, Domingo Ribeiro-Soriano and Daniel Palacious-Marques, "From User generated Data to Data-driven Innovation: A Research Agenda to Understand User Privacy in Digital Markets" (2021) 60 IJIM 1, 1.

3 Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC [2019] *OJ L 130/92*, art. 2 (2).

4 Maryna Manteghi, 'Overcoming Barriers to Text and Data Mining in the Era of ChatGPT: The Proposed Data Act as a Game-Changer' (2024) 73(1) GRUR International 34, 34.

5 For a general overview of TDM techniques and methods see Han Jiawei, *Data mining: Concepts and Techniques* (3rd edition, Elsevier 2011).

In 2019, the EU legislator adopted two specific TDM exceptions under Arts. 3 and 4 of the Directive on Copyright in the Digital Single Market (the CDSM Directive),[6] aimed at providing more legal certainty regarding the use of TDM within the Digital Single Market (DSM). These provisions emerged as a new hope for better regulation of TDM in the EU, which has been recognised as a crucial tool for research and innovation.[7] Even though the exceptions might mitigate copyright-related concerns to some extent, they may not completely resolve the conflict between competing interests involved in TDM.

The issue of TDM regulation has given rise to one of the most complicated and intensive debates in modern copyright law. Two main concerns can be identified: first, there is a fear that the broad protection afforded to copyright and *sui generis* database owners may impose extensive restrictions on TDM; second, many doubt that the current EU regulation of TDM is able to balance competing rights and interests at stake. In particular, new challenges brought by recent advancements in AI may require the readjustment of the legal framework for TDM in the EU. There have been extensive discussions on the topic, namely, attempts to find a resolution. This study contributes to these discussions by introducing novel critical perspectives, insights, and solutions in light of fundamental rights. The adopted approach offers practical guidance to all stakeholders involved, distinguishing this research from other studies.

1.2 Research Objectives and Questions

The primary objective of this dissertation is to examine the current EU regulation of TDM from both legal and technological perspectives in order to identify legislative flaws and propose essential improvements. Therefore, the main research question of this dissertation is:

> ***Does the EU's regulation of TDM sufficiently consider the interests related to technological developments and the fundamental freedoms of research and information?***

The main research question is approached through several specific sub-questions.

First, the dissertation examines how copyright and *sui generis* database rights could restrict the use of TDM in the EU. This includes the analysis of relevant

6 CDSM Directive, arts. 3 and 4.
7 CDSM Directive, recital 8.

exclusive rights, in particular, the copyright holder's broad right to reproduction[8] and the *sui generis* database right[9]. Moreover, the dissertation discusses the pre-existing exceptions and limitations provided under the InfoSoc Directive[10] and the Database Directive,[11] and their application to TDM. Analysing the situation before the adoption of the CDSM Directive may help to explain the reasons for adopting specific copyright exceptions targeted at TDM use in the EU.

Second, the study evaluates how the situation has changed with the introduction of TDM exceptions and whether these legal solutions have eased the restrictions on the right to TDM. Have they brought new hope or greater frustration in the AI-driven age? What advantages and disadvantages do they pose for research organisations and businesses? What are the specific conditions and limitations that users are required to comply with to enjoy the TDM exceptions under the CDSM Directive? In this context, the research examines the scope of the introduced provisions and their intended purpose. The overall aim is to provide an in-depth understanding of how these exceptions might function in practice and the effects they could have on various stakeholders.

Third, the dissertation examines how fundamental rights to information and the freedom of research could be affected by copyright and *sui generis* database right-based restrictions in the context of TDM. In Europe, the right to information originates from freedom of expression, which includes "freedom to hold opinions and to receive and impart information and ideas without interference by public authority and regardless of frontiers".[12] The right to information embraces a passive

8 Directive (EU) 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society [2001] *OJ L167/*10 (InfoSoc Directive), art.2. For more discussions on a broad definition of protected rights see, for example, Aikaterini Simopoulou, 'Text and Data Mining under EU Copyright Law' (International Hellenic University, Thessaloniki, 2020), 6; Thomas Margoni and Martin Kretschmer, 'The text and data exception in the proposal for a DSM: Why it is Not What EU Copyright Law Needs' (2018) CREATe Working paper, 4; 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 (8) GRUR International 685, 696.

9 Directive (EU) 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases [1996] *OJ L 77/20 (*Database Directive), art.7(1).

10 InfoSoc Directive (n 8).

11 Database Directive (n 9).

12 Art. 11 of the European Charter of Fundamental Rights (EUCFR) and art. 10 (1) of the European Convention of Human Rights (ECHR). In Europe, the right to information is also included in several national constitutions, such as art. 5 (1) of the German Basic law, art. 16(3) of the Federal Swiss Constitution and art. 11 of the French Declaration of Human Rights. At the international level, see art. 19 of the Universal Declaration of Human Rights (UDHR) and art. 19 (2) of the International Covenant on Civil and Political Rights (ICCPR) art. 19 (2) which states, "Everyone shall have the right to freedom of expression; this right shall include freedom to seek, receive and impart

freedom to receive information, which includes the right to access information and an active right to search for relevant information through existing materials.[13] In this sense, the freedom of research is generally considered to derive from the fundamental right to information.[14] However, freedom of research has also been explicitly recognised under Art. 13 of the Charter of Fundamental Rights of the European Union (EUCFR), which stipulates that "the arts and scientific research shall be free of constraint. Academic freedom shall be respected."[15] The study highlights that the conflict could be intensified by the fact that copyright, as part of other IP, is protected under property ownership in line with Art. 17(2) of the EUCFR[16] and Art. 1 of Protocol No. 1 of the ECHR.[17]

information and ideas of all kinds…". For more comments on art.11 of the EUCFR and art. 10 (1) see Steve Peers et al. (eds), *The EU Charter of Fundamental Rights: A Commentary* (Hart Publishing, 2021), 333-369.

13 The European Court of Human Rights (EctHR), *Youth Initiative for Human Rights v. Serbia*, no. 48135/06, 25 June 2013, paras 20 and 24, unreported. ECtHR, *Társaság a Szabadságjogokért v. Hungary*, no. 37374/05, 14 April 2009, unreported. *See also* ECtHR, *Kenedi v. Hungary*, no. 31475/05, 26 May 2009, unreported. In addition see ECtHR, *Timpul Info-Magazin and Anghel v. Moldova*, no. 42864/05, 27 November 2007, para. 31, unreported; and ECtHR, *WVgrzynowski and Smolczewski v. Poland*, no. 33846/07, 16 July 2013, para. 57, unreported. . In this sense, see also the Court of Justice of the European Union (CJEU) case *Stichting Greenpeace Nederland and Pesticide Action Network Europe (PAN Europe) v European Commission* (T-545/11) EU:T:2013:523. SSRN id3829 p. 7.

14 See, for instance, Christophe Geiger and Bernd Justin Jütte, 'Conceptualising a 'Right to Research' and Its Implications for Copyright Law: An International and European Perspective' (2022) Joint PIJIP/TLS Research Paper Series No.7-2022, 29–35; Steeve Peers et al. (n 12) 336, 405; Christophe Geiger and Bernd Justin Jütte, 'The Right to Research as Guarantor for Sustainability, Innovation and Justice in EU Copyright Law' (2022) in T. E. Pihlajarinne, J. Mähönen and P. Upreti (eds), *Intellectual Property Rights in the Post Pandemic World: An Integrated Framework of Sustainability, Innovation and Global Justice* (Cheltenham: Edward Elgar Publishing, forthcoming, 2023), 23–25. *EIT Nordisk Film & TV A/S v Denmark* (40485/02) 8 December 2005; *Dammann v Switzerland* (77551/01) 25 July 2006 at [52]. Opinion of AG Szpunar in *Pelham GmbH v Hutter* (C-476/17) EU:C: 2019:624 at [91].

15 Art. 13 of the EUCFR. See, for instance, Steeve Peers et al. (n 12) 349.

16 Art.17 (2) of the EUCFR states: "Intellectual property shall be protected."

17 Art. 1 of Protocol 1 of the ECHR: "Every natural or legal person is entitled to the peaceful enjoyment of his possessions. No one shall be deprived of his possessions except in the public interest and subject to the conditions provided for by law and by the general principles of international law." See also cases acknowledging the applicability of art 1 of Protocol 1 of the ECHR to intellectual property: ECtHR, *Melnychuk v. Ukraine* (dec.), no. 28743/03, 5 July 2005, ECHR 2005-IX. *See also* ECtHR, *Dima v. Romania*, no. 58472/00 (decision of 26 May 2005, unreported, and judgment of 16 November 2006, unreported). For earlier decisions by the European Commission of Human Rights, *see* ECommHR, *Sociéte nationale de programme*

Furthermore, the research analyses how the CDSM Directive's TDM exceptions have improved the situation in terms of fundamental rights and what limitations could dilute their efficiency. The analysis is conducted in the context of scientific research and the development of AI-related products.

Fourth, the dissertation aims to introduce the most fruitful means for the proper regulation of TDM in the EU. The research analyses legal and technical remedies that could address the current issues surrounding TDM, especially within the AI sector. A few questions arise: What can be done to ensure that the balance between competing interests is achieved more effectively than under the current law governing TDM? What solutions can bring a long-lasting impact during a time of rapid technological development? And finally, what are the future perspectives on the use of TDM within the realm of copyright law in light of recent technological developments?

The main research question, together with the relevant sub-questions, is addressed in the four publications. Each of the published articles has a specific context and objectives, and examines the key research questions from various perspectives.

The first research article, titled *"The Insufficiency of the EU's Text and Data Mining Exceptions for Using Artificial Intelligence"*, serves as an introduction to the topic. This paper examines how the EU legal frameworks on copyright, database, and computer programs may restrict the application of TDM. The article emphasises the need for a research-friendly and technology-neutral legal regime for TDM that could effectively address technological development and innovation within the DSM. It argues that the CDSM Directive's TDM exceptions appear insufficient to ensure a balance between competing rights and interests.

The TDM exception for scientific research under Art. 3 of the CDSM Directive is narrowly formulated and applies only to a few beneficiaries.[18] The second, broader TDM exception under Art. 4, which covers all types of beneficiaries and purposes, is diluted by the reservation right.[19] As the analysis unfolds, it becomes apparent that the national implementation of these provisions could provide an opportunity to create a more favourable framework for TDM in the EU. Since the article was written during the national transpositions of the CDSM Directive, it aimed to introduce meaningful improvements to the scope and applicability of TDM exceptions in a way that would secure a balance of power in the digital environment.

France 2 v. France, no. 30262/96, 15 January 1997, unreported; and ECommHR, *Oguz Aral, Galip Tekin and Inci Aral v. Turkey* (dec.).

18 Art.3 of the CDSM Directive.

19 Art. 4 of the CDSM Directive.

The article represents the starting point of my PhD journey, serving as an exploratory foundation for my future research. The subsequent publications provide a more rigorous and in-depth analysis of the issues at stake.

The second article, *"In Search of Balance: Text, Data Mining and Copyright in the Digital Single Market Directive from a Fundamental Rights Perspective",* discusses the legal framework of TDM in terms of fundamental rights and freedoms. The main objective of this article is twofold. First, the paper examines how the CDSM Directive's TDM exceptions improve the situation compared with the period before the CDSM Directive. The study indicates that the exceptions seek to balance copyright with other competing rights involved in TDM by providing greater accessibility to protected materials. It highlights that the mandatory provisions provide more legal certainty for researchers and businesses and a harmonised framework for TDM across the EU.

Second, the article evaluates the limitations of Arts. 3 and 4 and their impact on the fundamental rights to information and research. The study argues that, despite many positive aspects, these freedoms have been addressed only superficially. The article stresses that some limitations related to the scope and extent of the exceptions are likely to be disproportionate to the risks involved. Therefore, it proposes amendments to the legal framework of TDM so that fundamental rights can be better realised than under the current regulation. Although the study occasionally refers to US case law when discussing recent developments in TDM, its main focus remains on the EU legal framework for copyright and its implementation across EU Member States. The focused approach helps define more specific regulatory issues, conceptualise problems, and finally, offer practical solutions for setting common standards for TDM regulation in the EU.

The third research article, *"Overcoming Barriers to Text and Data Mining in the Era of ChatGPT: The Proposed Data Act as a Game-Changer"*, turns its focus to Art. 35 of the proposed Data Act,[20] which seeks to free databases comprising machine-generated data from the constraints of the *sui generis* database protection. The paper aims to evaluate whether this provision could offer a remedy for researchers facing challenges under the CDSM Directive's TDM exceptions. The article discusses real-life scenarios in which TDM researchers use ChatGPT to analyse large databases containing various types of medical data.

The findings indicate that researchers may still encounter legal uncertainty while relying on Art. 35 of the proposed Data Act, as many aspects require further clarification. In this regard, the article argues for refining the wording of this

[20] Proposal for a Regulation of the European Parliament and of the Council on harmonised rules on fair access to and use of data (Data Act), Brussels [2022] COM (2022) 68 final (proposed Data Act).

provision, considering the inclusion of derived data within the scope of the Data Act, and introducing new provisions that would explicitly address researchers' needs. Considering Art. 35 of the proposed Data Act as one of the alternative solutions, the paper demonstrates how strong copyright and database protection, combined with the limited exceptions, could impede data-driven research within the EU.

Finally, the fourth article, *"Can text and data mining exceptions and synthetic data training mitigate copyright-related concerns in generative AI?"*, discusses the growing copyright-related concerns in GenAI training and potential solutions. The rightsholders' concerns are twofold: (1) the use of protected works for mining without obtaining prior permission, and (2) the production of possibly derivative works based on the mining of protected content, which could replace those created by authors. The first point highlights the risk that authors may become less motivated to create new content due to reduced control over their works, while the second raises concerns about the future commercial viability of authors as creative workers.[21]

The study focuses on the first of these concerns, specifically discussing the unfair exploitation of protected works in the course of TDM, to examine the roots of the conflict. The paper addresses the most pressing issues from both legal and technical perspectives. In particular, the study evaluates the potential of the CDSM Directive's TDM exceptions, on the one hand, and synthetic data, on the other, to resolve the tension between rightholders and providers of GenAI systems. At the time of writing this article, there was no literature discussing the latter solution in the context of copyright and AI development, which added value to the current research. The findings indicate that the remedies in focus might mitigate the conflict to some extent, provided they are properly refined. In this regard, the paper emphasises the need to broaden the scope of the "commercial" TDM exception and to explore the synthetic data approach further to increase access to information while respecting rightholders' rights and interests.

Collectively, the articles provide a thoughtful analysis of the EU regulation of TDM that illustrates both its current state and potential improvements.

1.3 Methodology

The methodological choice in this research has been influenced by the article-based structure of this dissertation and the perspectives that it brings. The research is built on a combination of diverse legal methods and approaches. Methodological

[21] Maryna Manteghi, 'Can text and data mining exceptions and synthetic data training mitigate copyright-related concerns in generative AI?' (2024) 16 (2) Law, Innovation and Technology 664.

pluralism typically enriches the study, maximises the validity and reliability of the research, and strengthens the research findings.[22]

Against this background, ***doctrinal research*** constitutes the foundation of this dissertation. The doctrinal method enables a detailed analysis of legal norms and principles, including their interpretations, to identify and address uncertainties and gaps in the existing law.[23] The approach is thus analytical-descriptive.[24] It involves a two-step process: first, the identification of the sources of law, and second, the interpretation and analysis of relevant legal texts.[25] As such, the method is relevant and suitable for this research, which aims to examine existing legal norms within copyright and database law, to identify and address uncertainties and gaps in the current legal framework regulating the use of TDM in the EU.

The study, therefore, interprets legislation governing copyright and *sui generis* database protection, as well as legal acts, conventions, case law, and legal literature, analysing how these sources appear in the TDM context. The primary legislation for this research comprises the EU directives (the CDSM Directive, the InfoSoc Directive, the Database Directive, and the Software Directive[26]) and regulations concerning copyright and database law (the Artificial Intelligence Act (AIA)[27] and the Data Act). The EU Charter of Fundamental Rights and the European Convention on Human Rights also play a significant role in this dissertation, as many aspects are discussed from a fundamental rights perspective. As for case law, the relevant decisions of the CJEU and the ECtHR serve as valuable sources. Works that have provided important contributions to the research include those authored by T. Mylly,

22 Jörg Friedrichs and Friedrich Kratochwil, 'On Acting and Knowing: How Pragmatism Can Advance International Relations Research and Methodology' (2009) 63(4) International Organisation 701, 708-710.

23 Jan M. Smits, 'What is Legal Doctrine? On the Aims and Methods of Legal Dogmatic Research' (September 1, 2015). Rob van Gestel, Hans-W. Micklitz and Edward L. Rubin (eds.), Rethinking Legal Scholarship: A Transatlantic Dialogue, New York [Cambridge University Press] 2017, pp. 207-228, Maastricht European Private Law Institute Working Paper No. 2015/06, 5. Varga Csaba, 'Legal Dogmatic: Methodology and Ontology' (2011) 11-12 Law of Ukraine: Legal Journal, 82.

24 Jan M. Smits (n 23) 8.

25 Terry Hutchinson and Nigel Duncan, 'Defining and Describing What We Do: Doctrinal Legal Research' (2012) 17 Deakin Law Review, 83–119.

26 Directive (EU) 2009/24/EC of the European Parliament and of the Council of 23 April 2009 on the legal protection of computer programs (Codified version) (Text with EEA relevance) [2009] *OJ L 111/16*.

27 Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) [2024] OJ L 2024/1689, 12.7.2024.

C. Geiger, E. Izyumenko, M. Caspers, L. Guibault, E. Derclaye, R. Ducato, A. Strowel, J. Griffiths, B. Hugenholtz, and others.

However, the doctrinal method goes beyond a simple description and understanding of existing legal frameworks.[28] It helps to find practical solutions that align best with the current system and to show what the law ought to be.[29] Therefore, the analysis is complemented by a more prescriptive approach, reflecting on what various actors involved in text and data mining ought to do or abstain from doing.[30] More specifically, the study seeks to contribute to improving the current EU regulation of TDM, particularly by proposing ways to readjust it to fit the rapid development of AI and digital technologies.

Additionally, the dissertation examines alternatives that may alleviate the conflict between copyright and database owners and TDM users in legal (e.g., Art. 43 of the Data Act) and technological domains (synthetic data training). As *Hutchinson* rightly indicates, to propose what the law ought to be, it is first necessary to ascertain what the current law is.[31] Therefore, each article integrates analytical-descriptive and prescriptive legal research merged with a problem-based approach,[32] aiming to determine the relevant issues and evaluate existing solutions and their potential to fit the current system, as well as their possible refinements. Exceptions are at the core of TDM regulation. Thus, Arts. 3 and 4 of the CDSM Directive constitute the main focus of this dissertation. However, merely analysing existing norms and their interpretations, even in detail, is not enough to develop normative statements on how the law should be.

Therefore, the study relies on a combination of doctrinal legal research and ***constitutional analysis***. This implies the use of constitutional norms at the EU level to challenge prevailing interpretations and solutions in copyright and database law as they pertain to TDM. Significant technological changes often require updates to legislation due to the growing complexity of new interests and values.[33] However, since legislating is a long-term process, legislation often lags behind technological

28 Jan M. Smits (n 23) 10.

29 Jan M. Smits, *The mind and method of the legal academic* (Edward Elgar Publishing 2012), 44.

30 Jan M. Smits (n 23) 10; Jan M. Smits (2012) (n 29) 44.

31 Terry Hutchinson, 'Doctrinal Research: Researching the Jury in Research Methods in Law' in Mandy Burton and Dawn Watkins (eds.), *Research Methods in Law* (Routledge, 2017), 24, 28.

32 Terry Hutchinson (n 31)12-14.

33 Tuomas Mylly, 'The New Constitutional Architecture of Intellectual Property' in Jonathan Griffiths and Tuomas Mylly (eds), *Global Intellectual Property Protection and New Constitutionalism: Hedging Exclusive Rights* (Oxford University Press 2021), 52.

developments and may fail to function effectively.[34] Constitutional law may help establish a stable and solid legal framework amid technological evolution through enduring constitutional safeguards and principles.[35] The courts could use these fundamental principles to adjust lagging laws to rapid technological advances, dispensing with the need for constant legislative change.[36]

In the context of IP law, this implies that human rights may control the content of IP laws by limiting their "excesses", which is a prerequisite for balancing the interests of rightholders and users.[37] The dissertation specifically focuses on the concept of a fair balance, which could be employed to weigh the competing rights, including their inherent values and interests, against each other. Against this background, the study adopts a constitutional approach to examine how the fundamental rights to freedom of research and information are affected by copyright and *sui generis* database right-based restrictions. It also explores how the CDSM Directive's specific TDM exceptions have improved the situation in terms of European-level fundamental rights and what limitations could dilute their efficiency.

Furthermore, the study applies a ***law-in-context approach*** to certain aspects, such as the analysis of material consequences in the current technological environment. Therefore, the research studies TDM and AI to the extent required, also from a technological perspective, seeking to understand their economic and societal significance. The overall design of the study demonstrates an interest in the interactions within which the law exists in context, forms part of society, and changes in response to technological advancements and shifts in innovation.[38] To determine how the law should be amended, it is essential to understand the context in which it operates.[39] Therefore, discussing key aspects of the current TDM regulation (e.g.,

34 Rosa Hartmut, *Social Acceleration: A New Theory of Modernity* (Columbia University Press, 2013), 262-267.

35 Rosa Hartmut, 'Airports Built on Shifting Grounds? Social Acceleration and the Temporal Dimension of Law' in Luigi Corrias and Lyana Francor (eds), Temporal Boundaries of Law and Politics: Time Out of Joint (Routledge 2018) 73–87, 77; Tuomas Mylly (n 31) 54.

36 Tuomas Mylly (n 33) 55-57; Tuomas Mylly, 'Proportionality in the CJEU's Internet Copyright Case Law: Invasive or Resilient?' in Ulf Bernitz, Xavier Groussot, Jaan Paju, and Sybe de Vries (eds), *General Principles of EU Law and the EU Digital Order* (Wolters Kluwer 2020), 267–296.

37 Tuomas Mylly (n 33) 56.

38 Feenan Dermot, 'Exploring the 'Socio' of Socio-Legal Studies, in Dermot Feenan' in D Feenan (ed.), *Exploring the 'Socio' of Socio-Legal Studies* (Palgrave MacMillan 2013), 3-19.

39 Sara Blandy and Susan Bright, *Researching Property Law* (Bloomsbury Publishing 2015), 8; David N. Schiff, 'Socio-Legal Theory: Social Structure and Law' (1976) 39 (3) The Modern Law Review, 287.

the scope of exceptions, access limitations, etc.) through a sociological lens helps identify flaws or legislative errors and propose effective ways to amend it.

The study examines societal changes that could require the amendment or readjustment of the current TDM regulation to adapt to the evolving environment. It analyses the law by examining factual situations and practical legal problems to determine the rights and responsibilities of the actors involved.[40] The dissertation examines how the law affects different groups of stakeholders involved in TDM, from large digital giants and public research organisations to start-ups and researchers in private spheres. In particular, it focuses on AI development, including the rapidly emerging GenAI models, and the challenges that AI developers may face when relying on the current TDM regulation.

Accordingly, the research discusses the CDSM Directive's TDM exceptions and how effectively they address societal needs, such as innovation and technological development. Overall, the aim of applying the law-in-context approach is to identify various shortcomings of the current TDM regulation, such as limitations, legal uncertainty, and other flaws, that could impact society across various domains.

1.4 Earlier Research and Contribution of the Study

The topic of TDM and copyright has received considerable interest from academics. Many scholars, such as *Thomas Margoni* and *Martin Kretschmer,*[41] *Christophe Geiger et al*.,[42] *Ducato* and *Strowel,*[43] *Severine Dusollier,*[44] and *Triaille et al.,*[45] have argued for the need to properly regulate TDM in the EU. Their findings have been remarkably valuable for this dissertation. They rightly emphasised that the pre-existing laws and exceptions were inadequate for addressing the rapidly evolving digital environment and AI. Much of the existing literature has focused on the

40 Resa Banakar and Max Travers, *Theory and Method in Socio-Legal Research* (Hart Publishing 2005).

41 Thomas Margoni and Martin Kretschmer (n 8) 695, 705

42 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(7) IIC, 814, 824.

43 Rossana Ducato and Alain Strowel, 'Limitations to Text and data Mining and Consumer Empowerment: Making the Case for a Right to 'Machine Legibility'' (2019) 50 I.I.C. 649, 668.

44 Severine Dusollier, 'The 2019 Directive on Copyright in the Digital Single Market: Some Progress, a Few Bad Choices, and an Overall Failed Ambition' (2020) 57 (4) Common Market Law Review 979, 985.

45 Jean-Paul Triaille, Depreeuw, Dusollier, Hubin, de Francquen, *Study on the application of Directive 2001/29/EC on copyright and related rights in the information society* (European Commission, 150 et seq. 2013).

interpretation of the CDSM Directive's TDM exceptions from a doctrinal or policy-oriented perspective. In particular, *Margoni* and *Kretschmer*[46], *Geiger* and *Jütte*,[47] *Senftleben*[48] and *Ducato, Derclaye et al.,*[49] and *Strowel*[50] have acknowledged that the newly introduced regulation on TDM has poorly addressed the rights and interests of both public and private actors, prioritising those of rightholders instead.

These academic works have helped to contextualise the key issues and clarify the benefits and limitations of the exceptions provided under Arts. 3 and 4 of the CDSM Directive. Moreover, previous scholarship has played a key role in determining potential solutions for improving TDM regulation. For instance, *Hugenholtz* and *Senftleben,*[51]*Margoni,*[52] and *Reilly*[53] have focused on the adoption of an "open norm" system in the EU, similar to the US fair use doctrine, while *Geiger*[54] has advocated for the introduction of a statutory remuneration right. Overall, *Geiger's* earlier studies on TDM and copyright have been highly informative and thought-provoking for this research. He introduced a fundamental rights perspective to the topic and framed the right to TDM within the broader framework of the right to information and freedom of research.

The dissertation contributes to the existing literature by providing novel insights through a comprehensive analysis of TDM regulation. Applying a multi-methodological approach, the research examines copyright-related concerns in TDM from socio-legal, technological, and fundamental rights perspectives. Examining complex issues from multiple angles has enriched the academic discourse on the topic, offering more clarity regarding the balancing of competing interests at stake. Another distinctive feature of this research is the combination of theoretical analysis with practical application of the current TDM regulation in real-life scenarios (e.g.,

46 Thomas Margoni and Martin Kretschmer (n 8).
47 Christophe Geiger and Bernd Justin Jütte (7-2022) (n 14).
48 Martin Senftleben,'Generative AI and Author Remuneration' (2023) 54 IIC 1535.
49 Estelle Derclaye et al., 'Opinion of the European Copyright Society on selected aspects of the proposed Data Act' (European Copyright Society, 12 May 2022).
50 Rossana Ducato and Alain Strowel (n 43) 668.
51 Bernt Hugenholtz and Martin Senftleben, 'Fair use in Europe. In search of flexibility' (2011) Amsterdam Law School Research Paper, 8.
52 Thomas Margoni, 'To TDM or Not to TDM?' (OpenMinTeD, Brussels, 2018).
53 Susan Reilly (2014) *Libraries at the Centre of the Debate on Copyright and Text and Data Mining: the LIBER Experience*. In: IFLA WLIC 2014, 16-22 August 2014, Lyon, France, 4.
54 Christophe Geiger and Vincenzo Iaia,'The Forgotten Creator: Towards a Statutory Remuneration Right for Machine Learning of Generative AI' (2023) 52 Computer Law & Security Review (forthcoming, 2024); Christophe Geiger, 'Elaborating a Human Rights friendly Copyright Framework for Generative AI' (2024) Joint PIJIP/TLS Research Paper Series 123.

ChatGPT-based research on medical databases). This helps to fully address the challenges posed by the AI age and to propose effective solutions.

Unlike previous studies, this research focuses not only on how the TDM exceptions should be amended but also on discussing alternative solutions. Only a few academic works have analysed TDM regulation in the context of the Data Act and synthetic data training. Moreover, at the time of writing the early articles that constitute part of this dissertation, little scholarly attention had been given to the role of the AIA in addressing copyright-related concerns arising from the development and use of GenAI models.

Despite the growing body of research on TDM and copyright, this dissertation adds original insights and perspectives that are valuable for both academics and stakeholders involved in TDM. The research provides a comprehensive study of the EU regulation of TDM, successfully covering all relevant aspects and systematising them within a consolidated piece of work.

1.5 Structure of the Research

This is an article-based dissertation. The research consists of this introduction (summary section) and four articles, which have been published in peer-reviewed journals. The full-text publications are presented at the end of this summary. The summary is structured as follows. The introductory chapter (**Chapter 1**) provides an overview of the dissertation theme, identifying the main objectives and research questions, and explaining the choice of methodology and theoretical framework.

Then, **Chapter 2** introduces the research background and theoretical foundations that underpin this research in more substantive terms. The chapter begins with the societal embedding of the theme. First, it discusses the importance of TDM in different fields and points out developments such as AI that necessitate TDM regulation. The chapter highlights the need for a specific TDM exception arising from research and technological needs. It explains the reasons behind adopting new TDM exceptions under the CDSM Directive by focusing on the pre-existing laws and exceptions (both in the InfoSoc Directive and the Database Directive) and the problems they created.

Further, the chapter discusses in detail the scope and purpose of Arts. 3 and 4 of the CDSM Directive. It examines the advantages and limitations of these provisions. Furthermore, the chapter reviews the fundamental rights that may come into conflict in TDM contexts. In particular, the discussion revolves around the right to intellectual property (IP) on the one hand and the fundamental right to information and the freedom of research on the other. The intention is to analyse how these freedoms have been realised in the current regulation of TDM, particularly whether they have been properly taken into account in ensuring a fair balance of competing

interests. Finally, the chapter briefly summarises the main issues and findings related to the theoretical foundation and research background, setting the stage for the next chapter.

Chapter 3 introduces each of the articles constituting this dissertation, focusing on their research questions, purposes, main findings, and contributions to the overall research project. The chapter explains how the articles are connected to form a cohesive argument guided by the data underpinning the theoretical positioning of this thesis and the research questions. The chapter also discusses recent developments in the field and their relevance to my publications.

Chapter 4 summarises the overall findings of the research and their implications in relation to the main research question, outlines the key contributions of this work to the field, and provides directions for future research.

2 Research Background

2.1 Introduction

Text and data mining has emerged as one of the most powerful analytical tools, enabling the extraction of new, previously unnoticed patterns and knowledge from massive amounts of raw digital data stored around the world.[55] The use of this technique can revolutionise almost every domain of human life. TDM has enormous potential for both businesses and researchers, as it allows them to analyse large volumes of digital data to optimise and improve their performance. For instance, many businesses can employ TDM to analyse reviews on social media platforms to improve their products and services or to examine large volumes of financial data to identify trends or predict market behaviour.

Essentially, TDM contributes to advancing research by uncovering new patterns, insights, or correlations from large collections of scientific publications that are not immediately apparent through traditional manual analysis.[56] Such data analysis allows researchers to reveal, for instance, drug efficacy and interactions, vaccine candidates, correlations between diseases, cultural and historical aspects, and other important discoveries that help them to find answers to pressing societal problems.[57]

Furthermore, TDM is crucial for the development and functioning of different types of information-based services and AI applications.[58] Rapidly emerging GenAI models demonstrate the most recent AI developments that rely on TDM techniques. These systems are trained on huge amounts of existing data, and the insights and

55 Martin Senftleben, Thomas Margoni et al., '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 (1) JIPITEC 70.

56 For more examples of TDM use, see Sean Flynn and Lokesh Vyas, 'Examples of Text and Data Mining Research Using Copyrighted Materials' (Kluwer Copyright Blog, 6 March 2023).

57 Liana Colonna, 'Opportunities and Challenges to Utilizing Text-data Mining in Public Libraries: a Need for Legal Research' (2018) 50 years of Law and IT, 193.

58 Harry Surden, 'Machine Learning and Law: An Overview' in Roland Vogl (ed), *Research Handbook on Big Data Law* (Edward Elgar Publishing, 2021) 171.

correlations extracted are then used to generate new outputs.[59] Put simply, generative AI models such as ChatGPT, GitHub Copilot, DALL-E, Midjourney, and Stable Diffusion can perform various tasks, such as image or text generation, coding, sentiment analysis, and question answering, based on the analysed (mined) data.[60] Against this background, it is not surprising that TDM has been recognised as one of the significant advancements in the scientific and technological fields, and a driver of innovation across the globe.[61]

Although it offers numerous opportunities for researchers and commercial actors, copyright and sui generis database protection may hinder the application of this data-driven analytical technology. It is necessary to explain how TDM works to determine which exclusive rights might be affected during the data analysis process. TDM typically involves several steps, which may vary depending on the specific objectives of data analysis and the types of data involved. First, a computer collects data from various sources, such as websites, social media platforms, databases, and other digital repositories. Then, the obtained materials are usually pre-processed and transformed into a suitable format for analysis. The next stage is data analysis (or mining itself), which involves extracting new patterns or correlations through text summarisation, topic modelling, keyword extraction, clustering, or other tasks. Finally, the ultimate outcomes are generated. At this point, for instance, researchers may create a report or disseminate the results through publications or conference presentations.[62]

To perform text and data mining, a computer must make copies of all or part of the collected works, particularly at the stage of data analysis.[63] The processed works may be protected by copyright or database rights. Therefore, if the input data is mined without authorisation, the process may infringe copyright or database owners' exclusive rights. Concerning copyright, if a whole work or a substantial part of it is

59 Artha Dermawan,'Text and Data Mining Exceptions in the Development of Generative AI Models: What the EU Member States Could Learn from the Japanese 'Nonenjoyment' Purposes?' (2023) 27(1) The Journal of World Intellectual Property 44, 45; Van Lindberg,'Building and Using Generative Models Under US Copyright Law' (2023) 18(2) Rutgers Business Law Review 1, 1; Christopher Callison-Burch, 'Understanding Generative Artificial Intelligence and Its Relationship to Copyright' [2023].

60 Peter Henderson et al.,'Foundation Models and Fair Use' [2023] 1, 4. Avanika Narayan et al.,'Can Foundation Models Wrangle Your Data?' (2022) 16(4) PVLDB 738, 738. Ehsan Latif,'AGI: Artificial General Intelligence for Education' 2024, 1, 4.

61 CDSM Directive, recital 8.

62 For more discussions on the technical aspects of TDM see, for instance, Ken Yale, *Handbook of Statistical Analysis and Data Mining Applications* (Elsevier 2017); Han Jiawei, *Data mining: Concepts and Techniques* (3rd edition, Elsevier 2011).

63 Ibid.

reproduced by any means and in any form, there could be an infringement of the broad right to reproduction.[64] Further, if the data analysis is carried out on compilations of data, it might encroach upon the *sui generis* database right granted to the maker of a database who has made a substantial investment in either the obtaining, verification, or presentation of the contents of the database.[65]

The protection applies regardless of whether the database or its contents meet the originality requirements necessary for copyright protection.[66] The *sui generis* database owners may prohibit "extraction or re-utilization of the whole or of a substantial part, evaluated qualitatively or quantitatively, of the contents of that database".[67] Against this background, it could be presumed that TDM generally involves an act of copying (reproduction or extraction) of obtained data within the meaning of Art. 2 of the InfoSoc Directive and Art. 7 of the Database Directive.[68]

According to these regulations, the processing of protected works must always be lawful, meaning that users are obliged to obtain authorisation from rightholders or rely on relevant exceptions. However, in the case of TDM, obtaining licences for vast amounts of digital data, the ownership of which is usually difficult to identify, is a complex, time-consuming, and costly process.[69] This may place some actors, such as start-ups, small and medium-sized enterprises (SMEs), and modestly funded research institutions, in a less favourable position compared to financially robust actors (e.g., large commercial companies or Internet giants). Financially constrained actors, incapable of paying high licensing fees, would have to rely only on materials that are in the public domain or on data made freely available online through open-

64 InfoSoc Directive, art. 2 and recital 21; Michel Walter and Silke von Lewinski, *European Copyright Law: A Commentary* (Oxford University Press, 2010), 967-968. Jean-Paul Triaille et al., 'Study on the legal framework of text and data mining (TDM)' (2014), 30-31.

65 Database Directive, art. 7(1).

66 Database Directive, art. 7(4). See also Matthias Leistner and Lucie Antoine, 'Attention, Here Comes the EU Data Act! A Critical In-depth Analysis of the Commission's 2022 Proposal' (2022) 13 JIPITEC 339, 342.

67 Database Directive, art. 7(1).

68 Jean-Paul Triaille et al., 'Study on the legal framework of text and data mining (TDM)' (2014) 31; Rossana Ducato and Alain Strowel, 'Limitations to Text and data Mining and Consumer Empowerment: Making the Case for a Right to 'Machine Legibility'' (2019) 50 I.I.C. 649, 659.

69 Christopher Callison-Burch, 'Understanding Generative Artificial Intelligence and Its Relationship to Copyright' [2023] 13; Joao Pedro Quintais,'Generative AI, Copyright and the AI Act'(Kluwer Copyright Blog, 9 May 2023); Martin Senftleben,'Generative AI and Author Remuneration' (2023) 54 IIC 1535, 1544-1545.

source or Creative Commons licences. This would unfairly disadvantage research and innovation based on the market power of TDM users.[70]

As *Ducato* and *Strowel* rightly argue, if sources are limited in terms of quantity, scope, or quality, the outcome of TDM might be insecure and scarce.[71] Many scholars emphasise that such scenarios may result in, for instance, the development of inefficient or poor-quality AI systems.[72] Against this background, the need for appropriate TDM regulation, which should secure creativity while facilitating access to information and new technological developments, has become apparent in the age of AI.[73] The chapter proceeds by first providing the historical background to explain the need for specific TDM exceptions, then examining the current state of affairs, and finally proposing possible solutions.

2.2 The Need for Specific TDM Exceptions on the Theoretical Level

Were the EU TDM exceptions needed? This is a preliminary question that should be answered before delving into discussions on the effectiveness of the EU's current regulation of TDM. On the theoretical level, copyright protects original expressions, the form (container) of a creative work, whereas the information it contains, such as

70 Christophe Geiger et al., 'The Exception for Text and Data Mining' (2018) Centre for International Intellectual Property Studies Research Paper No.2018-02, 22.

71 Rossana Ducato and Alain Strowel (n 68) 668.

72 European Copyright Society, 'General Opinion on the EU Copyright Reform Package' (2017) 3; Carys J. Craig,'The AI-Copyright Challenge: Tech-Neutrality, Authorship, and the Public Interest' in Ryan Abbott (ed.), Research Handbook on Intellectual Property and Arti cial Intelligence (Edward Elgar Publishing 2022) 134, 150; Christopher Callison-Burch (n 50) 14; Apoorva Verma,'The Copyright Problem with Emerging Generative AI' [2023] SSRN, 5; Katherine Lee et al.,'Talkin' 'Bout AI Generation: Copyright and the Generative AI Supply Chain' Journal of the Copyright Society of the U.S.A. (forthcoming 2024) 4; Thomas Margoni and Martin Kretschmer (n 8) 687.

73 Christophe Geiger and Bernd Justin Jütte,'Digital Constitutionalism and Copyright Reform: Securing Access to through Fundamental Rights in the Online World' (Kluwer Copyright Blog, 25 January 2022). Even though the US has legalised the use of TDM as a matter of transformative use in *HathiTrust (2012)Authors Guild v HathiTrust* 902 F. Supp. 2d 445, 104 U.S.P.Q. 2d 1659 (2012) and *Google Book (2013 Authors Guild v Google* 05 Civ. 8136 (DC) (November 14, 2013) the topic gets a new breath of discussion with the emergence of GenAI models. See for instance *The New York Times Company v. Microsoft Corporation* (1:23-cv-11195), Daily News v Microsoft and OpenAI (consolidated with *The New York Times v OpenAI/Microsoft* and *CIR v OpenAI/Microsoft*), *The Intercept Media v OpenAI*, *Authors Guild v. OpenAI Inc.* (1:23-cv-08292).

ideas, facts, and data, is not protected.[74] These elements of copyrighted works are considered important to the public interest; therefore, access to them should be free from any constraints.[75]

In this way, copyright law seeks to facilitate the distribution of protected materials and to incentivise the creation of new, original works.[76] The principle is well known as the idea-expression dichotomy. The principle is acknowledged as one of the most important internal copyright mechanisms for balancing copyright with other competing rights.[77] The doctrine does not enjoy explicit general statutory recognition but is well established.[78]

At the EU level, the dichotomy is explicitly recognised in the Software Directive (Recital 11 and Arts. 1(2) and 5(3)), Database Directive (Recital 45), and the CDSM Directive (Recital 9), and it is restated in the case law of the CJEU[79]. Moreover, international treaties, such as the WIPO Copyright Treaty (Art. 2) and the WTO's TRIPs Agreement (Art. 9(2)), contain provisions acknowledging the general validity of the doctrine. The same is confirmed in the Berne Convention, which clarifies that only original expressions of human intellectual creations are protected and not the ideas embodied in those works.[80] Many authors emphasise the relevance of this concept to TDM.

74 Michael D. Birnhack, 'Acknowledging the Conflict between Copyright Law and Freedom of Expression under the Human Rights Act' (2003) 24 Entertainment Law Review 32; Thomas Margoni and Martin Kretschmer (n 8) 697; William Cornish, *Intellectual Property: Patents, Copyrights, Trademarks &Allied Rights* (London:Sweet & Maxwell, 1999), 382.

75 Loreto Corredoira and Rodrigo Cetina Presuel, 'Current Copyright Policy Tendencies in 2015: Further Weakening of Limits and Exceptions and the Ever-Reducing Public Domain' (2015), 3.

76 Loreto Corredoira and Rodrigo Cetina Presuel (n 75) 3.

77 Michael D. Birnhack (n 74) 9; Thomas Margoni and Martin Kretschmer (n 8) 696-697, 706; Sunimal Mendis, *Copyright, the Freedom of Expression and the Right to Information: The persisting Discord,* vol 8 (Nomos, Baden-Baden 2011) 22.

78 Michael D. Birnhack (n 74) 8.

79 See for instance Case C-833/18, of 11 June 2020, Brompton Bicycle , ECLI:EU:C:2020:461, at 27; Case C-683/17, of 12 September 2019, Cofemel, ECLI:EU:C:2019:721, at 29; Case C-393/09, BSA, of 22 December 2010, ECLI:EU:C:2010:816, at 49.

80 Berne Convention art. 2 (1), see also Sam Ricketson and Jane C. Ginsburg, *International Copyright and Neighbouring Rights*: The Berne Convention and Beyond (OUP 2022) 407, 406; Claude Masouyé and William Wallace, *Guide to the Berne convention for the protection of literary and artistic works (Paris act, 1971)* (WIPO publication No 615, 1978) 229; Mihaly Ficsor, *Guide to the Copyright and Related Rights Treaties Administered by WIPO and Glossary of Copyright and Related Rights Terms* (WIPO publication 891, 2003) 317, 23 – 24.

For instance, *Margoni* and *Kretschmer* argue that there should be no need for a copyright exception, as the tool extracts mere ideas and facts underlying protected materials, which are external to copyright protection.[81] As mentioned above, these elements can be freely circulated for the benefit of the public and should not be restricted. The purpose of TDM is not to infringe copyright by making unauthorised copies but to conduct data analysis in order to generate new, valuable information.[82] Many argue that if there is no real clash with copyright, there is no copyright infringement, and thus no safeguards for users' rights and interests are needed.[83]

However, some authors emphasise that the concept is not flawless in its endeavour to balance authors' interests with those of the public. For instance, *Birnhack* claims that the concept is rather vague.[84] *Mandic* argues that it is not clear how many ideas, when formed in an original combination, can amount to a substantial portion of a copyrighted work.[85] Practically, it may be difficult for courts to define the demarcation between ideas and their expressions, especially when they are intimately connected.[86] When the idea and expression overlap, the protected expression can become an obstacle to accessing non-protected (underlying) information due to restrictions arising from a broadly defined right to reproduction.[87] Furthermore, as *Birnhack* argues, in some cases, it may be necessary to borrow the copyrighted expression to use the "message" it conveys.[88]

In a time of rapid technological development, it might not be feasible in practice to access and extract the information embedded in a protected work without leaving

81 Thomas Margoni and Martin Kretschmer (n 8) 695, 705; Sean Flynn et al., 'Implementing User Rights for Research in the Field of Artificial Intelligence: A Call for International Action' (2020) 42(7) European Intellectual Property Review, 393-398; Matthew Sag, 'The New Legal Landscape for Text Mining and Machine Learning' (2019) 66 Journal of the Copyright Society of the USA, 3, 9-19; Carys J. Craig, 'Globalising User Rights-Talk: On Copyright Limits and Rhetorical Risks' (2017) 33(1) American University International Law Revew 1.) 4.

82 Jean-Paul Triaille et al., (n 68) 47; Carys J. Craig (2017) (n 81) 4.

83 Thomas Margoni and Martin Kretschmer (n 8) 705; Jean-Paul Triaille et al., (n 68) 47; Michael D. Birnhack (n 74) 26.

84 Michael D. Birnhack (n 74) 9, see also Yin Harn Lee et al., 'Copyright and Freedom of Expression: A Literature Review' (2015) (CREATe Working Paper 2015/04) 76-77.

85 Danilo Mandic, 'Balance: Resolving the conundrum between copyright and technology?' (University of Westminster, Working Paper, May 2011) 13.

86 Michael D. Birnhack (n 74) 9, 64; Jacob Zweig, 'Fair Use as Free Speech Fundamental: How Copyright Law Creates a Conflict between International Intellectual Property and Human Rights Treaties' (2013) 64 Hastings LJ 1549, 1553; Krisjanis Buss, 'Copyright and Free Speech: The Human Rights Perspective' (2015) 8(2) Baltic Journal of Law and Politics 182, 197; Yin Harn Lee et al., (n 84) 72.

87 Jean-Paul Triaille et al., (n 68) 109; Thomas Margoni and Martin Kretschmer (n 8) 690.

88 Michael D. Birnhack (n 74) 9, Yin Harn Lee et al., (n 84) 70.

the expression intact.[89] Reproductions may be needed to enable the development and functionality of information-based services and applications. *Ducato* and *Strowel* further elucidate that such reproductions may not hold intrinsic value but rather serve as a tool for enabling mere "reading" and analysis of protected content to extract underlying ideas and facts.[90]

In this regard, *Murray-Rust* argues that "the right to read is the right to mine,"[91] meaning that legitimate users should not be prevented from reading a work using their computers.[92] Similarly, *Reilly* states that there should be no unfair discrimination between human reading and reading by a machine.[93] However, many concede that since such reproduction is not fully transient or incidental, it may require explicit authorisation from copyright or *sui generis* database holders.[94]

Therefore, the complexities of the AI age have made it difficult to address copyright issues merely through the application of the idea-expression dichotomy. In this regard, many scholars suggest that when copyright protection gets closer to the embedded ideas and facts that underpin the original expressions, a compromise should be reached in the form of a specific exception.[95] Moreover, *Margoni* and *Kretschmer* emphasise that exceptions to copyright and the *sui generis* database right

89 Thomas Margoni and Martin Kretschmer (n 8) 690; Christophe Geiger, 'Author's Right, Copyright and the Public's Right to Information: A Complex Relationship (Rethinking Copyright in the Light of Fundamental Rights)'in Fiona Macmillan (ed.), *New Directions in Copyright Law* vol. 5 (Edward Elgar 2007) 24, 26; Sunimal Mendis (n 77) 23.

90 Rosana Ducato and Alain Strowel, 'Ensuring Text and Data Mining: Remaining Issues with the EU Copyright Exceptions and Possible Ways Out' (2021) 43 E.I.P.R. 20, 25.

91 Peter Murray-Rust, 'The Right to Read Is the Right to Mine' (Open Knowledge June 1 2012).

92 Rosana Ducato and Alain Strowel (2021) (n 90) 25; see also Thomas Margoni and Martin Kretschmer (n 8) 706; ECS (n 72) 5; Matthew Sag (n 81) 9-19; Carys J. Craig (n 81) 4; Severine Dusollier, 'The 2019 Directive on Copyright in the Digital Single Market: Some Progress, a Few Bad Choices, and an Overall Failed Ambition' (2020) 57 (4) Common Market Law Review 979, 985; See also recital 9 of the CDSM Directive.

93 Susan Reilly (2014) *Libraries at the centre of the debate on copyright and text and data mining: the LIBER experience*. Paper presented at: IFLA WLIC 2014 - Lyon - Libraries, Citizens, Societies: Confluence for Knowledge in Session 119 - Academic and Research Libraries with Serials and Other Continuing Resources and Committee on Copyright and other Legal Matters (CLM). In: IFLA WLIC 2014, 16-22 August 2014, Lyon, France, 4.

94 Marco Caspers and Lucie Guibault, 'A right to 'read' for machines: Assessing a black-box analysis exception for data mining' (2016) 53 (1) Proceedings of the Association for Information Science and Technology 1,1-2; Susan Reilly (n 93) 4, Sean Flynn et al., (n 81) 397, 399.

95 Pascale Chapdelaine, 'Copyright User Rights and Remedies: An Access to Justice Perspective' (2018) 7 LAWS 1-26, 13.

are needed to clarify the legality of technical processes (such as TDM) in principle.[96] The system of exceptions and limitations is an important balancing mechanism within copyright law, designed to mitigate tensions between rightsholders' and users' rights and interests. Such derogations normally permit certain uses of protected works without rightholders' authorisation, provided that these uses serve important public interests.[97]

Moreover, an exception may be needed when it is impractical or unreasonable to obtain such permission (e.g. temporary acts of reproduction, news reporting, etc.).[98] However, many claim that the pre-existing exceptions in the InfoSoc and Database Directives, which were designed during the initial phase of the digital age, were wholly insufficient to address the needs of the evolving digital environment.[99] They were restricted by numerous conditions and were not properly harmonised across the EU.[100] As *Triaille et al.* rightly emphasise, inadequate exceptions stand in the way of new data-driven applications and developments.[101]

One of the exceptions that could be relevant to TDM is the mandatory exception for temporary acts of reproduction provided under Art. 5(1) of the InfoSoc Directive.[102] To enjoy the exception, users of copyrighted works must comply with several conditions. The reproduction must be temporary, transient, or incidental and must constitute an integral and essential part of a technological process aimed at the lawful use of a protected work with no independent economic significance.[103] The reproductions made in the course of TDM may not qualify as "transient or incidental" since the copies are needed not only for data browsing but also for processing, uploading, transformation, and analysis of the obtained data.[104] In

96 Thomas Margoni and Martin Kretschmer (n 8) 708.

97 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 (7) IRIPCL 814.

98 Nicolas Jondet, 'The text and data mining exception in the proposal for a directive on copyright: why the European Union needs to go further than the laws of member states' (2018) 67 Propriétés Intellectuelles, p. 25-35, 30.

99 Severine Dusollier (n 73) 982; Jean-Paul Triaille, Depreeuw, Dusollier, Hubin, de Francquen, *Study on the application of Directive 2001/29/EC on copyright and related rights in the information society* (European Commission, 150 et seq. 2013); Bernt Hugenholtz et al., *The Recasting of Copyright & Related Rights for the Knowledge Economy* (Study for the European Commission 2006) 59–60.

100 Severine Dusollier (n 92) 982.

101 Jean-Paul Triaille et al., (n 99).

102 Article 5 (1) of the InfoSoc Directive.

103 Article 5 (1) of the InfoSoc Directive.

104 The CJEU clarified the scope of transient or incidental copies in *Infopaq International A/S v Danske Dagblades Forening* (C-5/08) EU:C:2009:465; [2009] E.C.D.R. 16 at [64].

addition, the copies are not automatically deleted after mining is finished and the outcome is generated.[105] Further, private actors, such as AI developers, may not meet the last condition of Art. 5(1) of the InfoSoc Directive, as their main purpose is to obtain financial benefits.[106]

The exception for scientific research provided under both the InfoSoc Directive and the Database Directive[107] represents another potential solution for proper TDM regulation. However, the exceptions are optional and have not been implemented uniformly across the EU, resulting in fragmented interpretations and applications of the provisions among Member States.[108] The provisions apply to the "sole purpose of illustration for teaching or scientific research" and for non-commercial purposes.[109] Moreover, they require users to indicate the source and the author's name of the reproduced work "unless this turns out to be impossible".[110] The latter notice is highly relevant to TDM, as the data analysis involves processing large volumes of data, which can be challenging, if not impossible, to attribute.

However, excluding private organisations from the scope of the provision could hinder developments in fields such as medicine or AI, where commercial actors play a significant role.[111] Moreover, the notion of "scientific research" may create legal uncertainty, as it has not been clearly defined by the Directives. According to Recital 36 of the Database Directive, the term should cover "both the natural sciences and the human sciences".[112] However, it would not be easy for TDM users to prove that their data analysis meets scientific criteria, taking into account different approaches, methods, and categories within the scientific disciplines.[113]

Another exception that could narrowly apply to TDM is the mandatory exception on the normal use of databases by "lawful users".[114] Users could enjoy

105 One of the requirements see *Public Relations Consultants Association Ltd v Newspaper Licensing Agency Ltd* EU: C:2014:1195 at [40]–[50].

106 Jean-Paul Triaille et al., (n 68) 47.

107 Article 5 (3)(a) of the InfoSoc Directive and Article 6(2)(b) and 9(b) of the Database Directive.

108 Severine Dusollier, 'The limitations and exceptions to copyright and related rights for libraries, research and teaching uses' (2013) in Study on the Application of Directive 2001/29/EC on Copyright and Related Rights in the Information Society (the InfoSoc Directive) (European Union 2013) 61.

109 Article 5 (3)(a) of the InfoSoc Directive and Article 6(2)(b), 9(b) and recital 36 of the Database Directive.

110 Ibid.

111 More discussion on the exceptions see e.g. Jean-Paul Triaille et al., (n 68) 50-71.

112 Recital 36 of the Database Directive.

113 See, for instance, the ERC panels' classification (2020), Anol Bhattacherjee, *Social Science Research: Principles, Methods, and Practices* (University of South Florida 2012).

114 Article 6(1) and 8(1) of the Database Directive.

the exception only if they have lawful access to and can use the content of a database based on the relevant exceptions or licensing.[115] Therefore, as *Caspers et al.* emphasise, in the absence of a legal exception, rightholders may, in principle, limit the application of TDM by not giving contractual permission to use the database.[116] Moreover, the provision only authorises reproduction to the extent necessary for "the purposes of access to the contents of the databases and normal use of the contents."[117] In this sense*, Triaille et al.* argue that the primary purpose of TDM, to extract new patterns, insights, or correlations from the contents of a database, would not normally fit within the purpose within the meaning of Article 6(1) of the Database Directive.[118]

Finally, the exception to the *sui generis* right provided under Art. 8 (1) of the Database Directive may also not be fully applicable to TDM.[119] The provision permits the extraction and/or re-utilisation of insubstantial parts (evaluated qualitatively and/or quantitatively) of the content of a database protected under the *sui generis* regime for any purpose.[120] The evaluation includes an assessment of the investment made by the maker of the database and the prejudice caused by these acts to that investment.[121] The exception provides significant possibilities for TDM. Moreover, the repeated and systematic extraction of insubstantial parts, which is typical in TDM, would still be lawful as long as it does not reconstitute the whole or substantial parts of the contents of a database, thus harming the investment made by the maker of a database.[122]

115 Jean-Paul Triaille et al., (n 68) 74-75. On the notion of lawful user se*e British Horseracing Board Ltd v William Hill Organization Ltd* (C-203/02) EU: C:2004:695; [2005] 1 C.M.L.R. 15 at [58] The CJEU clarifies that "lawful user" should be understood as covering users "whose access to the contents of a database for the purpose of consultation results from the direct or indirect consent of the right holder." Moreover, in *ACI Adam BV v Stichting de Thuiskopie* (C-435/12) EU: C:2014:254; [2014] E.C.D.R. 13 at [39] the CJEU clarified that lawful access to content is a necessary precondition for the normal exploitation of protected works.

116 Marco Caspers and Lucie Guibault et al, 'A Baseline report of policies and barriers of TDM in Europe' (2016), 33.

117 Article 6(1) of the Database Directive.

118 Jean-Paul Triaille et al., (n 68) 75–76; in this sense see also Christophe Geiger et al., (n 97) 824.

119 Art. 8 (1) of the Database Directive

120 Art. 8 (1) of the Database Directive

121 CJEU, 9 November 2004, case C-203/02, *The British Horseracing Board Ltd and Others v William Hill Organization Ltd*., par. 89.

122 CJEU, 9 November 2004, case C-203/02, *The British Horseracing Board Ltd and Others v William Hill Organization Ltd*., par. 89. Estelle Derclaye, *The Legal Protection of Databases: A Comparative Analysis* (Edward Elgar Publishing 2008) 111.

The purpose of TDM is not to reconstitute any part of a database but to obtain new knowledge from the content of the database. Therefore, users may potentially rely on Art. 8(1) of the Database Directive.[123] However, the application of this exception might be narrow, as it limits its beneficiaries to lawful users, implying that the use of TDM could be contractually overridden.[124]

To sum up, the pre-existing laws and exceptions discussed in this chapter were wholly insufficient to address the needs of rapidly evolving digital technologies and AI. Uncertainty in the regulation of TDM triggered the adoption of new mandatory, specific TDM exceptions under the CDSM Directive, intended to provide more legal certainty to both private and public actors in the digital environment. The next chapter will discuss these provisions in detail. It will consider how the TDM exceptions have improved the situation compared with that before the CDSM Directive and what aspects should be refined or readjusted in the age of AI.

2.3 TDM Exceptions under the CDSM Directive

With the adoption of the CDSM Directive, we have witnessed a significant transformation in the framework of copyright exceptions and limitations. The exhaustive list of exceptions and limitations to copyright and related rights has shifted from being narrow and mostly optional limitations on exclusive rights to becoming proper enabling mechanisms.[125] The CDSM Directive's copyright exceptions aim to demonstrate that an exception is not only a derogation from copyright but rather a self-contained rule that can facilitate publicly valuable uses of creative works (e.g. scientific research, teaching, access to information, etc.).[126]

Exceptions and limitations, if designed properly, can ensure a fair balance between the protection of creative works and their access and use, thereby aligning with fundamental constitutional values and principles.[127] In this sense, the CJEU has recognised that while copyright, as part of other IP rights, is protected under property ownership in the EUCFR, the right is not absolute and must be reconciled with other competing fundamental rights.[128]

123 Jean-Paul Triaille et al., (n 68) 79, Christophe Geiger et al., (n 97) 824.
124 Irinin Stamatoudi I, 'Text and Data Mining' In: Stamatoudi I (ed) *New Developments in EU and International Copyright Law* (Kluwer Law International 2016) 264–265; Christophe Geiger et al., (n 97) 824.
125 Severine Dusollier (n 92) 981.
126 Severine Dusollier (n 92) 982-983.
127 Christophe Geiger et al., (n 97) 908; see also Thomas Margoni and Martin Kretschmer (n 8) 696.
128 Ibid.

The concept of a fair balance derives from the principle of proportionality.[129] Proportionality is a general principle of EU law rooted in German constitutional law and is intended to balance conflicting interests.[130] The mechanism helps ensure that any limitations or restrictions on fundamental rights are reasonably justified and do not go beyond what is necessary to achieve the primary aim of the legislation.[131] The TDM exceptions under Arts. 3 and 4 of the CDSM Directive[132] seek to achieve that

129 Aharon Barak, *Proportionality: Constitutional Rights and their Limitations* (Cambridge: CUP 2012) 340, 343-344; Robert Alexy, 'Constitutional Rights, Balancing, and Rationality' (2003) 16 (2) Ratio Juris 131, 135; Jonas Christoffersen, 'Human Rights and Balancing: The Principle of Proportionality' in Christophe Geiger (ed), *Research Handbook on Human Rights and Intellectual Property* (Edward Elgar Publishing 2015) 19, 35; ECJ 22 January 2013, C-283/11, *Sky Österreich*, paras 50.

130 The proportionality test consists of three steps (sub-principles). First, the measure affecting the right must be suitable or appropriate, possessing at least the minimum degree of effectiveness, to reach the legitimate purpose (the principle of suitability). Second, the measure must be necessary, meaning that there is no other equally suitable measure available that is less onerous for achieving the pursued aim (the principle of necessity). Third, the benefits obtained from the restrictive measure must be proportionate to the harm caused to the restricted right (the principle of proportionality in the narrow sense). The latter sub-principle revolves around the concept of balancing, which, in turn, involves several stages. The process evaluates the degree of detriment to the first principle, the importance of satisfying the competing principle, and whether this importance could justify the harm to the first principle. The consideration is based on the "Law of Balancing", which implies that a conflict between competing values or principles can be resolved only through balancing. See more on the fair balance and proportionality Peter Teunissen, 'The Balance Puzzle. The ECJ's Method of Proportionality Review in EU Copyright Law' (2018) 5; Tuomas Mylly, 'Regulating With Rights Proportionality? Copyright, Fundamental Rights and Internet in the Case Law of the Court of Justice of the European Union' in Oreste Pollicino et al (eds), *Copyright and Fundamental Rights in the Digital Age : A Comparative Analysis in Search of a Common Constitutional Ground* (Edward Elgar Publishing 2020) 54-98; Jonas Christoffersen (n 129) 20; Christophe Geiger and Bernd Justin Jütte, 'Platform liability under Article 17 of the Copyright in the Digital Single Market Directive, Automated Filtering and Fundamental Rights: An Impossible Match' (2021) 70(6) GRUR International 517, 521-522; Robert Alexy (n 129) 135; Aharon Barak (n 129) 343; Tuomas Mylly (n 33) 55-56; See also ECtHR (5th section) *Fredrik Neij and Peter Sunde Kolmisoppi (The Pirate Bay) v Sweden App* no 40397/12 (ECtHR, 19 February 2013); *Delfi AS v Estonia*. ECJ 8 September 2016, C-160/15, *GS Media*, ECJ 1 December 2011, C-145/10, *Eva-Maria Painer*; ECJ 3 September 2014, C-201/13, *Deckmyn*; ECJ 21 October 2010, C-467/08, *Padawan*. ECJ 29 January 2008, C-275/06, *Promusicae*; ECJ 24 November 2011, C-70/11, *Scarlet Extended*; ECJ 16 February 2012 paras 45-46;C-360/10, *Netlog*; ECJ 27 March 2014, C-314/12, *UPC Telekabel*; ECJ 15 September 2016, C-484/14, *McFadden*.

131 Christophe Geiger and Bernd Justin Jütte (n 130) 521; Tuomas Mylly (n 130) 56; Art. 52 (1) of the EUCFR.

132 CDSM Directive arts. 3 and 4.

balance by providing improved access to copyright-protected materials while safeguarding authors' creativity and original expressions.[133]

The provisions were adopted as part of copyright reform to ease the control exercised by copyright and *sui generis* database holders over their works and to facilitate new developments and processes in the digital era.[134] This could, as *Pollicino* and others acknowledge, safeguard the public's general interests and foster the fundamental right to information and the freedom of research in the digital environment.[135] In this sense, *Geiger* and *Jütte* highlight that the permitted uses would create "breathing room" for AI developers, enabling significant developments within the European AI sector.[136] As *Geiger* and *Jütte* emphasise, the provisions clarify the lawfulness of TDM in principle.[137]

The unified and mandatory approach to TDM regulation can create a harmonised legal framework for TDM in Europe and help properly address challenges brought about by digitalisation.[138] Users would benefit from more legal certainty and clarity regarding the use of protected works for TDM-based data analysis.[139] This would facilitate, for instance, cross-border research collaboration and encourage companies using TDM tools to develop and establish their business models in the EU.[140] Overall, the exceptions have introduced a more robust regulatory framework for

133 CDSM Directive arts. 3 and 4 and recital 6; Christophe Geiger and Bernd Justin Jütte (n 73).

134 Christophe Geiger and Bernd Justin Jütte (n 14).

135 Oreste Pollicino et al.,'Cultural Rights, Cultural Diversity and the EU's Copyright Regime: the Battleeld of Exceptions and Limitations to Protected Content' in Oreste Pollicino (ed.), *Copyright and Fundamental Rights in the Digital Age* (Edward Elgar Publishing Limited, 2020) 124, 141–142.

136 Christophe Geiger and Bernd Justin Jütte,'Designing Digital Constitutionalism: Copyright Exceptions and Limitations as a Regulatory Framework for Media Freedom and the Right to Information Online' in Martin Senftleben et al. (eds), *Cambridge Handbook of Media Law and Policy in Europe* (Cambridge University Press, forthcoming), 4; see also Mauritz Kop,'The Right to Process Data for Machine Learning Purposes in the EU' (2021) 34 Harvard Journal of Law and Technology 1, 15

137 Christophe Geiger and Bernd Justin Jütte (n 14) 55.

138 Christophe Geiger et al., (n 70) 18-19.

139 Geiger, Christophe and Bulayenko, Oleksandr and Hassler, Théo and Izyumenko, Elena and Schönherr, Franciska and Seuba, Xavier, Reaction of CEIPI to The Resolution on the Implementation of Directive 2001/29/EC on the Harmonisation of Copyright in the Information Society Adopted by the European Parliament on the 9th July 2015 (2015) 37(11) European Intellectual Property Review 683, Centre for International Intellectual Property Studies (CEIPI) Research Paper No. 2015-01, 18.

140 More discussions see Christophe Geiger et al., (n 70); Christophe Geiger and Bernd Justin Jütte (n 14); Christophe Geiger and Bernd Justin Jütte (7-2022) (n 14) 54; Rosana Ducato and Alain Strowel (2021) (n 90).

TDM compared with the legal uncertainty that existed before the adoption of the CDSM Directive.

However, exceptions and limitations to copyright and related rights always limit and disable; they never just enable (everything) under the concept (for example, quotation, private use, parody, etc.). This applies to the CDSM Directive's TDM exceptions. Despite many benefits, the provisions come with certain flaws and limitations that could be problematic in terms of fundamental rights. In particular, many acknowledge that the introduced regulation on TDM has not fully realised the fundamental rights to information and the freedom of research.[141] Against this background, the following sections will discuss Arts. 3 and 4 of the CDSM Directive in the context of research and AI development, highlighting their benefits and limitations in terms of fundamental rights.

2.3.1 TDM Exception under Art. 3 of the CDSM Directive

Art. 3 of the CDSM Directive allows research organisations and cultural heritage institutions to reproduce and extract protected works to which they have lawful access for TDM, provided that the mining is carried out for the purposes of scientific research.[142] Research organisations are defined as "university, including its libraries, a research institute or any other entity, the primary goal of which is to conduct scientific research… on a not-for-profit basis or by reinvesting all the profits in its scientific research."[143] The term "cultural heritage institutions" refers to "a publicly accessible library or museum, an archive or a film or audio heritage institution" within the meaning of the CDSM Directive.[144] The exception in Art. 3 covers many positive aspects, along with certain limitations that may dilute its efficiency.

141 The Parliament's Constitutional Law Committee statement PeVL 58/2022 vp HE 43/2022 vp (14 November 2022). Thomas Margoni and Martin Kretschmer (n 8); Christophe Geiger and Bernd Justin Jütte (7-2022) (n 14); Christophe Geiger and Bernd Justin Jütte (n 136); Martin Senftleben (n 69)1535; Christophe Geiger,'Elaborating a Human Rights Friendly Copyright Framework for Generative AI' (2024) Joint PIJIP/TLS Research Paper Series 123, 22; On the technical aspects of machine learning see Josef Drexl, Reto Hilty et al.,'Technical Aspects of Artificial Intelligence: An Understanding from an Intellectual Property Law Perspective' (2019) Max Planck Institute for Innovation and Competition Research Paper No. 19–13, 28.

142 Art 3 of the CDSM Directive.

143 Art. 2 (1) of the CDSM Directive.

144 Art 2 (3) of the CDSM Directive.

No Contractual Override

According to Art. 7(1) of the CDSM Directive, the exception cannot be contractually overridden by rightholders.[145] Recital 17 further provides that the exception covers only entities carrying out scientific research; therefore, any potential harm caused to rightholders through the application of the exception would be minimal. For this reason, uses under the exception should not be subject to compensation for rightholders.[146] This could strengthen the position of the beneficiaries of Art. 3 by safeguarding their rights and interests from possible restrictive licensing terms imposed by rightsholders.

For instance, this may prevent publishers from using their contractual power to limit or restrict access to and the use of works required by researchers carrying out text and data mining.[147] Moreover, such a safeguard may provide more legal certainty and ensure that the exception is applied uniformly across EU Member States. This would facilitate cross-border research collaboration and ensure that new knowledge is disseminated freely across the EU.[148] The exchange of knowledge among research organisations is essential for advancing scientific research and addressing various societal concerns and challenges.[149]

The Beneficiaries of Art. 3

Art. 3 of the CDSM Directive has introduced more legal certainty for research organisations, which could open the door for new research projects and innovations. The provision does not require its beneficiaries to indicate the source and the author's name of the processed works, as was the case with the exceptions for scientific research in the InfoSoc Directive and the Database Directive.[150] The exception was needed, as it is practically difficult to attribute all materials used for TDM. Moreover, the final output of the data analysis typically does not include any excerpts from processed works but rather new insights and correlations. Thus,

145 Art. 7 (1) of the CDSM Directive.
146 Recital 17 of the CDSM Directive.
147 Jonathan Griffiths, Tatiana Synodinou, 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(1) GRUR International 22, 30.
148 See art.179(1) of the Consolidated Version of the Treaty on the Functioning of the European Union (TFEU) and Commission, 'Better Careers and More Mobility: A European Partnership for Researchers' COM(2008) 317 final; Christophe Geiger et al., (n 70) 20.
149 Christophe Geiger and Bernd Justin Jütte (7-2022) (n 14) 54.
150 Art.5(3)(a) of the InfoSoc Directive and arts. 6(2)(b) and 9(b) of the Database Directive.

listing protected works among thousands of sources may offer no real benefit to authors.

However, by limiting the scope of beneficiaries, the provision restricts access to information and hinders research in the private sector. The excluded actors, such as journalists, individual researchers, and businesses, would not be able to conduct data analysis based on TDM without prior authorisation. Obtaining licences for large volumes of mined data is a costly and complex process. This would place start-ups, SMEs, and similar commercial entities with limited financial or human resources in a less favourable position than corporate or internet giants. Limiting the scope of the exception in such a way would deprive the EU of essential research and innovation, as a considerable amount of public-interest research is carried out by private actors, especially in the field of medicine and AI.[151]

The limitation could be problematic in terms of the fundamental right to information and the freedom of research. The ECtHR clarified in *Tarsasag a Szabadsagjogokert v. Hungary* that the interference with the freedom of expression created by hindering access to information of public interest to certain users (e.g., journalists) cannot be justified and regarded as having been necessary in a democratic society.[152] Similarly, in *Kenedi v Hungary,* the Court emphasised that access to data for legitimate research purposes is an essential element of the freedom of expression.[153]

In the same vein, *Peers et al*. rightly argue that the fundamental rights to information and research should not be limited to particular types of data, research activity, or users.[154] Any person or entity should be permitted to analyse legitimately gathered data, particularly if the intention is to impart the underlying information to the public.[155] The permission is especially important when data analysis is conducted for the purpose of scientific research, as it could facilitate new developments in various fields and thereby contribute to overall public welfare. Against this background, extending the scope of Art. 3 of the CDSM Directive could be crucial for ensuring that users with legitimate access to protected data effectively exercise their fundamental rights to information and research.

151 Rossana Ducato and Alain Strowel (n 68) 651; Christophe Geiger and Bernd Justin Jütte (n 14) 13.

152 *Tarsasag a Szabadsagjogokert v Hungary* (37374/05) 14 April 2009 at [35]–[39].

153 *Kenedi v Hungary* (31475/05) 26 August 2009 para 43.

154 Steve Peers et al. (n 12) 344–345.

155 In this sense see and *Youth Initiative for Human Rights v Serbia* (48135/06) 25 September 2013.

The Purpose of "Scientific Research"

The purpose of Art. 3 is not limited to the "use for the sole purpose of illustration for teaching or scientific research"[156] but rather extends to scientific research *per se*. Many argue that "illustration" refers to teaching and not to scientific research.[157] For instance, *Stamatoudi* claims that scientific research requires a deeper examination and analysis than merely illustrating parts of (educational) works.[158] Moreover, *Triaille et al*. emphasise that TDM is used to advance scientific research, and that the whole work should typically be copied and used for data analysis.[159] They further argue that the expression "for the sole purpose…" can exclude research projects from the benefit of the exception if they incidentally carry other (supplementary) purposes such as statistical analysis or similar objectives.[160]

Although the CDSM Directive has broadened (to some extent) the purpose of the exception, it does not clarify the notion of "scientific research". Recital 12 only indicates that the term should be understood to cover "both the natural sciences and the human sciences".[161] The vague formulation can create legal uncertainty and lead to a narrow interpretation of the exception in Art. 3. As *Ducato* and *Strowel* rightly argue, there might be different nuances in the classification of science, which could also be subject to multiple interpretations.[162] The classification methods used to categorise the fields of science may cover other independent branches along with the natural sciences and humanities (e.g. engineering and technology, medical and health sciences, agricultural sciences)[163] or they may also recognise alternative groups of main branches (e.g. humanities and social sciences, physical sciences and engineering).[164]

156 InfoSoc Directive art. 5 (3) (a).

157 Jean-Paul Triaille et al., (n 68) 61; Irinin Stamatoudi (n 124) *273;* Christophe Geiger, Giancarlo Frosio and Oleksandr Bulayenko, 'The Exception for Text and Data Mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market - Legal Aspects' (2018) Centre for International Intellectual Property Studies (CEIPI) Research Paper No. 2018-02, 12; Severine Dusollier, 'The limitations and exceptions to copyright and related rights for libraries, research and teaching uses' (2013) *in Study on the Application of Directive 2001/29/EC on Copyright and Related Rights in the Information Society (the InfoSoc Directive)* (European Union 2013) 61*,* 378.

158 Irinin Stamatoudi (n 124) *273.*

159 Jean-Paul Triaille et al., (n 68) 61.

160 Jean-Paul Triaille et al., (n 68) 104.

161 Recital 12 of the CDSM Directive.

162 Rosana Ducato and Alain Strowel (2021) (n 90) 11.

163 See, OECD, Working Party of National Experts on Science and Technology Indicators, 'Revised field of science and technology (FOS) classification in the Frascati Manual' (2007).

164 See, for instance, the ERC panels' classification (2020).

Further, to qualify as "scientific," research should be conducted using established scientific methods and approaches. These instruments should meet various criteria (including replicability, precision, falsifiability, objectivity, reliability, parsimony, etc.) to be able to observe existing scientific data and test or cross-check the key research findings.[165] However, *Bhattacherjee* emphasises that some fields of study, such as arts, music, literature, and theology, cannot be tested using established scientific methods or techniques.[166] Against this background, there is no compelling reason to limit the application of the exception to certain scientific areas.[167]

Therefore, using the broader term "research" to define the purpose of Art. 3 could better address the research needs across various fields. The EU defines "research" as "creative work undertaken on a systematic basis in order to increase the stock of knowledge, including knowledge of man, culture, and society, and the use of this stock of knowledge to devise new applications".[168] Therefore, outcomes generated from such research would contribute to the advancement of science in either theoretical or practical ways and cover a broader scope of studies across various fields. This would be justified from the perspective of the fundamental freedom of research and would strengthen the EU's competitive position as a research area.[169]

Commercial and Non-Commercial Research Purposes

The exception in Art. 3 is not limited to non-commercial research purposes, which aligns with the purpose of the freedom of scientific research to protect all types of research activities.[170] Both commercial and non-commercial research may generate valuable outcomes that can provide significant benefits to the public. Moreover, research findings could be more accessible when commercialised.[171] For instance, in

165 Anol Bhattacherjee (n 113); Brian Kennett, *Planning and Managing Scientific Research: a Guide for the Beginning Researcher* (Canberra: ANU Press, 2014), 2.

166 For more discussions on what can be considered "science" see Anol Bhattacherjee (n 94).

167 Rosana Ducato and Alain Strowel (2021) (n 90) 11.

168 Directive (EU) 2016/801 of the European Parliament and of the Council of 11 May 2016 on the conditions of entry and residence of third-country nationals for the purposes of research, studies, training, voluntary service, pupil exchange schemes or educational projects and au pairing (recast) OJ L 132, 21/05/2016, p. 21–57 , art. 3 (9).

169 Recital 10 of the CDSM directive, Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions - Reviewing Community innovation policy in a changing world COM/2009/0442 final; Steeve Peers et al. (n 12) 420.

170 Steeve Peers et al. (n 12) 420–421.

171 LIBER, 'Busting TDM Myths and Misunderstandings' (2017).

public-private research collaborations, new patterns, correlations, or insights extracted from existing data can be transformed into marketable products, such as drugs, medical devices, software, apps, and other products.[172] The idea of innovation is to "take new ideas and translate them into commercial outcomes by using new processes, products or services in a way that is better and faster than the competition."[173] In this way, as *Bendis* and *Byler* emphasise, research would directly feed into social and economic development.[174]

However, many argue that the double limitation, introduced in Art. 3, is very close to the non-commercial requirement and could be even more restrictive in certain cases.[175] As *Margoni* and *Kretschmer* explain, the latter requirement would probably allow commercial entities carrying out scientific research for non-commercial purposes to benefit from the exception, a scenario that is not possible under Art. 3.[176]

The Public-Private Partnership

The public organisations falling within the scope of Art. 3 could also benefit from this exception when they carry out research in the framework of public-private partnerships, for instance, when using technological tools provided by their private partners for data analysis.[177] The CDSM Directive further clarifies that commercial actors should not have a decisive influence over the beneficiaries of Art. 3, which would grant them preferential access to research results.[178] Collaboration with the private sector offers researchers numerous benefits.

It can provide researchers with additional sources of research funding, supply them with complementary expertise or equipment, and help transform research into practical applications (research commercialisation).[179] In this sense, the European Commission emphasises the importance of research collaboration and the transfer of

172 Maryna Manteghi, 'Kneschke vs. LAION: Opening Fresh Perspectives on AI Training and TDM Exceptions' (*European Law Blog*, January 2025).
173 Richard Bendis and Ethan Byler, 'Creating a National Innovation Framework' (2009) Science Progress.
174 Richard Bendis and Ethan Byler (n 173).
175 Thomas Margoni and Martin Kretschmer (n 8) 705.
176 Thomas Margoni and Martin Kretschmer (n 8) 705.
177 Recital 11 of the CDSM Dircetive.
178 Recital 12 of the CDSM Directive.
179 Aurora Airaskorpi and Outi Vanharanta, 'The State of Industry-Academia Collaboration in the Social Sciences, Humanities and Arts: Perspectives from the Nordics and Beyond' Vaikuttavuussäätiö 2023; Sooho Lee and Barry Bozeman, 'The Impact of Research Collaboration on Scientific Productivity' (2005) *35* (5) Social Studies of Science 673–702.

knowledge between public and private sectors in addressing pressing societal challenges.[180] This also aligns well with the EU's goal of developing an integrated European Research Area with a more coherent research and innovation system.[181]

The Right to Store Copies

Art. 3(2) of the CDSM Directive allows the storage of copies made in the course of TDM "for the purposes of scientific research, including for the verification of research results".[182] Therefore, the provision does not limit the purpose of storage to current scientific activities but also allows researchers to improve the quality of their research by testing (verifying) hypotheses before knowledge production.[183] This would, as *Geiger* emphasises, greatly contribute to the realisation of the freedom of research.[184]

However, many aspects related to the status of copies of the protected works remain unclear. In particular, it remains unclear whether third parties (external users) can use the copies of mined works for the purpose of scientific peer review and joint research.[185] Recital 15 refers to the possibility of relying on the exception for scientific research provided under Art. 5(3)(a) of the InfoSoc Directive in such cases. However, these uses, which normally constitute an important part of scientific research, are not necessarily exempted under this optional and unspecified provision, as it is not fully adapted to the use of technologies in scientific research.[186]

180 European Commission, COM (2007) 182 - Improving knowledge transfer between research institutions and industry across Europe: embracing open innovation – Implementing the Lisbon agenda.

181 Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions - Reviewing Community innovation policy in a changing world COM/2009/0442 final 2 (2.3).

182 Art 3 (2) of the CDSM Directive.

183 Eleonora Rosati, *Copyright in the Digital Single Market: Article-by-Article Commentary to the Provisions of Directive 2019/790* (Oxford: Oxford University Press, 2021), 57. Reto Hilty and Heiko Richter, 'Position Statement of the Max Planck Institute for Innovation and Competition on the Proposed Modernisation of European Copyright Rules Part B Exceptions and Limitations (Art. 3 Text and Data Mining)' (2017) *Max Planck Institute for Innovation and Competition Research Paper* No.17-02, 11.

184 Christophe Geiger and Bernd Justin Jütte (n 14) 21.

185 Rosana Ducato and Alain Strowel (2021) (n 90) 12; Thomas Margoni and Martin Kretschmer (n 8) 697; Christophe Geiger, Giancarlo Frosio and Oleksandr Bulayenko (n 157) 23.

186 Rosana Ducato and Alain Strowel (2021) (n 90) 12; Recital 10 of the CDSM Directive.

Further, Art. 3 of the CDSM Directive requires that the copies of works made in the course of TDM be stored with "an appropriate level of security".[187] In addition, Recital 15 provides that Member States are entitled to decide on further specific arrangements for retaining the copies, including the appointment of trusted bodies that would be responsible for their storage.[188] Therefore, EU Member States may apply different storage conditions, creating legal uncertainty and thus hindering cross-border research activities.[189] Moreover, many argue that research organisations, such as universities and libraries, which allocate significant funds to subscriptions and retain copyright-protected works, should be responsible for the storage of copies rather than relying on "trusted intermediaries".[190]

Lawful Access

The exception in Art. 3 of the CDSM Directive allows the reproduction and extraction of protected works under certain conditions. However, it does not automatically grant access to such data. The provision requires that the "lawful access" to protected works pre-exists the TDM research.[191] The Directive defines the concept as "access to content based on an open access policy or through contractual arrangements between rightsholders and research organisations or cultural heritage institutions, such as subscriptions, or through other lawful means".[192] In addition, Recital 14 provides that the notion of lawful access should also cover access to content that is freely available online.[193]

The requirement of lawful access could limit the application of the exception in Art. 3 by subjecting TDM research to private ordering.[194] Put simply, rightholders may prevent or restrict the data analysis by refusing to grant licences to research organisations or raising the licensing or subscription fees.[195] The excessive related costs could make TDM research harder to perform, especially for research

187 Art 3 (2) of the CDSM Directive.
188 Recital 15 of the CDSM Directive.
189 REBIUN, 'Text and Data Mining: (Articles 3 and 4 of the EU-DSM) by REBIUN's Copyright Working Group' (*Ifla.org*,10 June 2020).
190 REBIUN (n 170); LACA, 'The Right to Read is the Right to Mine: But Not When Blocked by Technical Protection Measures' (1 August 2019).
191 Art 3 (1) of the CDSM Directive.
192 Recital 14 of the CDSM Directive.
193 Recital 14 of the CDSM Directive.
194 Christophe Geiger et al., (n 51) 29.
195 European Copyright Society (2017), ECS (n 72) 4, Christophe Geiger et al., (n 70), 22. Reto Hilty and Valentina Moscon, 'Modernisation of the EU Copyright Rules Position Statement of the Max Planck Institute for Innovation and Competition' (2017) Max Planck Institute for Innovation and Competition Research Paper No.17-12, 4.

institutions with limited financial resources.[196] These actors may not be able to obtain access to large volumes of digital data needed for mining, which could compromise the quality and value of TDM research.[197]

In this sense, *Ducato* and *Strowel* claim that if the sources are insufficient, the research outcome would be limited and unreliable.[198] Further, *Geiger* emphasises that the limitation could undermine the right to research by "discriminating research according to research organisations' market power".[199] This may exacerbate the divide between better- and less-funded research institutions and further reduce the research and innovative power of the latter actors.[200] The requirement of lawful access should not mean restricting TDM research but rather balancing competing interests by ensuring that authors receive fair compensation for the use of their works.

Therefore, the concept should be interpreted broadly and flexibly in light of the term "lawful use".[201] The CJEU clarified in *Infopaq International A/S* v. *Danske Dagblades Forening* that use should not be considered lawful only when it is authorised by rightholders through licensing, but also when it is not restricted by law.[202] In this sense, for instance, if protected works have been made freely available online without any TPMs, access to such works should be considered lawful.[203]

Technological Protection Measures

In light of a potentially high number of access requests to and downloads of their works, rightholders are permitted under Art. 3(3) of the CDSM Directive to apply measures to ensure the security and integrity of the systems or databases where such

196 Christophe Geiger et al., (n 70) 29.
197 Christophe Geiger et al., (n 70) 29.
198 Rossana Ducato and Alain Strowel (n 68) 668.
199 Christophe Geiger et al., (n 70) 22.
200 Christophe Geiger et al., (n 70) 29.
201 Jonathan Griffiths et al., (n 147) 24; Christophe Geiger, 'The Missing Goal-Scorers in the Artificial Intelligence Team: Of Big Data, the Fundamental Right to Research and the failed Text and Data Mining Limitations in the CSDM Directive' (2021) PIJIP/TLS Research Paper Series No 66, 7; Tatiana Eleni Synodinou, 'Lawfulness for Users in European Copyright Law: Acquis and Perspectives' (2019) 10 J.I.P.I.T.E.C. 20 para.1, 26.
202 *Infopaq International A/S* (C-302/10) EU:C:2012:16 at [42]. *Stichting Brein v Wullems (t/a Filmspeler)* (C-527/15) EU:C:2017:300; [2017] 3 C.M.L.R. 30 at [65]; *Football Association Premier League Ltd v QC Leisure* (C-403/08 and C-429/08) EU:C:2011:631; [2012] 1 C.M.L.R. 29 at [168].
203 *Infopaq International A/S* (C-302/10) EU:C:2012:16 at [44] and [45]; see AG Campos Sánchez-Bordona's Opinion in *Land Nordrhein-Westfalen v Renckhoff* (C-161/17) EU:C:2018:279. Ttiana Eleni Synodinou (201) 23.

works are hosted.[204] For instance, rightholders may apply TPMs to ensure that only users having lawful access to their materials can access them, including through IP address validation or user authentication.[205] Recital 16 of the CDSM Directive further clarifies that such measures should be proportionate to the risks involved, should not go beyond what is necessary to achieve their objective, and should not prevent the effective application of the exception.[206] The CDSM Directive indicates that rightholders should enjoy legal protection against the circumvention of any effective TPMs provided for in Art. 6 of the InfoSoc Directive.[207]

Even though the protection of TPMs remains essential for the effective protection and realisation of the rightholders' exclusive rights, rightholders should apply voluntary measures to ensure that the use of TPMs does not undermine the enjoyment of the exception.[208] While TPMs must be strictly targeted to effectively protect both copyright and the interests of lawful users, they must also be sufficiently effective to prevent unauthorised access to protected works or at least make it harder to achieve.[209]

In *Poland v. European Parliament,* the CJEU held that such measures must comply with the freedom of expression and information and secure the effective protection of the right to IP.[210] Otherwise, the rightholder's interference with the freedom of information would be unjustified in light of the objective pursued.[211] Following the Court's reasoning in *McFadden,* IP address validation or user authentication, examples provided by Recital 16 of the CDSM Directive, could be capable of striking a fair balance between competing rights at stake.[212]

However, automated technological measures cannot adequately distinguish between lawful (e.g. covered by relevant exceptions or limitations) and unlawful (infringing)

204 Recital 16 and art 3(3) of the CDSM Directive.
205 Recital 16 of the CDSM Directive.
206 Recital 16 of the CDSM Directive.
207 Art. 7 (2) of the CDSM Directive and art. 6 of the InfoSoc Directive.
208 Recital 7 of the CDSM Directive.
209 *Poland v European Parliament* (C-401/19) EU:C:2022:297; [2023] 4 W.L.R. 24 at [81], *UPC Telekabel Wien GmbH v Constantin Film Verleih GmbH* (C-314/12) EU:C:2014:192; [2014] Bus. L.R.541 at [55] and [56].
210 *Poland v European Parliament* (C-401/19) EU:C:2022:297; [2023] 4 W.L.R. 24 at [81]; In this sense see *UPC Telekabel Wien GmbH v Constantin Film Verleih GmbH* (C-314/12) EU:C:2014:192; [2014] Bus. L.R. 541 at [55] and [56]. *McFadden* (C-484/14) EU:C:2016:689 at [98]–[99].
211 *UPC Telekabel Wien GmbH* (C-314/12) EU:C:2014:192 at [56]; *McFadden v Sony Music Entertainment Germany GmbH* (C-484/14) EU:C:2016:689; [2017] 1 C.M.L.R. 38 at [93].
212 Recital 16 of the CDSM Directive; In this sense, see *McFadden* (C-484/14) EU:C:2016:689 at [98]–[99] 100.

access to protected data, which could lead to over-blocking.[213] As *Synodinou* argues, such measures would not pass the effectiveness test and would not justify the means required to achieve the intended aims.[214] In *Poland v. European Parliament,* the CJEU emphasised that blocking of lawful access would be incompatible with the rights to freedom of expression and information and would fail to strike a fair balance between these rights and the right to IP.[215] Therefore, the application of TPMs could lead to frequent lock-outs, which could prevent legitimate TDM users, under the exception of Art. 3, from accessing data for scientific research purposes.

Unreasonable delays in accessing automatically blocked content could create the risk that such data would lose its informative value and become outdated before TDM research takes place.[216] Moreover, even temporary unavailability of access could undermine the legitimate users' right to information and freedom of research.[217] In *Poland v. European Parliament,* the CJEU held that minimum procedural safeguards are needed, especially where interference with fundamental rights arises from automated processes, to ensure that such interference is strictly necessary.[218]

Clarifying the circumstances under which TPMs may be adopted and the conditions governing their application may reduce the risk of erroneous or unreasonable blocking of lawful access for legitimate TDM users.[219] Following the CJEU's reasoning in *Scarlet Extended SA* and *McFadden,* such safeguards would help strike a fair balance between the protection of copyright and the fundamental rights of those TDM users who are affected by such measures.[220]

213 *SABAM* (C-360/10) EU:C:2012:85 at [50] and [52]; *Scarlet Extended SA* (C-70/10) EU:C:2011:771 at [50]; *Ahmet Y ld r m v Turkey* (3111/10) 18 March 2013 at [66]–[68]; *Kharitonov v Russia* (10795/14) 16 November 2020 at [38]–[40]; *Poland v European Parliament* (C-401/19) EU:C:2022:297 at [86]. See Julia Reda, Joschka Selinger and Michael Servatius, 'Article 17 of the Directive on Copyright in the Digital Single Market: A Fundamental Rights Assessment' (2020) 26–27.

214 Tatiana Eleni Synodinou, 'Intermediaries' Liability for Copyright Infringement in the EU: Evolutions and Confusions' (2015) 31 Computer Law & Security Review 57, 62.

215 *Poland v European Parliament* (C-401/19) EU:C:2022:297 at [86].

216 *Poland v European Parliament* (C-401/19) EU:C:2022:297 at [60].

217 *Observer and Guardian v the United Kingdom* (13585/88) 26 November 1991 at [59(b)], 60

218 *Poland v European Parliament* (C-401/19) EU:C:2022:297 at [67]; see also *Data Protection Commissioner v Facebook Ireland Ltd* (C-311/18) EU:C:2020:559; [2021] 1 C.M.L.R. 14 at [176] *Kablis v Russia* (48310/16; 59663/17) 9 September 2019 at [92].

219 Ibid.

220 *Scarlet Extended SA* (C-70/10) EU:C:2011:771 at [45]; *McFadden* (C-484/14) EU:C:2016:689 at [100]. See also *UPC Telekabel Wien GmbH* (C-314/12) EU:C:2014:192 at [47].

2.3.2 TDM Exception under Art. 4 of the CDSM Directive

Art. 4 of the CDSM Directive introduces another mandatory TDM exception, which has also significantly improved the regulation of TDM compared to the situation that existed before the adoption of the CDSM Directive. The provision allows the reproduction and extraction of lawfully accessible works by any type of user and for any TDM-related purpose or activity.[221] Art. 4 provides that TDM can be carried out on protected works on the condition that the use of such works has not been expressly reserved by their rightholders in an appropriate manner.[222] The exception of Art. 4 has been adopted to recognise the significance of TDM techniques, whether performed by private or public entities, not only in the context of scientific research but also for other purposes across various fields.[223]

However, during the preparatory phase, the importance of TDM exceptions for the private sector (e.g. AI development) was mostly overlooked. The proposal for a Directive of the European Parliament and of the Council on copyright in the Digital Single Market[224] provided for a mandatory TDM exception covering only research organisations carrying out TDM for the purpose of scientific research.[225] Accordingly, private actors such as start-ups, SMEs, individual researchers, and journalists were excluded from the beneficiaries of this exception and still needed to obtain permission to carry out TDM for any purpose.

As *Geiger* rightly argues, there was a risk of formulating an "ineffective and therefore rapidly obsolete provision", which could impede AI development and hinder essential research activities and innovation in the public and private sectors.[226] Many criticised the restricted scope of the exception, particularly expressing concerns regarding the exclusion of private entities, which remained subject to exclusive rights.[227] As a result, a new so-called "commercial" TDM exception was introduced under the CDSM Directive, aimed at addressing the issues not covered

221 Art 4 (1) of the CDSM Directive.
222 Art. 4 (3) of the CDSM directive.
223 Recital 18 of the CDSM Directive, see also Thomas Margoni, 'Text and Data Mining in Intellectual Property Law: Towards an Autonomous Classification of Computational Legal Methods' (2020) CREATe working paper 01/2020; Christophe Geiger (n 182).
224 See the Proposal for a directive of the European Parliament and of the Council on copyright in the Digital Single Market, 14 September 2016, Brussels, COM(2016) 593 final, 2016/0280 (COD).
225 Art. 3, para. 1 of the Proposal (n 224).
226 Christophe Geiger (n 201) 7.
227 Christophe Geiger, Giancarlo Frosio and Oleksandr Bulayenko (n 157); for a critical evaluation of the directive proposal, see also Christophe Geiger, Giancarlo Frosio and Oleksandr Bulayenko (n 97) and from the same authors: 'The EU Commission's Proposal to Reform Copyright Limitations: A Good but Far Too Timid Step in the Right Direction' (2018) 40 (1) EIPR 4; ECS (n 72) 5.

by the TDM exception for scientific research. This major advance in the regulation of TDM could help resolve or mitigate copyright-related concerns, for instance, in rapidly emerging GenAI models.

Despite many positive aspects, the efficiency of Art. 4 may be diluted by certain limitations and specific conditions set in the CDSM Directive. While Art. 4(2) allows its beneficiaries to retain copies made in the course of TDM, the provision stipulates that such copies may only be stored "for as long as is necessary for the purposes of text and data mining".[228] Therefore, once the TDM process is completed, the copies should be deleted, unlike under Art. 3, which allows the storage of copies for verification purposes and potentially for future scientific research.

Further, Art. 4 does not specify the use of TPMs, however, as *Rosati* claims, such measures are also more broadly relevant to this provision unless rightholders have reserved the use of their works.[229] Recital 18 outlines the conditions under which the exception should take effect. The provision indicates that the exception can only be applied where the protected work is accessed lawfully by the beneficiaries, including materials that are freely available online, and provided that the rightholders have not reserved the use of their works in an appropriate manner.[230] The former requirement aligns well with the requirement of "lawful access" in Art. 3, which was critically discussed above. However, significant concerns have been raised regarding the latter condition, namely the so-called "opt-out" mechanism of Art. 4.

As indicated above, the condition allows rightholders to expressly reserve the use of their works for TDM in an "appropriate manner". Recital 18 of the CDSM Directive further clarifies that where a work has been made publicly available online, rightholders can reserve their rights through machine-readable means, including metadata and terms and conditions of a website or a service.[231] In other cases, rightholders can reserve their rights to make reproductions and extractions for TDM by other means, such as contractual agreements or a unilateral declaration.[232] The "opt-out" mechanism was intended to prevent the over-use of copyright-protected works and thus achieve a fair balance between rightholders' interests and those of TDM users.[233]

228 Art 4 (2) and recital 18 of the CDSM Directive.

229 Eleonora Rosati, *Copyright in the Digital Single Market: Article-by-Article Commentary to the Provisions of Directive 2019/790* (Oxford: Oxford University Press, 2021), 91; recital 16 of the CDSM Directive.

230 Recital 18 of the CDSM Directive.

231 Recital 18 of the CDSM Directive.

232 Recital 18 of the CDSM Directive.

233 Recital 6 of the CDSM Directive; Andres Guadamuz,'A Scanner Darkly: Copyright Liability and Exceptions in Articial Intelligence Inputs and Outputs' (2024) 73(2) GRUR International 111, 120.

However, in practice, it could further strengthen the position of copyright and *sui generis* database owners who are entitled to control TDM through contractual arrangements.[234] First, the exception of Art. 4 can be overridden by a contract under Art. 7(1) of the CDSM Directive, and second, the rightholders can control the use of their works in the course of TDM by relying on the lawful access requirement. In this context, if the rights are reserved, users would need to pay twice to carry out TDM ‑ first, to acquire lawful access to data, and then, to read and analyse such data using various automatic techniques.[235] This could allocate the research and innovative power into the hands of large private enterprises, leaving many small companies with limited budgets in a less favourable position.[236]

In this sense, *Hugenholtz* reasonably argues that the reservation right of Art. 4 could legitimate a derivative market in TDM, empowering rightholders to control, license or entirely prohibit text and data mining.[237] Further, it is not clear how an "opt-out" mechanism under Art. 4 would work in practice. In particular, more explanation is needed regarding the requirements that the use should be reserved "expressly" and "in an appropriate manner".[238]

Against this background, many argue that the limitations, including legal uncertainty, of Art. 4 of the CDSM Directive could have a chilling effect on the right to information and the freedom of research, as they are driven by copyright holders' commercial interests.[239] In this sense, some authors argue that it is necessary to adopt generally accepted protocols or standards that could clarify the work of an "opt-out" mechanism.[240] Arguably, rightholders should apply only effective technological

234 See Andrew Tyner,'The EU Copyright Directive:'Fit for the Digital Age' or Finishing It?' (2020) 26 J.I.P.L. 275, 281; Bernt Hugenholtz,'The New Copyright Directive: Text and Data Mining (Articles 3 and 4)' (*Kluwer Copyright Blog*, 24 July 2019).

235 Maryna Manteghi, 'Will the AI Act Bring More Clarity to the Regulation of Text and Data Mining in the EU?' (*EU Law Analysis*, 18 May 2024).

236 Gina Maria Ziaja,'The Text and Data Mining Opt-out in Article 4(3) CDSMD: Adequate Veto Right for Rightholders or a Su ocating Blanket for European Arti cial Intelligence Innovations?' (2024) 19(5) Journal of Intellectual Law and Practice 453, 454–453.

237 Bernt Hugenholtz (n 234).

238 Gina Maria Ziaja (n 236).

239 Christophe Geiger (n 201) 9; Christophe Geiger et al.,'Crafting a Text and Data Mining Exception for Machine Learning and Big Data in the Digital Single Market' in Xavier Seuba, Christophe Geiger and Julien Pénin (eds.), *Intellectual Property and Digital Trade in the Age of Arti cial Intelligence and Big Data* (A CEIPI-ICTSD, Issue 5, 2018) 95; Bernt Hugenholtz (n 234); Thomas Margoni and Martin Kretschmer (n 8) 685; Christophe Geiger and Bernd Justin Jütte (n 136) 13; Maryna Manteghi, 'Can text and data mining exceptions and synthetic data training mitigate copyright-related concerns in generative AI?' (2024), Law, Innovation and Technology, 16(2), 675.

240 Paul Keller,'Generative AI and Copyright: Convergence of Opt-Outs?' ((*Kluwer Copyright Blog*, 23 November 2023); Nicola Lucchi'ChatGPT: A Case Study on

measures, within the meaning of Art. 6(1) and (3) of the InfoSoc Directive, to reserve the use of their works. Failing that, it might be difficult for users to ascertain whether the use has been reserved or not.[241]

One option that rightholders might employ to reserve the use of their works for TDM could be the robots-exclusion protocol (a so-called robots.txt file).[242] The mechanism has been used by website owners for many years to prevent their content from being indexed by web robots or crawlers.[243] However, it would be less effective in the case of GenAI training. Rightholders might not be able to include detailed instructions on which materials may or may not be accessed or scraped by AI crawlers for training purposes. [244] This is because a robots.txt file is of limited size.[245] Moreover, the protocol is a voluntary measure, which does not protect access to and use of multiple works as such. Therefore, AI crawlers may just ignore the instructions or access protected works from other websites not employing a robots.txt file through embedded links.[246] Currently, many alternative solutions are available. For instance, websites such as Spawningai.txt[247] and Google-Extended[248] enable rightholders to control what content can or cannot be accessed and used for TDM, while the HaveIBeenTrained website allows copyright holders to search for their works in training datasets and opt them out.[249]

Copyright Challenges for Generative Arti cial Intelligence Systems' (2023) European Journal of Risk Regulation, 1,15; Paul Keller and Zuzanna Warso,'Dening Best Practices for Opting Out of ML Training' (*Open Future,* 28 September 2023); Péter Mezei,'A Saviour or a Dead End? Reservation of Rights in the Age of Generative AI' (2024) SSRN, 3.

241 *VG Bild-Kunst v Stiftung Preussischer Kulturbesitz* (C-392/19) EU:C:2021:181; [2021] Bus. L.R. 800 at [46]; In this sense see also *GS Media BV v Sanoma Media Netherlands BV* (C-160/15) EU:C:2016:644; [2017] 1 C.M.L.R.30 at [46].

242 OddyTech,'About /robots.txt'; In this sense see Martin Senftleben (n 69) 1544-1545; Chyan Yang and Hsien-Jyh Liao,'Using the Robots.txt and Robots Meta tags to implement online copyright and a related amendment' (2010) 28(1) Library Hi Tech 94, 97; Bernt Hugenholtz (n 234); Google Search Central,'Introduction to robots.txt'.

243 M Carl Drott,'Indexing aids at corporate websites: the use of robots.txt and META tags' (2002) 38 Information Processing & Management 209, 212.

244 Rossana Ducato and Alain Strowel (n 68) 675; Darren Ang,'The Web Scraper's World of Copyright Exceptions and Contractual Overrides' (2021) 13(22) Singapore Law Review 5.

245 Martijn Koster et al.,'RFC 9309 Robots Exclusion Protocol' (*RFC*, September 2022).

246 M Carl Drott (n 243) 212; Martijn Koster et al (n 245); Google Search Central (n 242); Darren Ang (n 244) 5; Cullen Miller,'Ai.txt: A New Way for Websites to Set Permissions for AI' (*Spawning Blog*, 30 May 2023);

247 See < https://site.spawning.ai/spawning-ai-txt > accessed 10 January 2025.

248 Danielle Romain,'An Update on Web Publisher Controls' (Google blog, 28 September 2023); Google Search Central,'Overview of Google Crawlers and Fetchers (User Agents)'.

249 See <https://haveibeentrained.com > accessed 10 January 2025.

Against this background, even though creating specific rules for the reservation of rights would provide more legal certainty for all stakeholders, it may not be able to change the rightholder-oriented approach of Art. 4 of the CDSM Directive. The purpose of TDM is not to create infringing copies of protected works but to generate new knowledge by analysing such materials.[250] In this sense, *Flynn* et al. rightly argue that the reproductions made in the course of TDM do not undermine the underlying rationale of copyright protection, which is to prevent unauthorised copying that could substitute for the author's work.[251] Therefore, a more radical solution is needed. Some scholars advocate for the adoption of an "open norm" system in the EU,[252] similar to the US fair use doctrine,[253] while others propose implementing a remuneration right for rightholders as an alternative to the reservation right.[254]

For instance, *Geiger* proposes introducing a statutory remunerated copyright limitation, which would impose a payment obligation on private actors for the use of protected works for TDM purposes.[255] *Geiger* analyses the conflict between providers of GenAI and creators of data used for training, claiming that the author's remuneration right could help strike a balance between the competing rights and interests.[256] He argues that the justifications for such remuneration can be found in Art. 5(2) (b) of the InfoSoc Directive, which provides a remunerated private copying

250 Mira T. Sundara Rajan,'Is Generative AI Fair Use of Copyright Works? NYT v. OpenAI' (*Kluwer Copyright Blog,* 29 February 2024).

251 Sean Flynn et al (n 81) 396; Christophe Geiger et al., (n 97) 817.

252 See for instance, Bernt Hugenholtz and Martin Senftleben 'Fair use in Europe. In search of flexibilities' (2011) Amsterdam Law School Research Paper , 8; Thomas Margoni, 'To TDM or not to TDM?' (OpenMinTeD, Brussels, 2018) 11; Susan Reilly (n 93) 4; Thomas Margoni and Giulia Dore, 'Why We Need a Text and Data Mining Exception (But it is Not Enough)' (2016) Working paper, 2016, 1-2.

253 The doctrine of fair use, is a doctrine in the law of the US that permits limited use of copyrighted material without having to first acquire permission from the copyright holder, codified in 17 U.S.C. § 107. Accordingly, in determining whether the use made of a work in any particular case is a fair use the factors to be considered shall include: (1) the purpose and character of the use, including whether it is commercial for non-profit educational purposes; (2) the nature of the copyrighted work itself; (3) whether the section used constitutes a substantial portion of the work as a whole; and (4) the effect of the use upon the potential market for, and value of, the copyrighted work.

254 Christophe Geiger (n 141); Martin Senftleben, (n 69); Christophe Geiger and Vincenzo Iaia,'The Forgotten Creator: Towards a Statutory Remuneration Right for Machine Learning of Generative AI' (2023) 52 Computer Law & Security Review (forthcoming, 2024); Christophe Geiger and Elena Izyumenko,'Towards a European'Fair Use' Grounded in Freedom of Expression' (2019) 35(1) American University International Law Review 1, 58.

255 Christophe Geiger and Vincenzo Iaia (n 254) 12.

256 Christophe Geiger and Vincenzo Iaia (n 254) 14.

exception,[257] and in the fundamental rights framework.[258] He emphasises that the functioning of the statutory remuneration instrument would ensure access to comprehensive data needed for TDM purposes while guaranteeing that authors receive fair remuneration for the commercial use of their intellectual creation protected under Art. 17(2) of the EUCFR.[259]

Indeed, the relationship between copyright holders and TDM users is complicated by the broad scope of copyright protection and the fact that copyright itself enjoys protection as part of other IP rights under the right to property.[260] However, as discussed above, the right to TDM may find a strong fundamental rights justification under the right to information and the freedom of research. Therefore, this research advocates for a revision of the exception in Art. 4 of the CDSM Directive to exclude the reservation of rights. This would increase access to information and research in the private sector and facilitate new developments and advances in various fields across the EU.[261] Rightholders would still receive remuneration for the use of their works and retain control under the lawful access requirement or through contractual agreements. Additionally, they would be able to apply TPMs to protect their content from mass invasion by web bots or crawlers.

The broader "commercial" TDM exception would encourage many private actors using advanced analytical techniques to set up and operate in the EU. This could increase the need for access to huge amounts of high-quality data, thereby motivating creators to produce new content. Overall, the refined TDM exception would align with fundamental rights and promote these values in the digital environment.[262]

2.4 TDM Exceptions in Light of the Three-Step Test

Broadening TDM exceptions is compatible with, and can be justified under, the three-step test, which constitutes an essential basis of copyright. The provision, which originates from Art. 9(2) of the Berne Convention, was subsequently integrated into multiple international instruments.[263] In EU copyright law, the three-

257 Art. 5(2) (b) of the InfoSoc Directive.
258 Christophe Geiger and Vincenzo Iaia (n 254) 13-14.
259 Christophe Geiger and Vincenzo Iaia (n 254) 14.
260 Dirk Voorhoof, 'Chapter 17: Freedom of expression and the right to information: Implications for copyright.' in C Geiger (ed), *Research Handbook on Human Rights and Intellectual Property* (Edward Elgar Publishing 2025) 340.
261 Sean Flynn et al (n 81) 403.
262 Thomas Margoni and Martin Kretschmer (n 8) 688, 695.
263 Arts. 5 and 11 of the WIPO Marrakesh Treaty to Facilitate Access to Published Works for Persons Who Are Blind, Visually Impaired or Otherwise Print Disabled (Marrakesh Treaty); Article 10(1) and (2) of the WIPO Copyright Treaty (WCT); Article 16(1) and (2) of the WIPO Performances and Phonograms Treaty (WPPT); Art. 13 of the Agreement on Trade-Related Aspects of Intellectual Property Rights (TRIPS).

step test was included in the InfoSoc Directive under Art. 5(5), which provides that copyright exceptions and limitations "shall only be applied in certain special cases which do not conflict with a normal exploitation of the work or other subject-matter and do not unreasonably prejudice the legitimate interests of the rightholder."[264]

The three-step test was designed to censor exceptions and limitations within IP law.[265] In this regard, the exceptions and limitations to copyright should not only be assessed through the lens of a broad interpretation of the reproduction right that seeks to ensure a high level of copyright protection, but must also comply with the three-step test, which may further narrow or limit their scope and application.[266]

The three-step test is also applicable to TDM exceptions under Art. 7(2) of the CDSM Directive. Recital 6 of the CDSM Directive clearly states that the test should be applied as a legal benchmark for evaluating the scope of exceptions and limitations to copyright, seeking to achieve a fair balance between competing interests.[267] Therefore, both the TDM exception for scientific research in Art. 3 and the so-called "commercial" TDM exception set out in Art. 4 of the CDSM Directive are subject to an additional assessment under the three-step test.[268] The test is cumulative in nature, meaning that each condition must be satisfied to ensure the acceptability of an exception or limitation.[269]

To examine compliance with the test, it should first be analysed whether the reproduction of protected works permitted under specific TDM exceptions constitutes a certain special case. In other words, the exceptions should be confined to certain special cases to align with the test. The term "certain" implies that exceptions or limitations must be clearly defined in national legislation, while there is no need to "identify explicitly each and every possible situation to which the

264 InfoSoc Directive art 5 (5).

265 Tuomas Mylly, 'Intellectual Property and European Economic Constitutional Law: The Trouble with Private Informational Power' (PhD thesis, University of Turku 2009) 199.

266 Ibid., p. 351; Daniel Mooney, 'Copyright in the Digital Age: Analysing the Achievements and Flaws in the EU Copyright Exceptions Domain' (2022) Cambridge J. L. Pol. & Art.133, 135.

267 Recital 6 of the CDSM Directive; Paulien Wymeersch, 'EU Copyright Exceptions and Limitations and the Three-Step Test: One Step Forward, Two Steps Back' (2023) 72 (7) GRUR International 631, 631.

268 Martin Senftleben, 'TDM, GenAI and the Copyright Three-Step Test' (2025) 2; Lucchi N 'Generative AI and Copyright – Training, Creation, Regulation' (2025) Study requested by the JURI Committee, Brussels: European Parliament, Policy Department for Justice, Civil Liberties and Institutional Affairs; Nicola Lucchi, 'Generative AI and Copyright – Training, Creation, Regulation,' (2025) Study requested bythe JURI Committee, Brussels: European Parliament, Policy Department for Justice, Civil Liberties and Institutional Affairs 34-35, 43.

269 Tuomas Mylly (n 265) 352.

exception could apply, provided that the scope of the exception was known and particularised."[270] Further, the term "special" connotes that exceptions or limitations should be "narrow in a quantitative as well as a qualitative sense".[271]

This interpretation suggests that the scope of beneficiaries must be sufficiently limited.[272] However, limitations or exceptions are not required to serve a specific purpose; instead, the focus should be on conceptual aspects, such as the categories of works affected by a copyright limitation and the circumstances under which it may be invoked.[273] Pursuant to these guidelines, it may be asserted that the TDM exception for scientific research provided under Art. 3 of the CDSM Directive satisfies the requirement of the first step, as it is narrowly drafted, applies to a limited set of beneficiaries such as research organisations and cultural heritage institutions, and clearly defines the purpose of the exception. Such conditions can meet the requirement of a sufficient degree of legal certainty and predictability.[274]

In contrast, many authors raise doubts as to whether the broader so-called "commercial" TDM exception under Art. 4 is fully compliant with the "certain special case" requirement because of its open-ended scope and dependence on rightholders' reservation.[275] However, it could be argued that TDM operates within a specific and clearly defined framework, using works as data input for computational (technical) analysis or processing rather than for creative expression. Moreover, the recognition of the temporary copying and private use exception illustrates that the use of works on a large scale is not inherently incompatible with the requirement of a certain special case.[276] Finally, the exception supports key societal interests and developments that could enhance the freedom of expression and the rights to information and research by, *inter alia,* increasing access to existing works.[277] Against this background, the TDM exception under Art. 4 of the CDSM Directive is likely to satisfy the first condition of the three-step test, particularly in relation to AI training, which may be regarded as sufficiently limited to fit "certain special cases."

270 See Report of the Panel, 15 June 2000, WTO Document WT/DS160/R, para. 6.108.
271 Ibid para. 6.109.
272 Report of the Panel (n 270) para. 6.127 and 6.143.
273 Report of the Panel (n 270) 6.109., 6.112; Christophe Geiger, Daniel Gervais and Martin Senftleben, 'The Three-Step Test Revisited: How to Use the Test's Flexibility in National Copyright Law' (2014) 29 American University International Law Review 581, 594.
274 Ibid.
275 Nicola Lucchi, 'Generative AI and Copyright: Training, Creation, Regulation' (Policy Department for Justice, Civil Liberties and Institutional Affairs, Directorate-General for Citizens' Rights, Justice and Institutional Affairs, European Parliament, PE 774.095, July 2025) 43.
276 Martin Senftleben (n 268) 5-7.
277 Ibid., p. 8.

The second step of the test implies that any limitations or exceptions must not conflict with the normal exploitation of the work. In this regard, the WTO Panel drew a distinction between the empirical and normative dimensions of the term "normal". Accordingly, the criterion of normal exploitation was interpreted as involving consideration of those forms of exploitation "that currently generate significant or tangible revenue", as well as those forms which, "with a certain degree of likelihood and plausibility, could acquire considerable economic or practical importance."[278] Therefore, an exception conflicts with normal exploitation if such uses enter into economic competition with the ways in which copyright owners normally extract value from their exclusive rights, thereby depriving them of significant or tangible commercial gains.[279] Attention was thus given to both existing and potential future markets in assessing a conflict with "a normal exploitation of the work."[280]

Considering the approaches outlined in the second step, discussions focus particularly on the TDM exception provided under Art. 4 of the CDSM Directive. Some authors argue that permitting AI training under this provision would deprive rightholders of a promising new source of income and cause a conflict with a core future aspect of "normal exploitation".[281] Their reasoning rests on the assumption that GenAI models can generate content capable of replacing human creative and original works, leading to a substantial risk that human authors might lose the incentive to create new works.[282]

However, many take the view that an "opt-out" mechanism averts a conflict with normal exploitation.[283] For instance, *Senftleben* asserts that the rights reservation option under Art. 4(3) of the CDSM Directive provides rightholders with the possibility to enter into licensing agreements with TDM users and thereby "restore" their exclusive rights. Accordingly, the "opt-out" mechanism should be regarded as a conflict-preventing measure.[284]

The broader "commercial" TDM exception proposed in this dissertation, which does not include a reservation right, is also unlikely to cause a conflict with the normal exploitation of the work. The TDM exception permitting AI training does

278 Report of the Panel (n 270) at para 6.180. see also Tuomas Mylly (n 265) 353.
279 Ibid at para 6.183.
280 Article 13 TRIPS; Article 5(5) of the InfoSoc Directive; Martin Senftleben (n 268) 10.
281 Nicola Lucchi (n 275) 43-44; Eleonora Rosati, 'No Step-Free Copyright Exceptions: The Role of the Three-Step in Defining Permitted Uses of Protected Content (Including TDM for Al-Training Purposes)' (2024) European Intellectual Property Review 46 (2024), 262, 271.
282 Ibid.
283 Martin Senftleben (n 48) 1544-1545; Martin Senftleben (n 268) 11-12.
284 Martin Senftleben (n 268) 11; Martin Senftleben (n 48) 1544_1545.

not deprive copyright holders of a significant source of income that plays a central role in the overall commercialisation of a work.[285] TDM licensing for AI training would merely complement the major profits and revenue derived from other modes of exploitation. Furthermore, the wording of the normal exploitation test focuses on the commercialisation of "the work." However, AI training requires the processing of vast amounts of data, and the role of any single work is insignificant both as training material for AI developers and as a source of income for authors.[286] Moreover, as *Samuelson* fairly argues, in the course of TDM, the work is used merely as a "data drop" within extensive training datasets.[287]

Finally, the third step of the test requires that limitations or exceptions should not unreasonably prejudice the legitimate interests of the author (Berne Convention and WCT)[288]or other rightholders (TRIPS)[289]. The third step constitutes a proportionality test involving a balancing exercise in which the legitimate interests of authors and rightholders are weighed against those of users.[290] The WTO Panel pointed out that the term "legitimate" covers both legal and broader normative perspectives.[291] Although the WTO Panel emphasised that legitimate interests are not limited to economic values, it chose not to discuss the concept of "legitimate interest" in greater detail.[292] In its analysis of the unreasonable-prejudice test, the Panel focused solely on "the economic value of the exclusive rights conferred by copyright on their holders."[293]

The key issue concerns the extent of prejudice that may reach unreasonable levels.[294] Accordingly, prejudice can be regarded as unreasonable if an exception to copyright results in, or is likely to result in, a disproportionate and unjustifiable loss of income for the rightholder.[295] Many authors emphasise that the reasonableness or justifiability element of the third step is decisive in balancing competing interests.[296] The TDM exception for scientific research can satisfy the requirements of the third

285 Martin Senftleben (n 268) 11
286 Ibid., p. 14.
287 Pamela Samuelson, 'Fair Use Defenses in Disruptive Technology Cases' (2024) 71 UCLA Law Review 1484, 1567.
288 Article 9(2) Berne Convention; Article 10 WCT.
289 Article 13 TRIPS.
290 Christophe Geiger et al (n 273) 595-596; Paulien Wymeersch (n 267) 634.
291 Report of the Panel (n 270) at para 6.224; Christophe Geiger et al (n 273) 596.
292 Report of the Panel (n 270) at para 6.226.
293 Report of the Panel (n 270) at para. 6.227.
294 Paulien Wymeersch (n 267) 635.
295 Report of the Panel (n 270) at paras 6.224-6.229; Tuomas Mylly (n 265) 353.
296 Daniel Gervais, 'Towards a New Core International Copyright Norm: the Reverse Three-Step Test' (2005) 9 Marq Intellectual Property L Rev 1,18-19; Paulien Wymeersch (n 267) 635.

step without remuneration. The conceptual contours of this provision suggest that the reproductions and extractions permitted under the exception do not result in harmful uses of protected works that would constitute unreasonable prejudice in the context of the three-step test.[297] The TDM exception of Art. 3 of the CDSM Directive pursues legitimate, reasonable, and well-founded policy objectives to increase access to information and to promote scientific research.

However, the balancing assessment regarding the TDM exception under Art. 4 of the CDSM Directive may require a stricter evaluation of potential market prejudice, as it applies to all types of users and purposes, including TDM for commercial uses.[298] In this regard, many argue that the potential prejudice to the legitimate interests of authors or other rightholders can be averted through a reservation right or an equitable remuneration scheme.[299] The former mechanism allows authors to reserve their rights for TDM, for instance, to exclude their works from AI training, and therefore, they cannot later claim harm or unreasonable prejudice. The latter mechanism likewise implies no unreasonable prejudice, as copyright holders who have licensed their works for TDM receive financial benefits.

Nevertheless, the broader "commercial" TDM exception, which does not include an "opt-out" mechanism, is unlikely to cause unreasonable prejudice to authors' legitimate interests. The aim of Art. 4 of the CDSM Directive is to pursue legitimate public goals, namely to promote innovation and technological development through increased access to information and knowledge by the public, which constitutes a manifestation of the fundamental rights to information and research. The test should not lead to a narrow interpretation of the provision that would extend copyright holders' legitimate interests in the sense of the third element of the three-step test to include unprotected ideas and data used in the course of TDM.

As *Mylly* fairly argues, the three-step test should be interpreted in light of applicable fundamental rights, which should be appropriately weighed in context.[300] It is also necessary to consider the values, interests, and public-policy reasons underlying copyright exceptions to avoid a merely formalistic evaluation. In this regard, the TDM exceptions should be assessed in light of the three-step test, in a manner that ensures the proper and effective realisation of the fundamental rights at stake.

To sum up, it may be argued that the current TDM framework, including the proposed broader exceptions, is likely to comply with the three-step test and can be justified on several grounds. First, these provisions cover TDM activities that involve

297 Martin Senftleben (n 268) 23.
298 Ibid., pp. 20, 23.
299 Paulien Wymeersch (n 267) 635; Martin Senftleben (n 268) 23-26.
300 Tuomas Mylly (n 265) 356.

merely the technical use of protected works. Second, they pursue social interests such as technological development and access to existing information. Third, the use of an individual work in the course of TDM does not cause any significant harm to the market. Finally, the lawful access requirement inherent in both TDM exceptions may ensure fair remuneration for authors and other rightholders.

2.5 The AIA and TDM

The original draft of the AIA did not cover copyright issues related to AI development.[301] The main purpose was to facilitate investment and innovation in AI while securing fundamental rights through the adoption of a risk-based approach without specifically addressing IP rights.[302] However, the increasing tension between copyright holders and providers of GenAI led the EU Parliament to include two provisions relevant to copyright in the final text of the AIA. Art. 3(63) of the AIA defines GenAI models as AI models that show significant generality, are capable of performing a wide range of different tasks, and can be integrated into diverse downstream systems or applications.[303]

The AIA imposes certain obligations on providers of GenAI models that are important in the context of TDM regulation. First, they should comply with "Union law on copyright and related rights…in particular to identify and comply with…a reservation of rights expressed pursuant to Article 4(3) of Directive (EU) 2019/790,"[304] and second, "draw up and make publicly available a sufficiently detailed summary about the content used for training of the general-purpose AI model…".[305]

The provisions were added to the final version of the AIA to address copyright-related concerns arising from the development and use of GenAI models. The former provision dispels all doubts regarding the relevance of the TDM exception in Art. 4 of the CDSM Directive to AI training, which relies heavily on TDM technologies.[306] The rationale for this, as many authors emphasise, could be derived from the broad definition of TDM in the CDSM Directive ("any automated analytical technique aimed at analysing text and data in digital form in order to generate information…")[307] and

301 Proposal for a Regulation of the European parliament and of the Council Laying Down Harmonized Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts COM/2021/206 final (AIA COM/2021/206 final).
302 Ibid., p. 1 (1.1).
303 Art.3 (63) of the AIA.
304 Art. 53 (1) (c) of the AIA.
305 Art. 53 (1) (d) of the AIA.
306 Christophe Geiger (n 141) 27-28; Péter Mezei (n 240) 8.
307 Article 2 (2) of the CDSM Directive.

the aim of Art. 4 of the CDSM Directive, which is to enable the use of TDM both by private and public actors for various purposes, including the development of new applications and technologies.[308] This aligns with the approach taken by the EC in its study on copyright and new technologies, which explicitly recognises the relevance of the TDM exception to the use of AI solutions in multiple domains.[309]

Further, the transparency obligation of the AIA, requiring providers of GenAI models to disclose data used for training their systems, can also provide more legal certainty in terms of AI training and copyright. [310] As *Geiger* and *Iaia* rightly argue, the transparency clause would enable authors and other rightholders to trace the use of their works for AI training and make a well-informed decision regarding the reservation of their rights.[311] Therefore, the transparency rule would enable a more effective practical exercise and control of an "opt-out" mechanism under Art. 4(3) of the CDSM Directive.[312]

Recital 107 of the AIA further clarifies that the providers of GenAI models should provide a comprehensive summary of the content used for training their algorithms, for instance, by "listing the main data collections or sets that went into training the model, such as large private or public databases or data archives, and by providing a narrative explanation about other data sources used."[313]

Therefore, the provision does not require providers of GenAI models to reveal an exhaustive list of all training data, which could make it less burdensome for AI developers to comply with the transparency obligation.[314] In this sense, *Quintais* argues that it could be problematic to itemise and attribute all copyrighted works used for AI training due to "the low threshold of originality, the territorial fragmentation of copyright and its ownership, the absence of a registration requirement for works, and in general the poor state of rights ownership

308 Recital 18 of the CDSM Directive. See e.g. Thomas Margoni and Martin Kretschmer (n 8); Rossana Ducato and Alain Strowel (n 68).

309 European Commission, 'Study on Copyright and new Technologies: Copyright data management and artificial intelligence' (February 2022), 196-228.

310 Article 53 (1) (d) of AIA and recital 107.

311 Christophe Geiger and Vincenzo Iaia (n 254) 8.

312 Communia,'Policy Paper #15 on Using Copyrighted Works for Teaching the Machine' (*Communia*, 26 April 2023); Authors and Performers Call for Safeguards Around Generative AI in the European AI Act (InitiativeUrheberrecht, 19 April 2023); Shabbir Merali and Ali Merali,'The Generative AI Revolution Opportunities, Shocks, and Risks' (2023), 46. For more discussions regarding the benefits of the transparency requirement see e.g., Leander Nielbock and Teresa Nobre,'The AI Act and the Quest for Transparency' (*Communia*, 28 June 2023); Thomas Margoni and Martin Kretschmer (n 8) 693-697; Martin Senftleben, Thomas Margoni et al.,' (n 55) 74, 82; Martin Senftleben (n 69) 12-19.

313 Recital 107 of the AIA.

314 Recital 107 of the AI Act; Gina Maria Ziaja (n 236) 458; Nicola Lucchi (n 240) 16.

metadata".[315] Moreover, *Lucchi* emphasises that the citation process for human writers differs from how AI trainers record and refer to mined works, thus it may not be easy to determine precisely what sources have been used.[316]

The copyright-related obligations of the AIA could bring more legal certainty regarding the applicability of the current TDM regulation to AI training and the transparency mechanism. However, the provisions might strengthen the rightholder-oriented approach of the "commercial" TDM exception, which could result in a massive exercise of the reservation of rights. Moreover, a strict transparency obligation rather than a "best efforts" standard might not achieve a reasonable level of technical feasibility.[317]

Therefore, there remain concerns and doubts regarding the practical functionality of the transparency clause and, more generally, the extent to which Art. 4 of the CDSM Directive can accommodate AI development.

2.6 EU Data Act and TDM

The new EU Regulation of the European Parliament and of the Council on harmonised rules on fair access to and use of data[318] (Data Act), as a key pillar of the European data strategy,[319] is relevant to TDM regulation. The Data Act entered into

315 Joao Pedro Quintais (n 69); see also Nicola Lucchi (n 240) 16; Christophe Geiger (n 141) 24-25.

316 Nicola Lucchi (n 240) 16; see also Jan Bernd Nordemann, Jonathan Pukas, 'Copyright exceptions for AI training data – will there be an international level playing field?' (2022) 17(12) Journal of Intellectual Property Law & Practice, 973-974.

317 Christophe Geiger and Vincenzo Iaia (n 254) 10.

318 Regulation (EU) 2023/2854 of the European Parliament and of the Council of 13 December 2023 on harmonised rules on fair access to and use of data and amending Regulation (EU) 2017/2394 and Directive (EU) 2020/1828 (Data Act) (Text with EEA relevance) PE/49/2023/REV/1 OJ L, 2023/2854, 22.12.2023.

319 Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions A European Strategy for Data, COM/2020/66 final. The EU strategy for data also covers other important elements such as Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on contestable and fair markets in the digital sector and amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act) (Text with EEA relevance) [2022] OJ L 265, 1-66. Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European data governance and amending Regulation (EU) 2018/1724 (Data Governance Act) (Text with EEA relevance) [2022] OJ L 152, 1-44. Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act) (Text with EEA relevance) [2022] OJ L 277, 1-102. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No

force on 11 January 2024 and will become applicable in September 2025. The Regulation aims to address challenges and unlock the potential of vast amounts of digital data held and controlled by a few actors to ensure fair access to and use of such data by consumers and businesses.[320]

The European Commission has acknowledged that access to data and the ability to use it are key drivers of economic and technological development in the rapidly evolving data-driven world.[321] The Data Act focuses on broadening access to and use of data generated by Internet-of-Things[322] (IoT) devices such as sensors and machines.[323] As *Kumar* et al. explain, IoT represents "an emerging paradigm that enables communication between electronic devices and sensors through the internet in order to facilitate our lives".[324]

In pursuit of its objectives, the Data Act includes a review of certain aspects of the Database Directive, focusing on the *sui generis* database right and its relationship with machine-generated data, without expressly amending the Directive.[325] Art. 43 of the Data Act (formerly Art. 35 of the proposed Data Act) provides that the *sui generis* protection should not apply "when data is obtained from or generated by a connected product or related service," to safeguard, in particular, the rights of users to access, use, and share such data.[326]

The Regulation defines a "connected product" as an "item that obtains, generates or collects data concerning its use or environment and that is able to communicate product data via an electronic communications service, physical connection or on-device access, and whose primary function is not the storing, processing or

300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance) PE/24/2024/REV/1 OJ L, 2024/1689, 12.7.2024.

320 European Commission, 'Data Act' (2024).

321 European Commission, 'Study to Support an Impact Assessment for the Review of the Database Directive' Final Report (2022), 1.

322 The Internet of Things is a term which refers to "connected products that obtain, generate or collect, by means of their components or operating systems, data concerning their performance, use or environment and that are able to communicate those data via an electronic communications service, a physical connection, or on-device access" (Recital 14 of the Data Act). Examples of the IoT include "smart home" devices and appliences, connected cars, smart TVs, trackers etc.

323 EC (n 321) 1; Giuseppe Colangelo, 'European Proposal for a Data Act-A First Assessment' (2022) CERRE Evaluation Paper 2022, 1, 6; Recital 14 of the Data Act.

324 Sachin Kumar, Prayag Tiwari and Mikhail Zymbler, 'Internet of Things is a Revolutionary Approach for Future Technology Enhancement: A Review' (2019) 6 Journal of Big Data 1, 1.

325 EC (n 321).

326 Art. 43 and Recital 112 of the Data Act. See also Arts. 4 and 5 of the Data Act specifying the users' rights intended for protection.

transmission of data on behalf of any party other than the user,"[327] and a "related service" as a "digital service, other than an electronic communications service, including software, which is connected with the product…in such a way that its absence would prevent the connected product from performing one or more of its functions."[328].

In the context of TDM, Art. 43 of the Data Act might provide a safe harbour for users who need to access and use databases made solely of machine-generated data (IoT data).[329] This would enable users to identify new insights and patterns, potentially facilitating greater discoveries in various fields. For instance, analysing (mining) databases consisting of data generated by wearable devices for healthcare monitoring (e.g., fitness trackers) could improve treatments, screening services, and overall patient well-being.[330]

The final text of the Data Act clarifies some controversial aspects found in the proposal for the Data Act.[331] In particular, it clarifies that the exclusion of databases made of machine-generated data from the *sui generis* database protection is absolute and not limited to cases of interference with users' rights under Arts. 4 and 5 of the Data Act.[332] Moreover, Recital 112 of the Data Act further specifies that the *sui generis* right should not apply to such databases in all cases. The Proposal provided that databases comprised of data obtained or generated by connected products should not be protected under Art. 7 of the Database Directive "where such databases do not qualify for the sui generis right," as "the requirements for protection would not be fulfilled."[333] The controversial wording raised concerns that databases of machine-generated data could still fall within the scope of this right under certain circumstances (e.g., the installation of sensors or expensive software).[334]

327 Art. 2(5) of the Data Act.
328 Art. 2(6) of the Data Act.
329 Arts. 4-5 and art. 43 of the Data Act.
330 Roberta De Michele and Marco Furini, 'IoT Healthcare: Benefits, Issues and Challenges' (GoodTechs '19: EAI International Conference on Smart Objects and Technologies for Social Good, Valencia, September 2019).
331 Proposal for a Regulation of the European Parliament and of the Council on harmonised rules on fair access to and use of data (Data Act), Brussels [2022] COM (2022) 68 final (proposed Data Act).
332 See art. 35 of the Proposal on Data Act (n 331).
333 Recital 84 of the Proposal on Data Act (n 331).
334 In this sense see eg, Estelle Derclaye et al., 'Opinion of the European Copyright Society on selected aspects of the proposed Data Act' (*European Copyright Society*, 12 May 2022) 5, 3; Inge Graef and Martin Husovec, 'Seven Things to Improve in the Data Act' 1, 4; Estelle Derclaye and Martin Husovec, 'Why the sui generis database clause in the Data Act is counter-productive and how to improve it?' 1, 2.

Although the refined final version of the Data Act introduces more legal certainty and clarity, various issues still need to be reconsidered.[335] First, the exclusion of derived or inferred data based on machine-generated data (e.g., statistical data) from the scope of the Regulation could place additional duties upon TDM users.[336] Particularly, they would need to determine which data is covered by the Regulation and which is not, a task which is not easy.[337] Once identified, they would be required to obtain authorisation from the *sui generis* database owners to lawfully analyse it.

Moreover, as *Noto La Diega* and *Derclaye* rightly argue, the exclusion is not well-justified.[338] Recital 15 indicates that to qualify as machine-generated, data should "represent the digitalisation of user actions and events" and be "valuable to the user and support innovation and the development of digital and other services protecting the environment, health and the circular economy."[339] The derived or inferred data may meet these requirements by providing valuable information needed, for example, for the integrity of research.

Second, excluding databases made of machine-generated data from the *sui generis* protection under Art. 43 of the Data Act does not automatically imply that TDM users can access and use such databases without any restrictions. The provision can be overridden by contractual agreements or by the application of TPMs (e.g., passwords, encryption, robots.txt file, etc).[340] The Regulation provides that such measures should not undermine a user's rights to use or access data; however, the provisions apply only to third parties' rights and data-sharing agreements.[341] Finally, the Data Act aims to facilitate access to machine-generated data by users, trade and business persons and, where there is an exceptional need to access such data, by public sector bodies.[342] Therefore, the Regulation is not focused on providing effective access to databases made of machine-generated data for research purposes.

As *Derclaye et al.* fairly highlight, the legislation lacks "robust access and use guarantees for scientific research,"[343] which are needed to ensure a research-friendly

335 Estelle Derclaye et al., (n 334).
336 Recital 15 of the Data Act.
337 Guido Noto La Diega and Estelle Derclaye, 'Opening Up Big Data for Sustainability: What Role for Database Rights in the Fourth Industrial Revolution?' in Ole-Andreas Rognstad, Taina Pihlajarinne and Jukka Mähönen (eds), *Promoting Sustainable Innovation and the Circular Economy: Legal and Economic Aspects* (Routledge 2022) 28.
338 Guido Noto La Diega and Estelle Derclaye (n 337) 28.
339 Recital 15 of the Data Act.
340 More discussions on this see Estelle Derclaye et al., (n 334) 5; Guido Noto La Diega and Estelle Derclaye (n 337) 28.
341 Data Act, arts 11(1) and 12(2). Recital 5 of the Data Act.
342 Data Act, art 1.
343 Estelle Derclaye et al., (n 334) 6.

regulatory framework at a time of growing data relevance. In this sense, researchers carrying out TDM should affiliate themselves with one of the above-mentioned categories of beneficiaries to enjoy access to data specified in the Regulation. They might also benefit from Art. 5 of the Data Act, which allows users to share machine-generated data with "third parties," which covers *inter alia* research organisations and not-for-profit actors.[344]

Moreover, researchers could potentially rely on the provisions requiring data holders, in situations of exceptional need, to make machine-generated data available to public sector bodies.[345] In this sense, Recital 63 of the Data Act clarifies that research organisations can also be organised as public sector bodies.[346] Further, researchers might be entitled to share data with "individuals or organisations in view of carrying out scientific research,"[347] provided that such actors "act either on a not-for-profit basis or in the context of a public-interest mission recognised by the State"[348]. In this context, for example, private research institutions would not be able to access databases made of machine-generated data, even when collaborating with public entities.[349]

Overall, the scope of the discussed provisions of the Data Act is narrow and applicable only in cases of exceptional need and under certain circumstances. Therefore, they provide limited benefits to researchers carrying out TDM for the purpose of scientific research. Against this background, the Data Act may not be the best solution for users seeking to lawfully conduct TDM on databases.

2.7 Synthetic Data

Synthetic data is a potential technical solution that could mitigate the tension between copyright and *sui generis* database holders on the one hand and TDM users on the other. Generally, synthetic data refers to "artificially annotated information generated by computer algorithms or simulations."[350] Simply put, synthetic data is so-called "fake" data generated by a computer that mimics real-world data. It can be generated by extracting the underlying features of actual data (e.g., interactions, patterns, or correlations) or without relying on original data sources by means of statistical models, simulations, the background knowledge of a data analyst, or other

344 Art. 5 and recital 33 of the Data Act.
345 Article 14 of the Data Act.
346 Recital 63 of the Data Act.
347 Article 21 (1) (a)) of the Data Act.
348 Article 21 (2) and recital 76 of the Data Act.
349 Estelle Derclaye et al., (n 334) 6.
350 Yingzhou Lu et al.,'Machine Learning for Synthetic Data Generation: A Review' (2021) 14(8) Journal of Latex Class Files 1, 1.

relevant computational techniques.[351] Access to vast amounts of diverse and high-quality data is a prerequisite for producing accurate and reliable outputs in the course of TDM.[352]

Synthetic data can be used to augment scarce or depleted reserves of data when, for instance, research in a chosen area is insufficient or when access to data is restricted due to various legal grounds (e.g. privacy and data protection, copyright).[353] In this sense, *Ducato* and *Strowel* claim that if TDM users have to rely only on data that is in the public domain or freely available online, the final output would be scarce and insecure.[354]

Moreover, many IT experts emphasise that generating synthetic data for TDM purposes requires less effort and fewer financial resources compared to obtaining real-world data.[355] Synthetic data, which is generated from actual data, could be of higher quality and more versatile, as it maintains the key features of real data and is validated or verified using it.[356] In theory, the use of synthetic data for TDM should not raise any concerns regarding copyright protection, as there is no direct reliance on actual data.[357] However, the process of data synthesis *per se* requires the use of

351 Khaled El Emam,'Accelerating AI with Synthetic Data Generating Data for AI Projects' (*NVIDIA, O'Reilly Media* 2020), 3; Aryan Jadon and Shashank Kumar,'Leveraging Generative AI Models for Synthetic Data Generation in Healthcare: Balancing Research and Privacy' (2023), 2; Tshilidzi Marwala et al.,'The Use of Synthetic Data to Train AI Models: Opportunities and Risks for Sustainable Development' (2023), 3-4; Joao Fonseca and Fernando Bacao,'Tabular and latent space synthetic data generation: a literature review' (2023) 10 (115) Journal of Big Data 1.

352 Yingzhou Lu et al (n 350) 1; Aryan Jadon and Shashank Kumar (n 351) 2-3.

353 Aryan Jadon and Shashank Kumar (n 351) 2; Abdul Majeed,'Attribute-Centric and Synthetic Data Based Privacy Preserving Methods: A Systematic Review' (2023) 3 Journal of Cybersecurity and Privacy 638, 639; Erroll Wood et al.,'Fake It till You Make It: Face Analysis in the Wild Using Synthetic Data Alone' (2021), 3682; Sina Alemohammad et al.,'Self-Consuming Generative Models Go MAD' (2023), 2; Joao Fonseca and Fernando Bacao (n 351) 6-7; Yingzhou Lu et al. (n 350) 1; Tshilidzi Marwala et al (n 351) 4; James Jordon et al.,'Synthetic Data– What, Why and How?' (2022), 6, 30;

354 Rossana Ducato and Alain M. Strowel (n 43) 668; Tshilidzi Marwala et al (n 351) 3.

355 See n 351.

356 Aryan Jadon and Shashank Kumar (n 351) 2-3; Khaled El Emam (n 351) 2; Tshilidzi Marwala et al (n 351) 3-4; Joao Fonseca and Fernando Bacao (n 351) 1; Mostly AI,'What is Synthetic Data?'.

357 Gareth Kristensen et al.,'Training AI models on Synthetic Data: No silver bullet for IP infringement risk in the context of training AI systems (Part 1 of 4)' (*Clearly IP and Technology Insights*, 12 December 2023).

advanced analytical techniques, such as TDM, to analyse existing information and learn the underlying patterns, insights, and correlations.[358]

Therefore, if the intention is not to make infringing copies of protected works that might substitute for the author's original expression but to create synthetic data samples that only imitate protected materials, then no copyright infringement would occur.[359] However, if AI-synthesised data closely resembles or includes excerpts of protected works, both its generation and subsequent use for mining could constitute copyright infringement.[360]

In this sense, for instance, AI developers would need to obtain authorisation from rightholders or comply with the requirements of Art. 4 of the CDSM Directive ("lawful access" and the reservation of rights) to use synthetic data for training their systems. However, once allowed, they could create more variations of AI-synthesised content, which might be used to train an unlimited number of AI models.[361] Using synthetic data derived from other synthetic data could save money, time, and effort for AI developers compared to obtaining lawful access to real data. Any subsequent generation of synthetic data would be less connected to actual data, which would eliminate the risk of copyright infringement.[362]

358 Peter Lee, 'Synthetic Data and the Future of AI' (2024) 110 Cornell Law Review (forthcoming) 15; Aryan Jadon and Shashank Kumar (n 351) 2. In this sense see: Sean Flynn et al (n 81) 4; Christophe Geiger et al (n 97) 817; Mira T. Sundara Rajan (n 250); Giulio Coraggio,'How synthetic data can address IP and privacy issues of arti cial intelligence' (*GamingTechLaw*, 4 April 2023).

359 In this sense, see Matthew Sag,'Fairness and Fair Use in Generative AI' (2024) 92 (5) Fordham Law Review 1887; See also famous US cases which hold that TDM constitute fair use: *Authors Guild v. HathiTrust*, 755 F.3d 87 (2d Cir. 2014) and *Authors Guild v. Google*, Inc., 804 F.3d 202 (2d Cir. 2015).

360 Peter Lee (n 358) 24; In this sense, see also GenAI pending cases. As of the time of writing this article, there are some ongoing GenAI cases that could bring some clarity see, for instance, in the US see *Sarah Andersen et al. v. Stability AI et al*Case No. 3:23-cv-00201-WHO; *The New York Times Company v. Microsoft Corporation et al*., plainti's complaint, Case 1:23-cv-11195 (27December 2023); *Richard Kadrey et al. v. Meta Platforms*, Case 3:23-cv-03417-VC at 2. (N.D. Cal. 20 November 2023). *P.M. v. GitHub Copilot, Inc., et al*. the case was led in the Northern District of California on June 28, 2023; *Paul Tremblay and Mona Awad v. OpenAI LP, et al.* (N.D. Cal., led July 5, 2023); *Sarah Silverman v. Meta Platforms, Inc. and OpenAI LP, et al.* (N.D. Cal., led July 7, 2023*). Google Bard (J.L. v. Alphabet Inc,* U.S. District Court for the Northern District of California, No. 3:23-cv-03440; In the UK see for instance *Getty Images (US) Inc. et al. v. Stability AI Ltd*., [2023] EWHC 3090 (Ch).

361 Katherine Lee et al (n 72) 51.

362 For technical aspects see Daniele Panfilo et al.,'A Deep Learning-Based Pipeline for the Generation of Synthetic Tabular Data' (2016) 11 IEEE.

However, there might be a risk that synthetic data could degrade over successive generations, leading to the development of less reliable GenAI systems.[363] In this sense, a combination of synthetic data, which is non-copyrighted, with limited amounts of "fresh" real-world data, could be one of the possible solutions for AI training.[364] This approach would ensure the generation of more effective and robust outcomes, and it might also make it easier to address them in terms of copyright.

To sum up, more research is needed to examine the potential of synthetic data to mitigate the tension between TDM users, including AI developers, and copyright holders. An interdisciplinary study on synthetic data involving IT and AI experts, as well as legal specialists, would help develop concrete legal standards and mature tools for the use of synthetic data in the course of TDM, for instance, to train GenAI models, without the risk of copyright infringement.

2.8 Concluding Remarks

TDM is one of the most powerful analytical techniques, enabling users to extract valuable information from huge amounts of digital data. TDM can be used in both public and private spheres to revolutionise almost every area of our lives. However, since the process involves copying obtained data, a conflict of rights and interests has arisen between copyright and *sui generis* database holders, on the one hand, and TDM users, on the other. There was a need for proper regulation of TDM, which would balance competing interests and thereby ensure free communication of ideas and facts while protecting artistic creativity. The pre-existing laws and exceptions proved insufficient to address the needs of rapidly evolving digital technologies and AI.

As a result, the long-awaited specific TDM exceptions under the CDSM Directive were intended to provide more legal certainty and clarify the lawfulness of TDM activities. However, the exceptions also came with some limitations, which could be problematic in terms of fundamental rights to information and research. To tie together the discussions in this chapter, it should be emphasised that the readjustment of the current EU regulation on TDM can only be achieved through amendments to the existing exceptions. Moreover, any reconsiderations or interpretations of these provisions should be carried out in light of fundamental rights. The next chapter will develop further conclusions while introducing the original publications.

363 Yuxi Ma et al.,'Brain in a Vat: On Missing Pieces Towards Arti cial General Intelligence in Large Language Models' (2023) 4; Thomas Göbel et al.,'Data for Digital Forensics: Why a Discussion on'How Realistic is Synthetic Data' is Dispensable' (2023) 4(3) Digit. Threat. Res. Pract. 1, 14.

364 Sina Alemohammad et al (n 353) 7; Thomas Göbel et al (n 363) 15.

3 Presentation of the Annexed Articles

In this chapter, I will introduce four published articles that constitute this dissertation in more detail. The articles contribute to the current discussions on the readjustment of the EU regulation of TDM in the age of AI. The articles present a cohesive analysis that begins with identifying and contextualising regulatory issues that could impede data analysis, then moves on to proposing feasible ways to improve them. The chapter contains a summary of each article, clarifying their contributions to the overarching research objective. The synopsis is followed by reflections on recent developments and their relation to the original publications.

Full-length articles are provided as annexes to this dissertation.

3.1 The Insufficiency of the EU's Text and Data Mining Exceptions for Using Artificial Intelligence

The first research article serves as an introduction to the topic. The regulatory issues are mainly discussed from the perspective of AI developers. Therefore, the paper first explains how the use of TDM contributes to the development and functioning of AI technologies. Next, the paper analyses how the EU legal frameworks on copyright, database, and computer programs could restrict the application of TDM in both public and private sectors. In particular, the article clarifies which exclusive rights may be affected in the course of TDM and which statutory exceptions may apply to prevent the infringement of those rights. Accordingly, the study examines the relevant provisions of the InfoSoc Directive, the Database Directive, and the Software Directive. The analysis reveals that a broadly defined right to reproduction and a "peculiar" *sui generis* database right could restrict the use of TDM, which could be detrimental, particularly to AI development.

Moreover, the paper emphasises that the closed and exhaustive list of statutory exceptions and limitations provided under those legal frameworks cannot be adjusted to advanced digital technologies such as TDM. The rationale behind this is that the provisions, which were designed during the initial phase of the digital era, are mostly

optional, leading to fragmented regulation of TDM and legal uncertainty due to numerous conditions and requirements. Against this background, the article welcomes the adoption of the specific TDM exceptions under Arts. 3 and 4 of the CDSM Directive as an internal balancing mechanism designed to address the conflict of competing interests at stake. The paper delves into these provisions by analysing their scope, purpose, and potential applications. The findings reveal many inherent shortcomings and limitations that could dilute the efficiency of the exceptions.

The paper was written during the national transpositions of the CDSM Directive. Therefore, it was intended to provide meaningful recommendations on how the provisions might have been implemented across the EU to ensure the best possible solution for TDM regulation. In particular, the article proposes extending the scope of Art. 3, which is limited to research organisations and cultural heritage institutions, to include additional groups of beneficiaries that also do not pursue commercial purposes (e.g., individual researchers and journalists).[365] Further, it argues that there is a need to clarify the term "scientific research", which identifies the purpose of the exception, by interpreting it broadly to allow TDM for other (e.g., educational or administrative) institutional objectives.

Additionally, the paper emphasises the need to allow third parties involved in scientific research to access copies created during the TDM process for verification purposes.[366] The last issue of Art. 3, addressed in the study, relates to the application of TPMs by rightholders to ensure the security and integrity of the networks where the original content is hosted. The article emphasises that Member States should define best practices for such applications. Finally, the paper suggests clarifying how the "opt-out" mechanism of Art. 4 should work in practice through the proper interpretation of the reservation of rights, which should be done "explicitly" and "in an appropriate manner".[367]

Therefore, the article's main contribution to the research is threefold. First, it serves as an entry point for understanding a concrete problem that results from the use of copyright or *sui generis*-protected works in the course of TDM. Second, the paper explains why pre-existing exceptions cannot be applied to TDM, and why the

365 This was subsequently realised in the German Implementation Law. See the German Implementation Law (Gesetz zur Anpassung des Urheberrechts an die Erfordernisse des digitalen Binnenmarktes), para.60d(3)2. Moreover, the Italian legislator went further by allowing researchers to not only reproduce or extract protected content in the course of TDM but also communicate the findings to the public. See The Italian legislative Decree art.70-ter (Decreto Legislativo 8 novembre 2021, n 177).

366 This was also the case covered by the German Implementation Act, however, with the requirement to terminate the public disclosure of the materials once the verification is completed. see Paragraph 60d(4)1 and 2 of the German Implementation Act (Gesetz zur Anpassung des Urheberrechts an die Erfordernisse des digitalen Binnenmarktes).

367 Art 4(3) of the CDSM Directive.

legal framework introduced by the CDSM Directive was necessary to mitigate the tension between competing rights and interests. Finally, the study clarifies how the CDSM Directive's TDM exceptions could be improved through national implementation. However, in practice, the overwhelming majority of EU Member States have implemented the CDSM Directive's TDM exceptions by simply copying and pasting the wording of the Directive.

As a result, many of the issues discussed in the article remain open. Overall, the article sets the stage for further, more specific, and contextualised research on the topic, as presented in the following three articles.

3.2 In Search of Balance: Text, Data Mining and Copyright in the Digital Single Market Directive from a Fundamental Rights Perspective

The second article aims to answer the question of how the CDSM Directive's TDM exceptions have improved the situation in terms of fundamental rights, and what limitations could dilute their efficiency. First, the paper introduces the right to information, the freedom of research, and the right to IP as recognised under the EUCFR. Next, it analyses how the exceptions have contributed to the realisation of these fundamental values.

The paper recognises many positive aspects of Arts. 3 and 4 of the CDSM Directive, which are crucial for balancing copyright and users' fundamental rights to information and research. First, the study highlights the importance of the mandatory approach to TDM regulation, which ensures a harmonised legal framework for TDM across the EU.

Next, Art. 3 is appreciated for providing more legal certainty for researchers, covering both commercial and non-commercial research purposes, permitting the use of TDM within the framework of public-private partnerships, allowing the storage of copies made in the course of TDM for subsequent scientific research purposes, including for the verification of research results, and more. Following that, the paper emphasises that the exception under Art. 4 of the CDSM Directive has also significantly improved the regulation of TDM. The provision addresses the crucial aspects not covered by Art. 3, as it applies to all types of users and all purposes of TDM activities, thereby benefiting private actors and, particularly, the EU AI sector. The article emphasises that the exceptions have been acknowledged as a major step and a new hope towards better regulation of TDM in the EU.

The second part of the article demonstrates that, despite many positive aspects, the specific TDM exceptions consist of certain limitations which could dilute their efficiency in terms of fundamental rights. The paper takes a critical view of the limitation on the scope of beneficiaries of Art. 3, its purpose limitation, the lawful

access requirement included in both provisions, the "opt-out" mechanism of Art. 4, the storage conditions, and the restrictive effect of TPMs. In this sense, the article argues that the right to IP, as protected under Art. 17 (2) of the EUCFR, is not an absolute right and should be balanced against other fundamental rights at stake. The analysis is conducted in light of the concept of a fair balance, derived from the principle of proportionality.

The article finds that, despite many positive aspects, the CDSM Directive's TDM exceptions do not sufficiently take into account the right to information and the freedom of research, as these values have only been addressed superficially. This has been clearly reflected in the Finnish implementation of the CDSM Directive, where the Parliament's Constitutional Law Committee rejected the government's draft implementation of the Directive on the grounds that it was not in line with the Finnish Constitution.[368] The Committee found, *inter alia,* that TDM exceptions had not taken the freedom of science into account in a balanced way.[369]

Finally, the article proposes amendments to the regulation of TDM that could be justified from a fundamental rights perspective. It is concluded that more balanced regulation could ensure that research and innovation in the context of data analysis are not unfairly restricted by excessive protection of copyright and the *sui generis* database right. The key contribution of the article pertains to its application of a fundamental rights approach, which helps determine what law on TDM ought to be.

3.3 Overcoming Barriers to Text and Data Mining in the Era of ChatGPT: The Proposed Data Act as a Game-Changer

The third article seeks to assess whether Art. 35 of the proposed Data Act (now Art. 43 of the Data Act) could provide a remedy for researchers encountering difficulties in relying on the CDSM Directive's TDM exceptions. The paper considers a real-life scenario in which TDM research is conducted using AI language models, such as ChatGPT, to analyse large medical databases and uncover valuable information for improving medical diagnosis and treatment. The illustration helps elucidate how the specific TDM exceptions would function in practice and which alternative solutions might be applied.

First, the article provides a brief technical background explaining the development and functioning of AI technologies, such as ChatGPT, which are based on TDM. Next, it indicates which legal issues may arise where original works or

368 The Finnish Parliament's Constitutional Law Committee statement PeVL 58/2022 vp HE 43/2022 vp (14 November 2022).

369 Ibid.

databases are copied without authorisation in the course of TDM. In particular, the article focuses on cases in which TDM is carried out on non-original databases that can be protected under the *sui generis* database regime.

Before delving into the main analysis, the study introduces the *sui generis* database right and discusses its relevance to, and impact on, TDM. The article argues that the protection restricts TDM by intensifying the control over access to data. Moreover, the exceptions to the *sui generis* protection provided under Arts. 8(1) and 9 (b) of the Database Directive are neither practical nor effective for the purposes of TDM. Therefore, the article elaborates on the potential application of the CDSM Directive's TDM exceptions to ChatGPT-based research under discussion. The study finds that both research organisations and private clinics using TDM tools to research medical databases could encounter significant obstacles and legal uncertainty when relying on these provisions.

Therefore, the article examines Art. 35 of the proposed Data Act as an alternative solution that could help prevent potential infringement claims. The provision excludes databases made of data generated by the use of a product or related service (IoT data) from protection under the *sui generis* database regime.

The article argues that while the proposed regulation intends to free machine-generated data from the constraints of that protection, many aspects should be clarified to ensure greater access to data for research and other purposes. In particular, the analysis suggests that it should be specified that the exclusion of databases made of machine-generated data from *sui generis* database protection is absolute. It should not be limited to cases of interference with only certain users' rights outlined in Arts. 4 and 5 of the proposed Data Act, or to instances where such databases fail to qualify for that protection.

Moreover, the paper argues that the inclusion of derived or inferred data, as well as researchers' specific needs, within the scope of the Regulation would provide additional benefits for users. Additionally, the article claims that the possibility of overriding Art. 35 of the proposed Data Act through contractual agreements or the application of TPMs could create an additional burden on users and undermine their rights to access and use data.

The study concludes that the Regulation could address certain problematic issues in the context of TDM regulation if properly amended. In this sense, the findings of the paper offer practical guidance to researchers using AI tools, such as ChatGPT, to mine databases on how best to avoid potential infringement of the *sui generis* database right. The main contribution of the article pertains to exploring alternative solutions for cases where the CDSM Directive's TDM exceptions are ineffective in addressing the needs of TDM users, particularly researchers. It presents a perspective on Art. 35 of the proposed Data Act by assessing its capacity to alleviate the tension between *sui generis* database holders and TDM users.

The article was prompted by arguments presented in previous publications, which indicate that the current regulation of TDM tends to prioritise the rightholders' rights and interests at the expense of TDM users. The paper, written while the Data Act was still at the proposal stage, suggests new solutions for creating a more balanced and research-friendly framework for TDM in the EU. If implemented, the amendments would promote cutting-edge scientific research based on TDM and advanced AI tools that could strengthen the EU's research and innovation capabilities.

In this sense, it should be emphasised that the final text of the Data Act includes some of the recommendations provided in the paper. For instance, Recital 112 of the Data Act clarifies that the *sui generis* right should not apply to databases made of machine-generated data in all cases. However, many other important aspects discussed above remain unaddressed.

3.4 Can Text and Data Mining Exceptions and Synthetic Data Training Mitigate Copyright-Related Concerns in Generative AI?

The fourth article focuses on the copyright-related concerns in the field of GenAI. The article aims to analyse both legal (the CDSM Directive's TDM exceptions) and technical (synthetic data) solutions, emphasising their strengths and weaknesses.

First, the paper familiarises the reader with GenAI models, which can generate new information based on insights and patterns derived from existing digital data. It explains how these AI systems are built. The process involves two key steps: training (a so-called "input" phase) and content generation (a so-called "output" phase). Put simply, a computer analyses massive volumes of existing data by extracting underlying knowledge (e.g., insights and trends), which is then used to feed AI algorithms and ultimately generate the desired output. The article focuses on the training of GenAI models, which relies on TDM techniques and may potentially lead to copyright infringement if performed on protected works without prior authorisation. The article emphasises that training these models only on public domain works, or on data freely available online, might not be sufficient to generate high-quality and accurate outputs.

Second, the article argues that the specific TDM exceptions under the CDSM Directive could help alleviate the tension between copyright holders and providers of GenAI systems. The analysis begins with the introduction of the right to train, which could be derived from the fundamental right to information and the freedom of research. The article highlights that by analysing existing digital data in the course of training, providers of GenAI models make it more accessible and understandable

to the public. The hidden knowledge derived from such data could effectively address the public's needs in various domains.

In this sense, the paper argues that unduly restricting reproductions occurring in the course of TDM may limit access to the very ideas and facts, thereby interfering with the fundamental values at stake. It further argues that copyright exceptions and limitations serve as an essential regulatory safeguard capable of balancing fundamental rights in the digital environment. Therefore, the study delves into a critical analysis of Arts. 3 and 4 of the CDSM Directive in the context of AI development.

However, since GenAI models are typically developed by private actors for commercial purposes, the main focus is on the latter, widely known as the "commercial" TDM exception. In particular, the study examines the reservation right, a so-called "opt-out" mechanism, of Art. 4, which raises a primary concern. The article argues that the mechanism has strengthened the position of copyright holders, allowing them to control the use of their works for TDM. AI developers would have to pay twice to be able to train their systems, first to obtain lawful access to data, and second, to analyse training datasets.

Moreover, the article argues that there are no generally accepted protocols or standards that could clarify the practical aspects of the reservation right, and this has led to the fragmentation of "opt-out" standards. AI developers offer various model-specific solutions to rightholders to restrict or permit access to their works for AI training (e.g., the Spawningai.txt website[370]) or to search for their works in training datasets and opt them out (e.g., the HaveIBeenTrained website[371]).

In this sense, the article emphasises that the AI Act's transparency obligation, requiring providers of GenAI models to reveal data used for training their systems, may clarify certain aspects of the reservation right of Art. 4. However, the paper argues that allocating the right to permit or prohibit the use of protected works for TDM to copyright holders may undermine the legitimate users' right to access information and analyse the underlying ideas and facts. These fundamental interests, which are crucial for the development of GenAI systems, should be effectively protected and realised. Against this background, the study finds that the best solution is not to refine the "opt-out" mechanism, but to establish an efficient legal framework for TDM in the form of a broader exception that does not include the reservation right.

Third, the article considers synthetic data as one of the technical solutions to the conflict. The idea of examining this approach and evaluating its strengths and weaknesses in the context of GenAI, TDM, and copyright is novel. The article argues

370 See <https://site.spawning.ai/spawning-ai-txt> accessed 10 January 2024.
371 See <https://haveibeentrained.com > accessed 10 January 2024.

that AI-synthesised data can be used for training when original data is scarce or when access to materials is subject to certain legal limitations (e.g., privacy or data protection). It is also cheaper and less demanding to generate synthetic data than to obtain original data.

However, the article emphasises that the approach may not fully resolve the conflict, as data synthesis *per se* relies on TDM performed on real-world data. On the one hand, the paper argues that once permitted, AI-synthesised data can be used to generate more variations of synthetic data, which would have fewer (or no) links to actual data and, accordingly, may not lead to copyright infringement. On the other hand, the study emphasises that there could be a risk of data degradation if data is generated only from "pure" synthetic data without the inclusion of fresh, original data. The article concludes that further interdisciplinary research is needed to explore the potential of synthetic data in relation to copyright and TDM.

Finally, the article finds that both legal and technical approaches to the regulation of TDM have their errors and omissions. The remedies would not be able to resolve the conflict between the rights and interests of copyright holders and those of providers of GenAI models. However, they could mitigate the tension if refined properly. The study offers a trampoline for further active actions in both legal and technical spheres.

3.5 Recent Developments

The dissertation focuses on current issues that trigger significant changes in society. Therefore, it is not surprising that, following the publication of the original articles, important developments have occurred.

3.5.1 Kneschke vs. LAION

At the time of writing this introduction, there were three national rulings that have opened fresh perspectives on, and clarified (to some extent), the application of the CDSM Directive's TDM exceptions. The first landmark ruling, *Kneschke vs. LAION,*[372] by the German Regional Court (Court), was discussed in one of my blog posts published as part of this research.[373] Kneschke, a photographer, sued LAION, a non-profit organisation that offers open datasets for training, for using one of his works for AI training without obtaining prior permission, potentially infringing copyright.

[372] *Kneschke v. LAION*, Landgericht München I (District Court of Munich I), Judgment of 27 September 2024, Case No. 23 O 00061/23.

[373] Maryna Manteghi, 'Kneschke vs. LAION: Opening Fresh Perspectives on AI Training and TDM Exceptions' (*European Law Blog*, January 2025).

The defendant claimed that they downloaded the photo for verification purposes, and that the training datasets provided only a hyperlink to a website where it was freely available to the public. LAION argued that their actions fell within the scope of TDM exceptions provided under the German Copyright Act[374] and the EU CDSM Directive. The Court dismissed the plaintiff's claim of unauthorised reproduction, finding that the actions were covered by the TDM exception for scientific research.[375]

The Court held that the preparatory phase of AI training is a prerequisite for acquiring new knowledge in the course of training, and there is no need to prove that such research contributes to later research success. The Court clarified that the process of building training datasets may be regarded as "scientific research". Moreover, the Court emphasised that LAION could benefit from Art. 3, as the company makes training datasets publicly available free of charge, and it is irrelevant that commercial entities may use such data for AI training.[376] In this sense, the Court's reasoning may strengthen the position of research organisations in their cooperation with private companies.[377]

Further, in *obiter dicta*, the Court provided some clarifications regarding the practical aspects of the reservation of rights under Art. 4 of the CDSM Directive. It held that the prohibition of content scraping by automated tools written in natural language in the terms and conditions of the website constitutes an "opt-out" mechanism within the meaning of the Directive. However, it remains unclear whether all machines can read and interpret such complex legal rules properly under different circumstances.[378]

Overall, the judgment holds considerable importance for future cases on AI training and copyright, even though the analysis is limited to the preparatory phase, and the reasoning may appear vague and raise additional questions.

374 German Copyright Act (*Urheberrechtsgesetz*, UrhG), Act on Copyright and Related Rights of 9 September 1965, BGBl. I 1965, 1273, as last amended by the Act on the Adaptation of Copyright Law to the Requirements of the Digital Single Market, BGBl. I 2021, 1204.

375 Art. 3 of the CDSM Directive and Sec. 60 d of the German Copyright Act.

376 Maryna Manteghi (n 373).

377 Maryna Manteghi (n 373).

378 Maryna Manteghi (n 373).

3.5.2 DPG Media et al vs. HowardsHome

Further, in the recent Dutch *DPG Media et al vs. HowardsHome* case,[379] ruled upon by the Amsterdam District Court (Court), the Court evaluated the CDSM Directive's press publishers' right[380], as well as the proper implementation of an "opt-out" mechanism in compliance with a "commercial" TDM exception.[381] The defendant provided a news alerts service, "HowardsHome Nieuws", to both public and private entities for many years. It obtained information mostly from publicly available internet RSS feeds of news articles, which included the title, publication date, thumbnails, a short description (snippets), and a hyperlink to the relevant text. The publishers (the plaintiffs) claimed that these activities infringed their exclusive rights to reproduction and making available to the public.

With regard to TDM, the Court held that since the plaintiffs had failed to prove that they effectively opted out in light of Art. 4 and Recital 18 of the CDSM Directive, the defendant's reliance on the "commercial" TDM exception was therefore justified. The publishers only excluded specific big AI bots such as ChatGPT-User, CCBot, Anthropic-AI and others. The interpretation of the requirement for a "machine-readable" rights reservation, which should be done explicitly in an appropriate manner, is not consistent and differs from that in *Kneschke vs. LAION.* It is anticipated that the Commission and the AI Office, which have already started working on a Code of Practice, will provide more clarity regarding the modalities and methods of the "opt-out" mechanism.[382]

3.5.3 Gamekapocs case

In the latest EU court decision related to TDM *Gamekapocs*[383]*,* the Hungarian Municipal Court of Appeals had to evaluate, *inter alia,* whether the scraping of the

379 *DPG Media et al vs. HowardsHome*, Rechtbank Amsterdam (Amsterdam District Court), Judgment of 30 October 2024, ECLI:NL:RBAMS:2024:6563.

380 Art. 15 of the CDSM Directive.

381 Article 15o of the Dutch 'Auteurswet' (Dutch Copyright Act), *Wet van 23 september 1912, houdende nieuwe regeling van het auteursrecht* (Stb. 1912, 308), as last amended by *Wet van 29 juni 2021 tot implementatie van Richtlijn (EU) 2019/790* (Stb. 2021, 305) (implementing art. 4 of the CDSM Directive).

382 ECS, 'Copyright and Generative AI: Opinion of the European Copyright Society' (2025), 7; Code of Practice on the EU AI Act: General Purpose AI (2025); EC, Third Draft of the General-Purpose AI Code of Practice published, written by independent experts (2025).

383 *Gamekapocs*, Municipal Court of Appeals (Hungary), Case No. 9 Pf. 20.353/2024/6-II, Judgment of 3 December 2024, see also Péter Mezei, 'The Multi-layered Regulation of Rights Reservation (Opt-out) Under EU Copyright Law and the AI Act -For the Benefit of Whom?' (v3.0) (March 31, 2025), 16-17 and Peter Mezei, 'Third European

plaintiff's website by Google's search engine for indexing available materials could be covered by Art. 4 of the CDSM Directive.[384] The Court held that the defendant lawfully accessed the plaintiff's press publications without bypassing TPMs. The Court highlighted that the robot exclusion protocol employed by the plaintiff on its website, which serves as a machine-readable "opt-out" mechanism, did not object to indexing by the search engine according to statutory requirements. The ruling determined that web scraping and search engine indexing should be regarded as a "form" of TDM, an approach also applied in the third draft of the General Purpose AI Code of Practice.[385]

In this sense, *Mezei* fairly argues that such a "conflation" cannot be viewed as reasonable in light of the purpose of the CDSM Directive.[386] Although the Directive broadly defines TDM[387] to include research, the development of new technologies, and more, it does not exempt all forms of automated data analysis, including web scraping or indexing, from liability.[388] Including such processes under Art. 4 would conflate distinct technologies under a single legal rule, which may not comply with the three-step test, as it could exceed the limits of "certain special cases". In this sense, *Mezei* rightly emphasises that this would "effectively transform the exception into a general rule".[389]

Further, he argues that rightholders are normally interested in the indexing of their content to increase visibility, but they may not want their content indexed for broader commercial purposes. Therefore, conflating indexing with automated data analysis would undermine the rightholders' ability to control distinct technologies.[390] Additionally, Art. 4(2) of the CDSM Directive permits the storage of copies made in the course of TDM only for as long as necessary for the purpose of TDM.[391]

Court Decision on the General Purpose TDM Exception is Out' (*Kluwer Copyright Blog*, May 8, 2025).

384 Ibid.

385 Ibid.; EC, Third Draft of the General-Purpose AI Code of Practice published, written by independent experts (2025) (n 44); Marianna Foerg, 'Second and Third Dratfs of the general-Purpose AI Code of Practice Have Been Released' (*Kluwer Copyright Blog*, April 4 2025).

386 Peter Mezei (n 383) 16-17; Peter Mezei, 'Third European Court Decision on the General Purpose TDM Exception is Out' (*Kluwer Copyright Blog*, May 8, 2025) (n 383).

387 Art 2 (6) of the CDSM Directive.

388 Recitals 8 and 18 of the CDSM Directive.

389 Peter Mezei (n 383) 17.

390 Hanjo Hamann, 'Artificial Intelligence and the Law of Machine-Readability: A Review of Human-to-Machine Communication Protocols and their Compatibility with Article 4(3) of the Copyright DSM Directive' (2024) 15(2) JIPITEC 102-117.

391 Art 4(2) of the CDSM Directive.

However, in the present case, the Court appears to legitimise the permanent storage of data collected for indexing purposes, which would exceed the scope of the "commercial" TDM exception.[392] Against this background, the approach would create legal uncertainty regarding the interpretation of the TDM exception under Art. 4 of the CDSM Directive and deprive rightholders of the right to opt out of TDM.

Overall, the EU case law on TDM has not provided effective practical guidance on the issues discussed in this dissertation so far. The rulings employ different approaches to the interpretation of the CDSM Directive's TDM exceptions and leave many questions open. In particular, the reservation of rights dilemma of Art. 4 of the CDSM Directive has not been properly addressed, and there are still no generally accepted protocols or standards that clarify the "machine-readable" format of the "opt-out" mechanism. Therefore, a uniform interpretation and framing of the concept by the CJEU is needed.

3.5.4 The Opinion of the European Copyright Society on Copyright and GenAI

Following the above-discussed EU courts' decisions related to TDM, 2025 opened with fresh perspectives and debates on TDM, copyright, and GenAI. Notably, in January 2025, the European Copyright Society (ECS), led by prominent scholars and academics, published its Opinion on copyright and GenAI.[393] The ECS argues that certain aspects of the development of GenAI have not been sufficiently considered under the CDSM Directive and the AIA, which could lead to legal uncertainty.

First, in the view of the ECS, it is important to clarify the exact scope of the CDSM Directive's TDM exceptions in the context of GenAI. This would provide more legal certainty regarding the copyright status of reproductions and other acts performed at each phase of GenAI development and operation.[394] Further, the ECS argues that it is necessary to identify the technologies that rightholders can employ to opt out according to Art. 53(1)(c) of the AIA. As discussed above, the provision requires providers of general-purpose AI models to comply with the reservation of rights of Art. 4(3) of the CDSM Directive.

Moreover, the ECS emphasises that it is also necessary to determine the rightholders entitled to the reservations of rights, as well as the timing and the location of such reservations.[395] For instance, it is not clear whether the reservation

392 Peter Mezei (n 383) 17.
393 ECS (n 382).
394 ECS (n 382) 1.
395 ECS (n 382) 7.

should be expressed where the protected work is stored or where it has been made lawfully accessible.[396]

Furthermore, the ECS argues that more clarification is needed regarding the transparency obligation provided under Art. 53(1)(d) of the AIA and its impact on the lawful access requirement of Arts. 3 and 4 of the CDSM Directive. Finally, the ECS highlights the importance of research and open-source models in the AI sector and also the need to ensure fair remuneration for authors and other rightholders in the context of GenAI developments and performances.[397]

Most of the concerns raised in the Opinion were discussed in this dissertation. The ECS reaffirms the assumptions regarding the relationship between the AIA and the CDSM Directive, which may prove fruitful, particularly in relation to the realisation of the "commercial" TDM exception and its "opt-out" mechanism.

3.5.5 The EU IP Office's Study on GenAI and Copyright

Finally, the EU IP office's study on GenAI and copyright[398] (hereafter the study) represents the most recent development on TDM at the time of writing this introduction. The study explores key technical and legal issues related to the development of GenAI in terms of EU copyright law.[399] The main focus is on the specific TDM exceptions provided under the CDSM Directive and the copyright-relevant provisions of the EU AIA, particularly the obligation to comply, identify and respect the reservations of rights, and the transparency requirement.

The study also analyses the ongoing legal challenges and litigation in the EU, the US, and other non-EU countries (such as China, the UK, Canada, India and South Korea). Regarding the EU *Kneschke vs. LAION* case, discussed earlier in this section, the study highlights that there could be a risk of "data laundering", where training datasets are built by relying on the TDM exception for scientific research, which is not limited by the reservation of rights, but subsequently used commercially.[400]

In this sense, the study argues that, as direct licensing markets are rapidly emerging and developing, the main challenge lies in the lack of clarity and accuracy in licensing terms (e.g., pricing benchmarks or remuneration models).[401] Many issues may arise, including a shortage of high-quality data, market imbalance where

396 ECS (n 382) 8.
397 ECS (n 382) 8-9, 11-12.
398 EU IPO, The Development of Generative Artificial Intelligence from a Copyright Perspective (2025).
399 EU IPO (n 398) 21-22.
400 EU IPO (n 398) 14, 118.
401 EU IPO (n 398) 87-107.

smaller or less financially stable companies cannot afford to gain access to data, lack of transparency, and others.

Further, the study reinforces the issue also examined in this dissertation, related to the lack of a uniform standard for the "opt-out" mechanism of Art. 4, by comparing the advantages and limitations of different rights reservation measures.[402] Next, the study also explores challenges that might arise in the "output" phase, such as the risk of memorisation and the generation of infringing content. It analyses various output transparency measures and their capability to ensure that generative content is detectable in light of the AIA (e.g., through content tagging, digital fingerprinting, watermarking, etc.).[403]

Overall, the study intends to clarify the interaction between GenAI and copyright from a technical perspective, to increase public awareness of copyright-related concerns that may arise during the development and operation of GenAI systems.

402 EU IPO (n 398) 230-234.
403 EU IPO (n 398) 13, 65-66.

4 Conclusions of the Dissertation

In this doctoral dissertation, I examined whether the EU's current regulation of TDM sufficiently considers the interests related to technological developments and the fundamental freedoms of research and information. The dissertation consists of this summary and four substudies. During the research project, I focused on three key aspects: first, I explored how copyright and the *sui generis* database right could restrict the use of TDM in the EU. The focus was on the copyright holder's exclusive right to reproduction and the *sui generis* database right to extraction, as well as the exceptions and limitations to these rights. Second, I examined the reasons for adopting specific TDM exceptions and subsequently analysed how the situation has changed after the introduction of Arts. 3 and 4 of the CDSM Directive. The main aim was to identify and analyse the advantages and limitations of these provisions in terms of the fundamental rights to information and research, particularly in the context of AI development.

Finally, I discussed the most effective legal and technical means for the proper EU regulation of TDM to ensure a long-lasting balance between competing interests amid rapid technological developments. As an example, I explored the developments and operations of GenAI models from the perspective of EU copyright law and provided fresh insights into the application of TDM exceptions to these processes.

The study found that the specific TDM exceptions provided under Arts. 3 and 4 of the CDSM Directive have improved the regulation of TDM in terms of fundamental rights. The provisions introduced more legal clarity and certainty for both researchers and commercial actors. However, the analysis revealed certain limitations in the scope of these exceptions, which could dilute their effectiveness and, thus, hinder, *inter alia,* crucial developments within the EU AI sector.

The research found that certain constitutional values, such as the right to information and the freedom of research, were only addressed superficially. In particular, the study demonstrated how the limitation on the scope of beneficiaries and the purpose of Art. 3, the lawful access requirement and the storage conditions included in both provisions, the "opt-out" mechanism of Art. 4, and the restrictive effect of TPMs could preclude the application of these exceptions in various fields.

As a result, the study explored alternative solutions for balancing rightholders' rights and interests with those of TDM users, such as the Data Act and synthetic data training. The research identified significant potential in these approaches, but it also highlighted the need for legal amendments concerning the formal solution and further interdisciplinary research with respect to the latter.

Based on my findings, I argue that Arts. 3 and 4 of the CDSM Directive, which constitute the core of the EU regulation of TDM, should be amended to ensure that fundamental freedoms are taken into account in a balanced way in the age of AI. The proper revision of the provisions would ensure that strong protection of copyright and *sui generis* database rights does not unduly restrict access to information and advanced data analysis in the EU.

First, I argue that it is necessary to extend the scope of beneficiaries of Art. 3 to permit any person or entity with lawful access to protected works to carry out TDM for the purpose of scientific research. Second, it may be necessary to use the term "research" instead of "scientific research" when referring to the purpose of Art. 3, which would cover a broader range of systematic studies. Third, the concept of "lawful access" referred to in both provisions should be interpreted in light of the term "lawful use," to ensure that the lawfulness of access can be determined not only by contractual terms but also when access is not restricted by law.

Fourth, it is recommended that the conditions under which copies of mined materials can be stored be clarified, and that their sharing be permitted for scientific peer review, the verification of research outcomes, and joint research. Fifth, more clarification is needed regarding the application of TPMs to prevent the risk of erroneous or unjustified blocking of lawful access. Finally, it is necessary to clarify the machine-readable format of the reservation of rights in Art. 4 of the CDSM Directive. In particular, it might be reasonable to establish a standardised "opt-out" mechanism that would apply to all uses falling within the scope of this provision and meet the transparency obligations of the AIA. However, enabling rightholders to control the use of their works for TDM could undermine lawful users' fundamental rights to access information and conduct research on unprotected ideas and facts underlying such materials. Therefore, ideally, a broader "commercial" TDM exception, not including the reservation of rights, might be the optimal solution.

To sum up, the study highlights the need to readjust the current EU regulation of TDM to ensure that it keeps pace with rapid AI development. The synthesis of my published articles indicates that the CDSM Directive's TDM exceptions should be amended and interpreted in light of fundamental rights. This would help address copyright-related concerns raised by the development and functionality of game-changing AI applications and systems, while also strengthening the research power of the EU at a global level.

Therefore, much remains to be done to ensure a fair balance between the rights and interests of copyright and *sui generis* database owners and those of TDM users. It would be reasonable to close this introductory part of my dissertation with a famous quote by Jimmy Carter, which perfectly conveys the central message of this research: "We must adjust to changing times and still hold to unchanging principles."[404]

[404] Inaugural Address of Jimmy Carter (January 20, 1977).

Abbreviations

AI	Artificial Intelligence
AIA	Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024, laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) [2024] OJ L 2024/1689, 12.7.2024.
Berne Convention	Berne Convention for the Protection of Literary and Artistic Works, adopted 9 September 1886, as revised at Paris on 24 July 1971.
CJEU	Court of Justice of the European Union
CDSM Directive	Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on Copyright and Related Rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC [2019] *OJ L 130*, 17.5. 2019, p. 92–125.
Data Act	Regulation (EU) 2023/2854 of the European Parliament and of the Council of 13 December 2023 on harmonised rules on fair access to and use of data and amending Regulation (EU) 2017/2394 and Directive (EU) 2020/1828 (Data Act) (Text with EEA relevance) PE/49/2023/REV/1 *OJ L, 2023/2854,* 22.12.2023.
Database Directive	Directive (EU) 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases [1996] *OJ L 77/20*.
DSM	Digital Single Market
EU	European Union
EU IPO	European Union Intellectual Property Office
EUCFR	European Charter of Fundamental Rights
ECHR	European Convention of Human Rights
EcommHR	European Commission of Human Rights
ECS	European Copyright Society
ECtHR	European Court of Human Rights

GenAI	Generative AI
ICCPR	International Covenant on Civil and Political Rights
InfoSoc Directive	Directive 2001/29 on the harmonisation of certain aspects of copyright and related rights in the information society [2001] OJ L167/10, 22 June 2001, pp. 10-19.
ICCPR	International Covenant on Civil and Political Rights
ICESCR	International Covenant on Economic, Social and Cultural Rights
IP, IPR	Intellectual Property, intellectual property right (s)
IoT	Internet of Things
IT	Information Technology
Software Directive	Directive 2009/24/EC of the European Parliament and of the Council of 23 April 2009 on the legal protection of computer programs (Codified version) (Text with EEA relevance) *OJ L 111*, 5.5.2009, p. 16–22
TDM	Text and Data Mining
TPMs	Technological Protection Measures
TRIPS	Agreement on Trade-Related Aspects of Intellectual Property Rights, Annex 1C of the Marrakesh Agreement Establishing the World Trade Organisation, signed in Marrakesh, Morocco on 15 April 1994.
UK	United Kingdom
UDHR	Universal Declaration of Human Rights
US	United States
WIPO	World Intellectual Property Organisation
WTO	World Trade Organisation

List of References

Literature

Abdul Majeed, 'Attribute-Centric and Synthetic Data-Based Privacy Preserving Methods: A Systematic Review' (2023) 3 Journal of Cybersecurity and Privacy 638.

Aikaterini Simopoulou, 'Text and Data Mining under EU Copyright Law' (International Hellenic University, Thessaloniki, 2020) <https://repository.ihu.edu.gr/xmlui/bitstream/handle/11544/29743/Text%20and%20Data%20Mining%20under%20EU%20Copyright%20Law.pdf?sequence=1> accessed 12 February 2025.

Aharon Barak, *Proportionality: Constitutional Rights and their Limitations* (Cambridge University Press 2012).

Andres Guadamuz, 'A Scanner Darkly: Copyright Liability and Exceptions in Artificial Intelligence Inputs and Outputs' (2024) 73(2) GRUR International 111

Andrew Tyner, 'The EU Copyright Directive: 'Fit for the Digital Age' or Finishing It?' (2020) 26 Journal of Intellectual Property Law 275.

Anol Bhattacherjee, *Social Science Research: Principles, Methods, and Practices* (University of South Florida, 2012) <https://digitalcommons.usf.edu/oa_textbooks/3/> accessed 12 February 2025.

Apoorva Verma, 'The Copyright Problem with Emerging Generative AI' (2023) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4537389 > accessed 12 February 2025.

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?' (2023) 27(1) The Journal of World Intellectual Property 44.

Aryan Jadon and Shashank Kumar, 'Leveraging Generative AI Models for Synthetic Data Generation in Healthcare: Balancing Research and Privacy' (*arXiv*, 2023) <https://arxiv.org/abs/2305.05247 > accessed 21 January 2025.

Avanika Narayan et al., 'Can Foundation Models Wrangle Your Data?' (2022) 16(4) PVLDB 738.

Aurora Airaskorpi and Outi Vanharanta, 'The state of industry-academia collaboration in the social sciences, humanities and arts: Perspectives from the Nordics and beyond' (Vaikuttavuussäätiö 2023) <https://www.vaikuttavuussaatio.fi/wp-content/uploads/2023/11/Shape-report-FINAL.pdf> accessed 12 April 2025.

Authors and Performers Call for Safeguards Around Generative AI in the European AI Act (*InitiativeUrheberrecht*, 19 April 2023) <https://urheber.info/media/pages/diskurs/call-for-safeguards-around-generative-ai/c93a5ab197-1681904353/nal-version_authors-and-performers-call-for-safeguards-around-generative-ai_19.4.2023_12-50.pdf> accessed 1 March 2025.

Ben Depoorter, Peter Menell, and David Schwartz (eds.). *Research Handbook on the Economics of Intellectual Property Law* (Edward Elgar Publishing (2019) <https://doi.org/10.4337/9781789903997> accessed 11 March 2025.

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-datamining-articles-3-and4/> accessed 11 March 2025.

Bernt Hugenholtz et al., *The Recasting of Copyright & Related Rights for the Knowledge Economy* (Study for the European Commission 2006) 59.

Bernt Hugenholtz and Martin Senftleben, '*Fair use in Europe. In search of flexibility*' Amsterdam Law School Research Paper 2011 <http://ssrn.com/abstract=1959554>accessed 12 May 2025.

Brian Kennett, *Planning and Managing Scientific Research: a Guide for the Beginning Researcher* (Canberra: ANU Press 2014) <https://library.oapen.org/handle/20.500.12657/33421>accessed 23 March 2025.

Carys J. Craig, 'The AI-Copyright Challenge: Tech-Neutrality, Authorship, and the Public Interest' in Ryan Abbott (ed.), *Research Handbook on Intellectual Property and Artificial Intelligence* (Edward Elgar Publishing 2022) 134.

Carys J. Craig, 'Globalising User Rights-Talk: On Copyright Limits and Rhetorical Risks' (2017) 33(1) American University International Law Revew 1.

Christophe Geiger and Vincenzo Iaia, 'The Forgotten Creator: Towards a Statutory Remuneration Right for Machine Learning of Generative AI' (2023) 52 Computer Law & Security Review (forthcoming, 2024) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4594873> accessed 23 March 2025.

Christopher Callison-Burch, 'Understanding Generative Artificial Intelligence and Its Relationship to Copyright' (2023) <https://docs.house.gov/meetings/JU/JU03/20230517/115951/HHRG-118-JU03-Wstate-Callison-BurchC-20230517.pdf> accessed 27 January 2025.

Christophe Geiger and Elena Izyumenko, 'Towards a European 'Fair Use' Grounded in Freedom of Expression' (2019) 35(1) American University International Law Review 1, 58.

Christophe Geiger and Bernd Justin Jütte, 'Conceptualising a 'Right to Research' and Its Implications for Copyright Law: An International and European Perspective' (2022) Joint PIJIP/TLS Research Paper Series No.7-2022 <https://digitalcommons.wcl.american.edu/cgi/viewcontent.cgi?article=1079&context=research> accessed 11 February 2025

Christophe Geiger and Bernd Justin Jütte, 'The Right to Research as Guarantor for Sustainability, Innovation and Justice in EU Copyright Law' in T. E. Pihlajarinne, J. Mähönen and P. Upreti (eds), *Intellectual Property Rights in the Post Pandemic World: An Integrated Framework of Sustainability, Innovation and Global Justice* (Cheltenham: Edward Elgar Publishing, forthcoming, 2023).

Christophe Geiger and Bernd Justin Jütte, 'Digital Constitutionalism and Copyright Reform: Securing Access to through Fundamental Rights in the Online World' (*Kluwer Copyright Blog*, 25 January 2022) <https://copyrightblog.kluweriplaw.com/2022/01/25/digital-constitutionalism-and-copyright-reform-securing-access-to-through-fundamental-rights-in-the-online-world/> accessed 11 February 2025.

Christophe Geiger, 'Author's Right, Copyright and the Public's Right to Information: A Complex Relationship (Rethinking Copyright in the Light of Fundamental Rights)' in Fiona Macmillan (ed.), *New Directions in Copyright Law* vol. 5 (Edward Elgar 2007) 24.

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(7) IIC, 814.

Christophe Geiger et al., 'The Exception for Text and Data Mining' (2018) Centre for International Intellectual Property Studies Research Paper No.2018-02.

Christophe Geiger and Bernd Justin Jütte, 'Platform liability under Article 17 of the Copyright in the Digital Single Market Directive, Automated Filtering and Fundamental Rights: An Impossible Match' (2021) 70(6) GRUR International 517.

Christophe Geiger and Bernd Justin Jütte, 'Designing Digital Constitutionalism: Copyright Exceptions and Limitations as a Regulatory Framework for Media Freedom and the Right to Information Online' in Martin Senftleben et al. (eds), *Cambridge Handbook of Media Law and Policy in Europe* (Cambridge University Press, forthcoming), <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4548510> accessed 13 February 2025.

Christophe Geiger, 'Elaborating a Human Rights-friendly Copyright Framework for Generative AI' (2024) Joint PIJIP/TLS Research Paper Series 123 <https://digitalcommons.wcl.american.edu/research/123>accessed 12 March 2025.

Christophe Geiger, 'The Missing Goal-Scorers in the Artificial Intelligence Team: Of Big Data, the Fundamental Right to Research and the failed Text and Data Mining Limitations in the CSDM Directive" (2021) PIJIP/TLS Research Paper Series No 66 <https://digitalcommons.wcl.american.edu/research/66/> accessed 10 February 2025.

Christophe Geiger, and Bulayenko, Oleksandr and Hassler, Théo and Izyumenko, Elena and Schönherr, Franciska and Seuba, Xavier, Reaction of CEIPI to The Resolution on the Implementation of Directive 2001/29/EC on the Harmonisation of Copyright in the Information Society Adopted by the European Parliament on the 9th July 2015 (2015) 37(11) European Intellectual Property Review 683, *Centre for International Intellectual Property Studies (CEIPI) Research Paper* No. 2015-01< https://ssrn.com/abstract=2970507> accessed 10 February 2025.

Christophe Geiger, Giancarlo Frosio and Oleksandr Bulayenko, 'The Exception for Text and Data Mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market - Legal Aspects' (2018) Centre for International Intellectual Property Studies (CEIPI) Research Paper No. 2018-02 <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3160586> accessed 10 February 2025.

Christophe Geiger et al., 'Crafting a Text and Data Mining Exception for Machine Learning and Big Data in the Digital Single Market' in Xavier Seuba, Christophe Geiger and Julien Pénin (eds.), *Intellectual Property and Digital Trade in the Age of Artificial Intelligence and Big Data* (A CEIPI-ICTSD, Issue 5, 2018) 95.

Christophe Geiger, Giancarlo Frosio and Oleksandr Bulayenko, "The EU Commission's Proposal to Reform Copyright Limitations: A Good but Far Too Timid Step in the Right Direction" (2018) 40 (1) EIPR 4.

Christophe Geiger, Daniel Gervais and Martin Senftleben, 'The Three-Step Test Revisited: How to Use the Test's Flexibility in National Copyright Law' (2014) 29 American University International Law Review 581.

Chyan Yang and Hsien-Jyh Liao, 'Using the Robots.txt and Robots Meta Tags to Implement Online Copyright and a Related Amendment' (2010) 28(1) Library Hi Tech 94.

Code of Practice on the EU AI Act: General Purpose AI (2025) <https://code-of-practice.ai/?section=summary#commitment-i-2-copyright-policy> accessed 15 May 2025;

Commission, 'Better Careers and More Mobility: A European Partnership for Researchers' COM (2008) 317 final.

Communia, 'Policy Paper #15 on using copyrighted works for teaching the machine' (*Communia*, 26 April 2023) <https://communia-association.org/policy-paper/policy-paper-15-on-using-copyrighted-works-for-teaching-the-machine/> accessed 10 February 2025.

Claude Masouyé and William Wallace, Guide to the Berne Convention for the Protection of Literary and Artistic Works (Paris Act 1971) (WIPO publication No 615 (E.) 229) <https://doi.org/10.34667/tind.28751> accessed 30 March 2025.

Cullen Miller, 'Ai.txt: A New Way for Websites to Set Permissions for AI' (*Spawning Blog*, 30 May 2023) <https://spawning.substack.com/p/aitxt-a-new-way-for-websites-to-set> accessed 30 March 2025.

Danielle Romain, 'An Update on Web Publisher Controls' (*Google blog*, 28 September 2023) <https://blog.google/technology/ai/an-update-on-web-publisher-controls/> accessed 28 January 2025.

Daniele Panfilo et al. 'A Deep Learning-Based Pipeline for the Generation of Synthetic Tabular Data' (2016) 11 IEEE <https://www.researchgate.net/publication/371828083_A_Deep_Learning-based_Pipeline_for_the_Generation_of_Synthetic_Tabular_Data> accessed 12 April 2025.

Danilo Mandic, 'Balance: Resolving the conundrum between copyright and technology?' (University of Westminster, Working Paper, May 2011)

<https://www.wipo.int/edocs/mdocs/mdocs/en/wipo_ipr_ge_11/wipo_ipr_ge_11_topic2-related2.pdf accessed 12 March 2025> accessed 12 April 2025.

Daniel Mooney, 'Copyright in the Digital Age: Analysing the Achievements and Flaws in the EU Copyright Exceptions Domain' (22 November 2022) Cambridge Journal of Law, Politics and Art 133.

Daniel Gervais, 'Towards a New Core International Copyright Norm: the Reverse Three-step Test' (2005) 9 Marquette Intellectual Property Law Review 1.

Darren Ang, 'The Web Scraper's World of Copyright Exceptions and Contractual Overrides' (2021) 13(22) Singapore Law Review 5.

David N. Schiff, 'Socio-Legal Theory: Social Structure and Law' (1976) 39 (3) The Modern Law Review, 287.

Dirk Voorhoof, 'Chapter 17: Freedom of Expression and the Right to Information: Implications for Copyright' in C Geiger (ed), *Research Handbook on Human Rights and Intellectual Property* (Edward Elgar Publishing 2025) < https://doi.org/10.4337/9781783472420.00030> accessed 23 March 2025.

Ehsan Latif, 'AGI: Artificial General Intelligence for Education' (2024) <https://arxiv.org/pdf/2304.12479.pdf > accessed 23 March 2025.

Eleonora Rosati, *Copyright in the Digital Single Market: Article-by-Article Commentary to the Provisions of Directive 2019/790* (Oxford: Oxford University Press, 2021).

Eleonora Rosati, 'No Step-Free Copyright Exceptions: The Role of the Three-Step Test in Defining Permitted Uses of Protected Content (Including TDM for AI-Training Purposes)' (2024) European Intellectual Property Review 46 (2024), 262.

ERC panels' classification (2020), <https://erc.europa.eu/sites/default/files/document/file/ERC_Panel_structure_2020.pdf> accessed 10 January 2025.

Erroll Wood et al., 'Fake It till You Make It: Face Analysis in the Wild Using Synthetic Data Alone' (2021) <https://arxiv.org/abs/2109.15102 > accessed 10 January 2025.

Estelle Derclaye, *The Legal Protection of Databases: A Comparative Analysis* (Edward Elgar Publishing 2008) 111.

Estelle Derclaye and Martin Husovec, 'Why the sui generis database clause in the Data Act is counterproductive and how to improve it?' (2022) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4052390> accessed 12 May 2025

Estelle Derclaye and others, 'Opinion of the European Copyright Society on selected aspects of the proposed Data Act' (*European Copyright Society*, 12 May 2022) <https://europeancopyrightsociety-dotorg.files.wordpress.com/2022/05/opinion-of-the-ecs-on-selected-as-pects-of-the-data-act-1.pdf > accessed 20 March 2025.

ECS, 'Copyright and Generative AI: Opinion of the European Copyright Society' (2025) <https://europeancopyrightsociety.org/opinions/> accessed 12 May 2025.

European Copyright Society, 'General Opinion on the EU Copyright Reform Package' (2017) <https://europeancopyrightsocietydotorg> accessed 20 February 2025.

European Commission, COM (2007) 182 - Improving knowledge transfer between research institutions and industry across Europe: embracing open innovation – Implementing the Lisbon Agenda, <https://op.europa.eu/en/publication-detail/-/publication/47e7bcfd-6493-4f0c-b4ee-da54fe9e2e86> accessed 13 March 2025.

European Commission, 'Study on Copyright and new Technologies: Copyright data management and artificial intelligence' (February 2022) <https://op.europa.eu/en/publication-detail/-/publication/cc293085-a4da-11ec-83e1-01aa75ed71a1/language-en> accessed 23 April 2025

European Commission, 'Data Act' (2024) <https://digital-strategy.ec.europa.eu/en/policies/data-act> accessed 12 May 2025.

European Commission, 'Study to Support an Impact Assessment for the Review of the Database Directive', Final Report (2022), <https://copenhageneconomics.com/wp-

content/uploads/2022/02/study-to-support-an-impact-assessment-for-the-review-of-the-database-directive.pdf > accessed 23 March 2025.

EU IPO, *The Development of Generative Artificial Intelligence from a Copyright Perspective* (2025) <https://www.europarl.europa.eu/meetdocs/2024_2029/plmrep/COMMITTEES/JURI/DV/2025/05-12/2025.05.12_item6_Study_GenAIfromacopyrightperspective_EN.pdf> accessed 15 May 2025.

Feenan Dermot, 'Exploring the 'Socio' of Socio-Legal Studies, in Dermot Feenan' in D Feenan (ed.), *Exploring the 'Socio' of Socio-Legal Studies* (Palgrave MacMillan 2013), 3-19.

Gareth Kristensen et al. 'Training AI models on Synthetic Data: No Silver Bullet for IP Infringement Risk in the Context of Training AI Systems (Part 1 of 4)' (*Clearly IP and Technology Insights*, 12 December 2023) <https://www.clearyiptechinsights.com/2023/12/training-ai-models-on-synthetic-datanosilverbulletforipinfringement-risk-in-the-context-of-training-ai-systems-part-1-of-4/> accessed 12 March 2025.

Gina Maria Ziaja, 'The Text and Data Mining Opt-out in Article 4(3) CDSMD: Adequate Veto Right for Rightholders or a Suffocating Blanket for European Artificial Intelligence Innovations?' (2024) 19(5) Journal of Intellectual Property Law and Practice 453.

Giulio Coraggio, 'How synthetic data can address IP and privacy issues of artificial intelligence' (*GamingTechLaw*, 4 April 2023) <https://www.gamingtechlaw.com/2023/04/how-synthetic-data-can-address-ip-and-privacy-issues-of-artificial-intelligence/> accessed 12 March 2025.

Giuseppe Colangelo, 'European Proposal for a Data Act-A First Assessment' (20 July 2022) CERRE Evaluation Paper 2022, 1, 6 <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4199565> accessed 10 April 2025.

Google Search Central, 'Introduction to Robots.txt' <https://developers.google.com/search/docs/crawling-indexing/robots/intro> accessed 20 April 2025.

Google Search Central, 'Overview of Google Crawlers and Fetchers (User Agents)' <https://developers.google.com/search/docs/crawling-indexing/overview-google-crawlers> accessed 01 March 2025.

Guido Noto La Diega and Estelle Derclaye, 'Opening Up Big Data for Sustainability: What Role for Database Rights in the Fourth Industrial Revolution?' in Ole-Andreas Rognstad, Taina Pihlajarinne and Jukka Mähönen (eds), *Promoting Sustainable Innovation and the Circular Economy: Legal and Economic Aspects* (Routledge 2022) 28

Han Jiawei, *Data mining: Concepts and Techniques* (3rd edition, Elsevier 2011).

Hanjo Hamann, 'Artificial Intelligence and the Law of Machine-Readability: A Review of Human-to-Machine Communication Protocols and their (In)Compatibility with Article 4(3) of the Copyright DSM Directive' (2024) 15 (2) JIPITEC 102.

Harry Surden, 'Machine Learning and Law: An Overview' in Roland Vogl (ed), *Research Handbook on Big Data Law* (Edward Elgar Publishing 2021).

Inge Graef and Martin Husovec, 'Seven Things to Improve in the Data Act' <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4051793> accessed 12 March 2025.

Irinin Stamatoudi I, 'Text and Data Mining' In: Stamatoudi I (ed) *New Developments in EU and International Copyright Law* (Kluwer Law International 2016).

Jacob Zweig, 'Fair Use as Free Speech Fundamental: How Copyright Law Creates a Conflict between International Intellectual Property and Human Rights Treaties' (2013) 64 Hastings LJ 1549.

James Jordon et al., 'Synthetic Data– What, Why and How?' (*Arxiv*, 2022) < https://arxiv.org/abs/2205.03257 > accessed 23 February 2025.

Jan Bernd Nordemann and Jonathan Pukas, 'Copyright exceptions for AI training data – will there be an international level playing field?' (2022) 17(12) Journal of Intellectual Property Law & Practice 973.

Jan M. Smits, 'What is Legal Doctrine? On the Aims and Methods of Legal Dogmatic Research' (September 1, 2015). Rob van Gestel, Hans-W. Micklitz and Edward L. Rubin (eds.), *Rethinking*

Legal Scholarship: A Transatlantic Dialogue (Cambridge University Press 2017) 207-228, Maastricht European Private Law Institute Working Paper No. 2015/06 <https://ssrn.com/abstract=2644088> accessed 12 January 2025.

Jan M. Smits, *The mind and method of the legal academic* (Edward Elgar Publishing 2012).

Jean-Paul Triaille et al., 'Study on the legal framework of text and data mining (TDM)' (2014) <https://www.fosteropenscience.eu/sites/default/files/pdf/3476.pdf > accessed 12 January 2025.

Jean-Paul Triaille, Depreeuw, Dusollier, Hubin, de Francquen, *Study on the application of Directive 2001/29/EC on copyright and related rights in the information society* (European Commission 150 et seq. 2013).

Jiawei Han, Micheline Kamber and Jian Pei, *Data Mining: Concepts and Techniques* (3rd edn, Elsevier 2012).

Joao Pedro Quintais, 'Generative AI, Copyright and the AI Act'(*Kluwer Copyright Blog*, 9 May 2023) <https://copyrightblog.kluweriplaw.com/2023/05/09/generative-ai-copyright-and-the-ai-act/> accessed 12 January 2025.

Joao Fonseca and Fernando Bacao, 'Tabular and Latent Space Synthetic Data Generation: a Literature Review' (2023) 10(115) Journal of Big Data 1.

Jonas Christoffersen, 'Human Rights and Balancing: The Principle of Proportionality' in Christophe Geiger (ed), *Research Handbook on Human Rights and Intellectual Property* (Edward Elgar Publishing 2015) 19 <https://doi.org/10.4337/9781783472420.00010>accessed 3 February 2025.

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(1) GRUR International 22 < https://doi.org/10.1093/grurint/ikac137>accessed 3 February 2025.

Josef Drexl, Reto Hilty et al., 'Technical Aspects of Artificial Intelligence: An Understanding from an Intellectual Property Law Perspective' (2019) Max Planck Institute for Innovation and Competition Research Paper No. 19–13 <https://ssrn.com/abstract=3465577> accessed 11 January 2025.

Jose Ramon Saura, Domingo Ribeiro-Soriano and Daniel Palacious-Marques, 'From User-generated Data to Data-driven Innovation: A Research Agenda to Understand User Privacy in Digital Markets' (2021) 60 IJIM 1.

Jörg Friedrichs and Friedrich Kratochwil, 'On Acting and Knowing: How Pragmatism Can Advance International Relations Research and Methodology' (2009) 63(4) International Organisation 701 < https://doi.org/10.1017/S0020818309990142>accessed 3 February 2025.

Julia Reda, Joschka Selinger and Michael Servatius, 'Article 17 of the Directive on Copyright in the Digital Single Market: A Fundamental Rights Assessment' (2020) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3732223>accessed 3 February 2025.

Katherine Lee et al., 'Talkin' 'Bout AI Generation: Copyright and the Generative AI Supply Chain' Journal of the Copyright Society of the U.S.A. (forthcoming 2024) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4523551> accessed 12 February 2024.

Khaled El Emam, 'Accelerating AI with Synthetic Data Generating Data for AI Projects' (*NVIDIA, O'Reilly Media*, 2020) <https://www.nvidia.com/content/dam/en-zz/Solutions/deep-learning/resources/accelerating-ai-with-synthetic-data-ebook/accelerating-ai-with-synthetic-data-nvidia_web.pdf> accessed 3 February 2025.

Ken Yale, *Handbook of Statistical Analysis and Data Mining Applications* (Elsevier 2017).

Krisjanis Buss, 'Copyright and Free Speech: The Human Rights Perspective' (2015) 8(2) Baltic Journal of Law and Politics 182.

LACA, 'The Right to Read is the Right to Mine: But Not When Blocked by Technical Protection

Measures' (1 August 2019) <https://libereurope.eu/article/the-right-to-read-is-the-right-to-mine-but-not-when-blocked-by-technical-protection-measures/#:~:text=Matters%20Working%20Group-,The%20Right%20to%20Read%20is%20the%20Right%20To%20Mine%3A%20But,Blocked%

20by%20Technical%20Protection%20Measures&text=Tell%20us%20and%20the%20Libraries,the%20context%20of%20data%20mining > accessed 22 April 2025.

Leander Nielbock and Teresa Nobre, 'The AI Act and the Quest for Transparency' (*Communia*, 28 June 2023) <https://communia-association.org/2023/06/28/the-ai-act-and-the-quest-for-transparency/> accessed 2 March 2025.

Liana Colonna, 'Opportunities and Challenges to Utilising Text-data Mining in Public Libraries: a Need for Legal Research' (2018) 50 Years of Law and IT.

LIBER, 'Busting TDM Myths and Misunderstandings' (2017) <https://libereurope.eu/article/busting-tdm-myths-misunderstandings/> accessed 12 April 2025.

Loreto Corredoira and Rodrigo Cetina Presuel, 'Current Copyright Policy Tendencies in 2015: Further Weakening of Limits and Exceptions and the Ever-Reducing Public Domain' **(**2015) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2642088> accessed 12 March 2025

M Carl Drott, 'Indexing Aids at Corporate Websites: the Use of Robots.txt and META Tags' (2002) 38 Information Processing and Management 209.

Marco Caspers and Lucie Guibault, 'A Right to 'Read' for Machines: Assessing a Black-Box Analysis Exception for Data Mining' (2016) 53 (1) Proceedings of the Association for Information Science and Technology 1.

Marco Caspers and Lucie Guibault, 'A Baseline Report of Policies and Barriers of TDM in Europe' (2016) <https://project.futuretdm.eu/wp-content/uploads/2017/05/FutureTDM_D3.3-Baseline-Report-of-Policies-and-Barriers-of-TDM-in-Europe.pdf> accessed 12 May 2025.

Marianna Foerg, 'Second and Third Drafts of the general-purpose AI Code of Practice Have Been Released' (*Kluwer Copyright Blog*, April 4 2025) <https://copyrightblog.kluweriplaw.com/2025/04/04/second-and-third-drafts-of-the-general-purpose-ai-code-of-practice-have-been-released/> accessed 18 May 2025.

Martin Senftleben, Thomas Margoni et al., '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 (1) JIPITEC 67.

Martin Senftleben,'Generative AI and Author Remuneration' (2023) 54 IIC 1535.

Martin Senftleben, 'TDM, GenAI and the Copyright Three-Step Test' (July 24, 2025) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5373903 > accessed 23 August 2025.

Martijn Koster et al., 'RFC 9309 Robots Exclusion Protocol' (RFC, September 2022) <https://www.rfc-editor.org/rfc/rfc9309.html#KiB> accessed 12 January 2025.

Maryna Manteghi, 'In Search of Balance: Text, Data Mining and Copyright in the Digital Single Market Directive from a Fundamental Rights Perspective' (2023) 48 European Law Review 422.

Maryna Manteghi, 'Overcoming Barriers to Text and Data Mining in the Era of ChatGPT: The Proposed Data Act as a Game-Changer' (2024) 73(1) GRUR International 34.

Maryna Manteghi, 'Can text and data mining exceptions and synthetic data training mitigate copyright-related concerns in generative AI?' Law, Innovation and Technology, (2024), 16(2), 663 https://doi.org/10.1080/17579961.2024.2392928.

Maryna Manteghi, 'Kneschke vs. LAION: Opening Fresh Perspectives on AI Training and TDM Exceptions' (*European Law Blog*, January 2025) https://doi.org/10.21428/9885764c.5f1ce2dd.

Maryna Manteghi, 'Will the AI Act Bring More Clarity to the Regulation of Text and Data Mining in the EU?' (*EU Law Analysis*, 18 May 2024) <https://eulawanalysis.blogspot.com/2024/05/will-ai-act-bring-more-clarity-to.html> accessed 11 March 2025.

Matthias Leistner and Lucie Antoine, 'Attention, Here Comes the EU Data Act! A Critical In-depth Analysis of the Commission's 2022 Proposal' (2022) 13 JIPITEC 339.

Matthew Sag, 'The New Legal Landscape for Text Mining and Machine Learning ' (2019) 66 Journal of the Copyright Society of the USA 3.

Matthew Sag, 'Fairness and Fair Use in Generative AI' (2024) 92 (5) Fordham Law Review 1887.

Mauritz Kop, 'The Right to Process Data for Machine Learning Purposes in the EU' (2021) 34 Harvard Journal of Law and Technology 1.

Michael D. Birnhack, 'Acknowledging the Conflict between Copyright Law and Freedom of Expression under the Human Rights Act' (2003) 24 Entertainment Law Review <https://ssrn.com/abstract=368961> accessed 11 February 2025.

Michel Walter and Silke von Lewinski, *European Copyright Law: A Commentary* (Oxford University Press 2010).

Mihaly Ficsor, *Guide to the Copyright and Related Rights Treaties Administered by WIPO and Glossary of Copyright and Related Rights Terms* (WIPO publication 891, 2003) 317.

Mira T. Sundara Rajan, 'Is Generative AI Fair Use of Copyright Works? NYT v. OpenAI' (*Kluwer Copyright Blog*, 29 February 2024) <https://copyrightblog.kluweriplaw.com/2024/02/29/is-generative-ai-fair-use-of-copyright-works-nyt-v-openai/> accessed 2 March 2025.

Mostly AI, 'What is Synthetic Data?' <https://mostly.ai/syntheticdata/whatissyntheticdata#:~:text=Synthetic%20data%20is%20generated%20by,create%20statistically%20identical%2C%20synthetic%20data> accessed 14 February 2025

Nicola Lucchi, 'ChatGPT: A Case Study on Copyright Challenges for Generative Artificial Intelligence Systems' (2023) European Journal of Risk Regulation <https://www.cambridge.org/core/services/aop-cambridge-core/content/view/CEDCE34DED599CC4EB201289BB161965/S1867299X23000594a.pdf/chatgpt-a-case-study-on-copyright-challenges-for-generative-artificial-intelligence-systems.pdf> accessed 11 February 2025.

Nicola Lucchi, '*Generative AI and Copyright: Training, Creation, Regulation*' (Policy Department for Justice, Civil Liberties and Institutional Affairs, Directorate-General for Citizens' Rights, Justice and Institutional Affairs, European Parliament, PE 774.095, July 2025).

Nicolas Jondet, 'The text and data mining exception in the proposal for a directive on copyright: why the European Union needs to go further than the laws of member states' (2018) 67 Propriétés Intellectuelles<https://www.researchgate.net/publication/327505385_The_text_and_data_mining_exception_in_the_proposal_for_a_directive_on_copyright_why_the_European_Union_needs_to_go_further_than_the_laws_of_member_states> accessed 13 February 2025.

OddyTech, 'About /robots.txt' <http://www.robotstxt.org/robotstxt.html> accessed 12 March 2025.

Oreste Pollicino et al., 'Cultural Rights, Cultural Diversity and the EU's Copyright Regime: the Battlefield of Exceptions and Limitations to Protected Content' in Oreste Pollicino (ed.) *Copyright and Fundamental Rights in the Digital Age* (Edward Elgar Publishing Limited, 2020) 124

OECD, Working Party of National Experts on Science and Technology Indicators, 'Revised field of science and technology (FOS) classification in the Frascati Manual' (2007) <https://www.oecd.org/science/inno/38235147.pdf > accessed 13 February 2025.

Pamela Samuelson, 'Fair Use Defences in Disruptive Technology Cases,' (2024) 71 UCLA Law Review 1484.

Pascale Chapdelaine, 'Copyright User Rights and Remedies: An Access to Justice Perspective' (2018) 7 LAWS 1-26 <https://ssrn.com/abstract=3204902> accessed 13 February 2025.

Paul Keller, 'Generative AI and Copyright: Convergence of Opt-Outs?' (*Kluwer Copyright Blog*, 23 November 2023) <https://copyrightblog.kluweriplaw.com/2023/11/23/generative-ai-and-copyright-convergence-of-opt-outs/ > accessed 23 March 2025.

Paul Keller and Zuzanna Warso, 'Defining Best Practices for Opting Out of ML Training' (*Open Future*, 28 September 2023) <https://openfuture.eu/publication/defining-best-practices-for-opting-out-of-ml-training/> accessed 13 February 2025.

Paulien Wymeersch, 'EU Copyright Exceptions and Limitations and the Three-Step Test: One Step Forward, Two Steps Back' (2023) 72(7) GRUR International 631.

Peter Henderson et al., 'Foundation Models and Fair Use' (2023) arXiv <https://arxiv.org/pdf/2303.15715.pdf > accessed 13 February 2025.

Peter Lee, 'Synthetic Data and the Future of AI' (2024) 110 Cornell Law Review (forthcoming) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4722162 > accessed 12 March 2025.
Péter Mezei, 'A Saviour or a Dead End? Reservation of Rights in the Age of Generative AI' (2024) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4695119> accessed 20 March 2025.
Peter Mezei, 'Third European Court Decision on the General Purpose TDM Exception is Out' (*Kluwer Copyright Blog*, May 8, 2025) <https://copyrightblog.kluweriplaw.com/2025/05/08/third-european-court-decision-on-the-general-purpose-tdm-exception-is-out/ > accessed 18 May 2025.
Péter Mezei, 'The Multi-layered Regulation of Rights Reservation (Opt-out) Under EU Copyright Law and the AI Act -For the Benefit of Whom?' (v3.0) (March 31, 2025) <https://ssrn.com/abstract=5064018> accessed 23 April 2025.
Peter Murray-Rust, 'The Right to Read Is the Right to Mine" (*Open Knowledge*, June 1 2012) < https://blog.okfn.org/2012/06/01/the-right-to-read-is-the-right-to-mine/ > accessed 23 April 2025.
Peter Teunissen, 'The Balance Puzzle. The ECJ's Method of Proportionality Review in EU Copyright Law' (2018) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5020903> accessed 23 April 2025.
REBIUN, 'Text and Data Mining: (Articles 3 and 4 of the EU-DSM) by REBIUN's Copyright Working Group' (*Ifla.org* 10 June 2020) <https://blogs.ifla.org/lpa/2020/06/10/text-and-data-mining-articles-3-and-4-of-the-eu-dsm/> accessed 23 April 2025.
Reto Hilty and Heiko Richter, 'Position Statement of the Max Planck Institute for Innovation and Competition on the Proposed Modernisation of European Copyright Rules Part B Exceptions and Limitations (Art. 3 Text and Data Mining)' (2017) Max Planck Institute for Innovation and Competition Research Paper No.17-02, 11 <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2900110> accessed 23 April 2025.
Resa Banakar and Max Travers, *Theory and Method in Socio-Legal Research* (Hart Publishing 2005).
Richard Bendis and Ethan Byler, 'Creating a National Innovation Framework' (2009) Science Progress <https://www.innovationamerica.us/images/stories/pdf/bendis_innovation.pdf> accessed 12 April 2025.
Robert Alexy, 'Constitutional Rights, Balancing, and Rationality' (2003) 16(2) Ratio Juris 131.
Roberta De Michele and Marco Furini, 'IoT Healthcare: Benefits, Issues and Challenges' (*GoodTechs '19: EAI International Conference on Smart Objects and Technologies for Social Good*, Valencia, September 2019).
Rosa Hartmut, '*Social Acceleration: A New Theory of Modernity*' (Columbia University Press 2013).
Rosa Hartmut, 'Airports Built on Shifting Grounds? Social Acceleration and the Temporal Dimension of Law' in Luigi Corrias and Lyana Francor (eds), *Temporal Boundaries of Law and Politics: Time Out of Joint* (Routledge 2018) 73–87.
Rossana Ducato and Alain Strowel, 'Limitations to Text and Data Mining and Consumer Empowerment: Making the Case for a Right to 'Machine Legibility' (2019) 50 IIC. 649.
Rossana Ducato and Alain Strowel, 'Ensuring Text and Data Mining: Remaining Issues with the EU Copyright Exceptions and Possible Ways Out' (2021) 43 EIPR. 20.
Sachin Kumar, Prayag Tiwari and Mikhail Zymbler, 'Internet of Things is a Revolutionary Approach for Future Technology Enhancement: A Review' (2019) 6 Journal of Big Data 1.
Shabbir Merali and Ali Merali, 'The Generative AI Revolution Opportunities, Shocks, and Risks' (2023) <https://www.ukonward.com/wp-content/uploads/2023/05/Generative-AI-Revolution-Final.pdf> accessed 1 March 2025.
Sam Ricketson and Jane C. Ginsburg, *International Copyright and Neighbouring Rights: The Berne Convention and Beyond* (OUP 2022) 407.
Sara Blandy and Susan Bright, *Researching Property Law* (Bloomsbury Publishing 2015).
Sean Flynn et al., 'Implementing User Rights for Research in the Field of Artificial Intelligence: A Call for International Action' (2020) 42(7) European Intellectual Property Review 393.
Sean Flynn and Lokesh Vyas, 'Examples of Text and Data Mining Research Using Copyrighted Materials' (Kluwer Copyright Blog, 6 March 2023)

<https://copyrightblog.kluweriplaw.com/2023/03/06/examples-of-text-and-data-mining-research-using-copyrighted-materi-als/ > accessed 20 March 2025.

Severine Dusollier, 'The 2019 Directive on Copyright in the Digital Single Market: Some Progress, a Few Bad Choices, and an Overall Failed Ambition' (2020) 57 (4) Common Market Law Review 979.

Severine Dusollier, 'The limitations and exceptions to copyright and related rights for libraries, research and teaching uses' in Study on the Application of Directive 2001/29/EC on Copyright and Related Rights in the Information Society (the InfoSoc Directive) (European Union 2013).

Sina Alemohammad et al., 'Self-Consuming Generative Models Go MAD' (2023) <https://arxiv.org/abs/2307.01850 > accessed 13 February 2025.

Sooho Lee and Barry Bozeman, 'The Impact of Research Collaboration on Scientific Productivity' (2005) 35 (5) Social Studies of Science 673–702 <https://doi.org/10.1177/0306312705052359> accessed 13 February 2025.

Steve Peers et al. (eds), *The EU Charter of Fundamental Rights: A Commentary* (Hart Publishing, 2021).

Sunimal Mendis, *Copyright, the Freedom of Expression and the Right to Information: The persisting Discord,* vol 8 (Nomos, Baden-Baden 2011).

Susan Reilly, *Libraries at the Centre of the Debate on Copyright and Text and Data Mining: the LIBER Experience*. Paper presented at: IFLA WLIC 2014 - Lyon - Libraries, Citizens, Societies: Confluence for Knowledge in Session 119 - Academic and Research Libraries with Serials and Other Continuing Resources and Committee on Copyright and other Legal Matters (CLM). In: IFLA WLIC 2014, 16-22 August 2014, Lyon, France <https://library.ifla.org/id/eprint/1007/1/119-reilly-en.pdf > accessed 10 January 2025.

Tatiana Eleni Synodinou, 'Lawfulness for Users in European Copyright Law: Acquis and Perspectives' (2019) 10 JIPITEC 20.

Tatiana Eleni Synodinou, 'Intermediaries' Liability for Copyright Infringement in the EU: Evolutions and Confusions (2015) 31 Computer Law & Security Review 57.

The Parliament's Constitutional Law Committee statement PeVL 58/2022 vp HE 43/2022 vp (14 November 2022), <https://www.eduskunta.fi/FI/vaski/Lausunto/Sivut/PeVL_58+2022.aspx> accessed 13 February 2025.

The ERC panels' classification (2020) <https://erc.europa.eu/sites/default/files/document/file/ERC_Panel_structure_2020.pdf> accessed 13 February 2025.

The Finnish Parliament's Constitutional Law Committee, Statement PeVL 58/2022 vp HE 43/2022 vp (14 November 2022), <https://www.eduskunta.fi/FI/vaski/Lausunto/Sivut/PeVL_58+2022.aspx > accessed 13 February 2025.

Terry Hutchinson and Nigel Duncan, 'Defining and Describing What We Do: Doctrinal Legal Research' (2012) 17 Deakin Law Review.

Terry Hutchinson, 'Doctrinal Research: Researching the Jury in Research Methods in Law' in Mandy Burton and Dawn Watkins (eds.), *Research Methods in Law* (Routledge, 2017).

Thomas Göbel et al., Data for Digital Forensics: Why a Discussion on How Realistic is Synthetic Data' is Dispensable' (2023) 4(3) Digital Threats Research and Practice 1.

Thomas Margoni and Martin Kretschmer, 'The text and data exception in the proposal for a DSM: Why it is Not What EU Copyright Law Needs' CREATe Working paper 2018.

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 (8) GRUR International 685.

Thomas Margoni and Giulia Dore, 'Why We Need a Text and Data Mining Exception (But it is Not Enough)' Working paper, 2016 <https://zenodo.org/record/248048#.W2iL3FM2vIU> accessed 24 July 2024.

Thomas Margoni, 'Text and Data Mining in Intellectual Property Law: Towards an Autonomous Classification of Computational Legal Methods' CREATe Working Paper 01/2020.

Thomas Margoni, 'To TDM or Not to TDM?' (OpenMinTeD, Brussels, 2018) <http://openminted.eu/wpcontent/uploads/2018/06/ThomasMargoniOMTD_BrussMAy2018.pdf > accessed 1 May 2025.

Tshilidzi Marwala et al., 'The Use of Synthetic Data to Train AI Models: Opportunities and Risks for Sustainable Development' (2023) <https://www.semanticscholar.org/reader/5c593e30366748a764190e4d80533c6e1aaef865> accessed 12 March 2025.

Tuomas Mylly, 'Proportionality in the CJEU's Internet Copyright Case Law: Invasive or Resilient?' in Ulf Bernitz, Xavier Groussot, Jaan Paju, and Sybe de Vries (eds), *General Principles of EU Law and the EU Digital Order* (Wolters Kluwer 2020).

Tuomas Mylly, 'The New Constitutional Architecture of Intellectual Property' in Jonathan Griffiths and Tuomas Mylly (eds), *Global Intellectual Property Protection and New Constitutionalism: Hedging Exclusive Rights* (Oxford University Press 2021).

Tuomas Mylly, 'Regulating With Rights Proportionality? Copyright, Fundamental Rights and Internet in the Case Law of the Court of Justice of the European Union' in Oreste Pollicino et al (eds), *Copyright and Fundamental Rights in the Digital Age: A Comparative Analysis in Search of a Common Constitutional Ground* (Edward Elgar Publishing 2020) 54.

Tuomas Mylly, 'Intellectual Property and European Economic Constitutional Law: The Trouble with Private Informational Power' (PhD thesis, University of Turku 2009) <https://www.finna.fi/Record/uef.997351523705966> accessed 10 January 2025.

Van Lindberg, 'Building and Using Generative Models Under US Copyright Law' (2023) 18(2) Rutgers Business Law Review 1.

Varga Csaba, 'Legal Dogmatic: Methodology and Ontology' (2011) 11-12 Law of Ukraine: Legal Journal.

William Cornish, *Intellectual Property: Patents, Copyrights, Trademarks &Allied Rights* (London: Sweet & Maxwell 1999).

Yin Harn Lee et al., 'Copyright and Freedom of Expression: A Literature Review' (2015) CREATe Working Paper 2015/04.

Yingzhou Lu et al., 'Machine Learning for Synthetic Data Generation: A Review' (2021) 14(8) Journal of Latex Class Files 1.

Yuxi Ma et al., 'Brain in a Vat: On Missing Pieces Towards Artificial General Intelligence in Large Language Models' (2023) <https://arxiv.org/abs/2307.03762> accessed 12 February 2025.

Law

Directives

Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC [2019] *OJ L 130/92*.

Directive (EU) 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society [2001] *OJ L167*/10.

Directive (EU) 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases [1996] *OJ L 77/20*.

Directive (EU) 2009/24/EC of the European Parliament and of the Council of 23 April 2009 on the legal protection of computer programs (Codified version) (Text with EEA relevance) [2009] *OJ L 111/16*.

Directive (EU) 2019/1937/EC of the European Parliament and of the Council on the protection of persons who report breaches of Union law (Whistleblower Protection Directive) [2019] *OJ L 305/17*.

Directive (EU) 2020/1828/EC of the European Parliament and of the Council on representative actions for the protection of the collective interests of consumers [2020] *OJ L 409/1*.

Directive (EU) 2016/801/EC of the European Parliament and of the Council of 11 May 2016 on the conditions of entry and residence of third-country nationals for the purposes of research, studies, training, voluntary service, pupil exchange schemes or educational projects and au pairing (recast) [2016] *OJ L 132/21*.

Others

European Commission, ‘Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions: Reviewing Community Innovation Policy in a Changing World’ COM (2009) 442 final.

European Commission, ‘Communication to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions: A European Strategy for Data’ COM (2020) 66 final.

European Commission, ‘Proposal for a Directive of the European Parliament and of the Council on Copyright in the Digital Single Market’ COM (2016) 593 final, 14 September 2016.

European Commission, ‘Proposal for a Regulation of the European Parliament and of the Council on Harmonised Rules on Fair Access to and Use of Data (Data Act)’ COM (2022) 68 final.

European Commission, ‘Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts’ COM (2021) 206 final.

Regulations

Regulation (EU) 2023/2854 of the European Parliament and of the Council of 13 December 2023 on harmonised rules on fair access to and use of data and amending Regulation (EU) 2017/2394 and Directive (EU) 2020/1828 (Data Act) [2023] *OJ L 2023/2854*.

Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European data governance and amending Regulation (EU) 2018/1724 (Data Governance Act) [2022] *OJ L 152/1*.

Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act) [2022] *OJ L 277/1*.

Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on contestable and fair markets in the digital sector (Digital Markets Act) [2022] *OJ L 265/1*.

Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024, laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) [2024] OJ L 2024/1689, 12.7.2024.

National Law

German Copyright Act (*Urheberrechtsgesetz*, UrhG), Act on Copyright and Related Rights of 9 September 1965, BGBl. I 1965, 1273, as last amended by the Act on the Adaptation of Copyright Law to the Requirements of the Digital Single Market, BGBl. I 2021, 1204.

Federal Constitution of the Swiss Confederation of 18 April 1999 (SR 101), as amended to 1 January 2024.

Declaration of the Rights of Man and of the Citizen (*Déclaration des droits de l'homme et du citoyen*), adopted 26 August 1789.

Gesetz zur Anpassung des Urheberrechts an die Erfordernisse des digitalen Binnenmarktes (Germany), 31 May 2021, BGBl. I 2021, 1204 (entered into force 7 June 2021).

Decreto Legislativo 8 novembre 2021, n. 177, Art. 70-ter, in *Attuazione della direttiva (UE) 2019/790*, Gazzetta Ufficiale 27 novembre 2021, n. 283 (Italia).

Auteurswet (Dutch Copyright Act), *Wet van 23 september 1912, houdende nieuwe regeling van het auteursrecht* (Stb. 1912, 308), as last amended by *Wet van 29 juni 2021 tot implementatie van Richtlijn (EU) 2019/790* (Stb. 2021, 305)

Conventions and Treaties

Charter of Fundamental Rights of the European Union [2012] OJ C 326/391.

Convention for the Protection of Human Rights and Fundamental Freedoms (*European Convention on Human Rights*, ECHR), adopted 4 November 1950, entered into force 3 September 1953, ETS No 5.

Universal Declaration of Human Rights, adopted 10 December 1948, UNGA Res 217 A (III).

International Covenant on Civil and Political Rights, adopted 16 December 1966, entered into force 23 March 1976, 999 UNTS 171.

Berne Convention for the Protection of Literary and Artistic Works, adopted 9 September 1886, as revised at Paris on 24 July 1971.

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

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

WIPO Marrakesh Treaty to Facilitate Access to Published Works for Persons Who Are Blind, Visually Impaired or Otherwise Print Disabled, adopted 27 June 2013, entered into force 30 September 2016.

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

WIPO Performances and Phonograms Treaty (WPPT), adopted 20 December 1996, entered into force 20 May 2002.

Agreement on Trade-Related Aspects of Intellectual Property Rights (*TRIPS Agreement*), Annexe 1C to the Marrakesh Agreement Establishing the World Trade Organisation, signed 15 April 1994, entered into force 1 January 1995.

Case law

ECtHR

Ahmet Yıldırım v Turkey, App. No. 3111/10, Judgment of 18 March 2013, ECtHR 2013.

Dammann v Switzerland, App. No. 77551/01, Judgment of 25 July 2006, ECtHR 2006.

Delfi AS v Estonia, App. No. 64569/09, Judgment (Grand Chamber) of 16 June 2015, ECtHR 2015.

Dima v Romania, App. No. 58472/00, Decision of 26 May 2005 (unreported), and Judgment of 16 November 2006 (unreported), ECtHR 2006.

EIT Nordisk Film & TV A/S v Denmark, App. No. 40485/02, Judgment of 8 December 2005, ECtHR 2005.
Fredrik Neij and Peter Sunde Kolmisoppi (The Pirate Bay) v Sweden, App. No. 40397/12, Judgment of 19 February 2013, ECtHR 2013.
Kablis v Russia, App. Nos. 48310/16 and 59663/17, Judgment of 9 September 2019, ECtHR 2019.
Kenedi v Hungary, App. No. 31475/05, Judgment of 26 May 2009 (unreported), ECtHR 2009.
Kharitonov v Russia, App. No. 10795/14, Judgment of 16 November 2020, ECtHR 2020.
Melnychuk v Ukraine (dec.), App. No. 28743/03, Decision of 5 July 2005, ECHR 2005-IX, ECtHR 2005.
Observer and Guardian v the United Kingdom, App. No. 13585/88, Judgment of 26 November 1991, ECtHR 1991.
Társaság a Szabadságjogokért v Hungary, App. No. 37374/05, Judgment of 14 April 2009 (unreported), ECtHR 2009.
Timpul Info-Magazin and Anghel v Moldova, App. No. 42864/05, Judgment of 27 November 2007, ECtHR 2007.
Węgrzynowski and Smolczewski v Poland, App. No. 33846/07, Judgment of 16 July 2013, ECtHR 2013.
Youth Initiative for Human Rights v Serbia, App. No. 48135/06, Judgment of 25 June 2013, paras 20 and 24 (unreported), ECtHR 2013.
Société nationale de programme France 2 v France, App. No. 30262/96, Judgment of 15 January 1997 (unreported), ECtHR 1997.
Oguz Aral, Galip Tekin and Inci Aral v Turkey (dec.), App. No. [number unavailable], Decision (date unavailable), ECtHR 1995.

CJEU

Infopaq International A/S v Danske Dagblades Forening, C-5/08, 16 July 2009, ECLI:EU:C:2009:465.
British Horseracing Board Ltd v William Hill Organization Ltd, C-203/02, 9 November 2004, ECLI:EU:C:2004:695; [2005] 1 C.M.L.R. 15.
Promusicae v Telefónica de España SAU, C-275/06, 29 January 2008, ECLI:EU:C:2008:54.
Football Association Premier League Ltd v QC Leisure (joined Cases C-403/08 and C-429/08), 4 October 2011, ECLI:EU:C:2011:631; [2012] 1 C.M.L.R. 29.
Padawan SL v Sociedad General de Autores y Editores de España (SGAE), C-467/08, 21 October 2010, ECLI:EU:C:2010:620.
Société belge des auteurs, compositeurs et éditeurs (SABAM) v Netlog NV, C-360/10, 16 February 2012, ECLI:EU:C:2012:85.
Sky Österreich GmbH v Österreichischer Rundfunk, C-283/11, 22 January 2013, ECLI:EU:C:2013:28.
UPC Telekabel Wien GmbH v Constantin Film Verleih GmbH, C-314/12, 27 March 2014, ECLI:EU:C:2014:192.
ACI Adam BV v Stichting de Thuiskopie, C-435/12, 10 April 2014, ECLI:EU:C:2014:254; [2014] E.C.D.R. 21.
Public Relations Consultants Association Ltd v Newspaper Licensing Agency Ltd, C-360/13, 5 June 2014, ECLI:EU:C:2014:1195.
Eva-Maria Painer v Standard VerlagsGmbH and Others, C-145/10, 1 December 2011, ECLI:EU:C:2011:798.
Deckmyn and Vrijheidsfonds VZW v Vandersteen and Others, C-201/13, 3 September 2014, ECLI:EU:C:2014:2132.
Scarlet Extended SA v Société belge des auteurs, compositeurs et éditeurs (SABAM), C-70/10, 24 November 2011, ECLI:EU:C:2011:771.
GS Media BV v Sanoma Media Netherlands BV and Others, C-160/15, 8 September 2016, ECLI:EU:C:2016:644.
McFadden v Sony Music Entertainment Germany GmbH, C-484/14, 15 September 2016, ECLI:EU:C:2016:689.

Stichting Brein v Wullems (t/a Filmspeler), C-527/15, 26 April 2017, ECLI:EU:C:2017:300.
Opinion of AG Campos Sánchez-Bordona in Land Nordrhein-Westfalen v Renckhoff, C-161/17, 25 April 2018, ECLI:EU:C:2018:279.
Opinion of AG Szpunar in Pelham GmbH v Hutter, C-476/17, 12 December 2018, ECLI:EU:C:2019:624.
VG Bild-Kunst v Stiftung Preußischer Kulturbesitz, C-392/19, 9 March 2021, ECLI:EU:C:2021:181.
Poland v European Parliament and Council of the European Union, C-401/19, 26 April 2022, ECLI:EU:C:2022:297; [2023] 4 W.L.R. 24.
Data Protection Commissioner v Facebook Ireland Ltd and Maximillian Schrems, C-311/18, 16 July 2020, ECLI:EU:C:2020:559.
Stichting Greenpeace Nederland and Pesticide Action Network Europe (PAN Europe) v European Commission, T-545/11, 8 October 2013, ECLI:EU:T:2013:523.
Brompton Bicycle Ltd v Chedech/Get2Get C-833/18, 11 June 2020, ECLI:EU:C:2020:461.
BSA – Business Software Alliance v Ministerstvo kultury, C-393/09, Judgment of 22 December 2010, ECLI:EU:C:2010:816.
Cofemel – Sociedade de Vestuário SA v G-Star Raw CV, C-683/17, Judgment of 12 September 2019, ECLI:EU:C:2019:721.

National Case Law

Kneschke v LAION, Landgericht München I (District Court of Munich I), Judgment of 27 September 2024, Case No. 23 O 00061/23.
DPG Media et al v HowardsHome, Rechtbank Amsterdam (Amsterdam District Court), Judgment of 30 October 2024, ECLI:NL:RBAMS:2024:6563.
Gamekapocs, Municipal Court of Appeals (Hungary), Case No. 9 Pf. 20.353/2024/6-II, Judgment of 3 December 2024

US Case Law

Authors Guild v Google Inc., 05 Civ. 8136 (S.D.N.Y.), Judgment of 14 November 2013 (Google Books).
Authors Guild v HathiTrust, 902 F. Supp. 2d 445, 104 U.S.P.Q.2d 1659 (S.D.N.Y. 2012).
Daily News v Microsoft and OpenAI (consolidated with *The New York Times v OpenAI/Microsoft* and *Center for Investigative Reporting (CIR) v OpenAI/Microsoft*), U.S. District Court for the Southern District of New York, pending.
The Intercept Media v OpenAI, U.S. District Court for the Southern District of New York, pending.
Authors Guild v OpenAI Inc., Case No. 1:23-cv-08292 (S.D.N.Y.), pending.
The New York Times Company v Microsoft Corporation and OpenAI LP, Case No. 1:23-cv-11195 (S.D.N.Y.), pending.
Sarah Andersen et al v Stability AI Ltd et al, Case No. 3:23-cv-00201-WHO (N.D. Cal.), filed 2023.
Richard Kadrey et al v Meta Platforms Inc., Case No. 3:23-cv-03417-VC (N.D. Cal.), Judgment of 20 November 2023.
P.M. v GitHub Copilot Inc. et al, Case No. 4:22-cv-06823 (N.D. Cal.), Decision of 28 June 2023.
Paul Tremblay and Mona Awad v OpenAI LP et al, Case No. 3:23-cv-03223 (N.D. Cal.), filed 5 July 2023.
Sarah Silverman v Meta Platforms Inc. and OpenAI LP et al, Case No. 3:23-cv-03416 (N.D. Cal.), filed 7 July 2023.
J.L. v Alphabet Inc. (Google Bard), Case No. 3:23-cv-03440 (N.D. Cal.), pending.

UK Case Law

Getty Images (US) Inc. and Others v Stability AI Ltd, [2023] EWHC 309 (Ch).

TURUN
YLIOPISTO
UNIVERSITY
OF TURKU

Painosalama, Turku, Finland 2025

ISBN 978-952-02-0424-2 (PRINT)
ISBN 978-952-02-0425-9 (PDF)
ISSN 0082-6987 (Print)
ISSN 2343-3191 (Online)