Stanford – Vienna
Transatlantic Technology Law Forum

A joint initiative of
Stanford Law School and the University of Vienna School of Law

European Union Law Working Papers

No. 124

Legality and Challenges of Web Scraping Databases in the European Union: Focus on Non-Personal Data and Copyright

Mariia Boiko

2026

European Union Law Working Papers

Editors: Siegfried Fina and Roland Vogl

About the European Union Law Working Papers

The European Union Law Working Paper Series presents research on the law and policy of the European Union. The objective of the European Union Law Working Paper Series is to share "works in progress". The authors of the papers are solely responsible for the content of their contributions and may use the citation standards of their home country. The working papers can be found at http://ttlf.stanford.edu.

The European Union Law Working Paper Series is a joint initiative of Stanford Law School and the University of Vienna School of Law's LLM Program in European and International Business Law.

If you should have any questions regarding the European Union Law Working Paper Series, please contact Professor Dr. Siegfried Fina, Jean Monnet Professor of European Union Law, or Dr. Roland Vogl, Executive Director of the Stanford Program in Law, Science and Technology, at:

Stanford-Vienna Transatlantic Technology Law Forum
http://ttlf.stanford.edu

Stanford Law School
Crown Quadrangle
559 Nathan Abbott Way
Stanford, CA 94305-8610

University of Vienna School of Law
Department of Business Law
Schottenbastei 10-16
1010 Vienna, Austria

About the Author

Mariia Boiko is a graduate of the University of Vienna School of Law (Austria) and The Kyiv University of Market Relations (Ukraine).

General Note about the Content

The opinions expressed in this student paper are those of the author and not necessarily those of the Transatlantic Technology Law Forum, or any of TTLF's partner institutions, or the other sponsors of this research project.

Suggested Citation

This European Union Law Working Paper should be cited as:
Mariia Boiko, Legality and Challenges of Web Scraping Databases in the European Union: Focus on Non-Personal Data and Copyright, Stanford-Vienna European Union Law Working Paper No. 124, http://ttlf.stanford.edu.

Copyright

Abstract

This thesis examines how copyright and sui generis database rights currently protect database owners against the use of their data for web scraping and data mining, while also considering the legitimate interests of data miners and AI model developers. It analyzes the relevant EU legal framework, including the Collective Rights Management Directive, the Database Directive, the DSM Directive, the InfoSoc Directive, and the AI Act. Particular attention is paid to collective management mechanisms and opt-out solutions, assessing how they function in practice. The study evaluates the advantages and limitations of these models from the perspective of rights-holders, data miners and AI model providers, and draws conclusions on how effectively each approach balances these competing interests.

Table of Contents

Introduction

Relevance of the research

Over the past decade, the progressive digital transformation of economic and social life has made data one of the core productive resources in the European Union (**EU**). As markets and public services become increasingly data-driven, questions surrounding the scope and effectiveness of intellectual property protection have gained renewed urgency. Within this wider regulatory context, the protection of databases under copyright and *sui generis* rights has a particularly sensitive role. This is mainly due to the fact that modern artificial intelligence (**AI**) technologies, including large language models and other machine-learning systems, depend on access to immense amounts of data, which is most often collected, structured and made available through databases.

The rapid technical progress of such AI systems in recent years has not been matched by an equivalent evolution in mechanisms for compensating the creators and investors behind databases. In practice, databases are frequently mined, scraped and re-used for training models without clear attribution or adequate remuneration to copyright holders and *sui generis* rights-holders. This reveals a structural imbalance in the current regulatory and compensation framework: on the one hand, the law formally grants robust rights to database makers and, on the other hand, the enforcement and monetisation of those rights in the context of large-scale automated data collection remain highly uncertain.

At the same time, the interests of bodies engaging in web scraping and data mining cannot be ignored. For new generations of AI systems and other data-driven innovations to emerge, these bodies require access to sizeable datasets in a manner that is efficient, predictable and not unduly burdened by the risk of infringement claims. The legal regime should therefore not only safeguard legitimate expectations of rights-holders, but also provide scrapers and

AI model providers with clear, practicable rules under which data can be lawfully collected and re-used.

In the contemporary information environment, scraping is often the most practical, and sometimes the only, method for obtaining large volumes of structured information from online databases. This reality makes it essential to design a system that reconciles these competing interests: database rights-holders must be able to secure fair compensation for the exploitation of their investments, while data mining bodies need legal certainty that their activities fall within a lawful framework.

Formally, database makers and other copyright holders can, in principle, turn to collective management organisations (**CMOs**) to license and monetise uses of their works or they may attempt to rely on individual negotiation and enforcement, including through various opt-out mechanisms. However, database rights-holders have historically not organised themselves in CMOs to any significant extent, and the practical experience of recent years demonstrates that databases are regularly used, including for AI training, with little or no remuneration to their owners. This suggests that the existing arrangements are failing in practice both from the perspective of rights-holders and from that of scrapers who often operate in a grey area of legal uncertainty.

Introducing entirely new layers of regulation might appear unnecessary, given that the EU has already sought to address the commercial exploitation of databases through the *sui generis* right. Yet, the core issue of fair and effective compensation remains unresolved. In addition, the full range of future applications and configurations of AI models is not yet clear, which justifies a cautious approach from regulators and arguably calls for solutions that are more responsive to market realities and industry practice rather than purely mandatory legal intervention.

The resulting gap between the formal legal framework and actual market practices creates a tension that must be addressed. Reconciling this imbalance is crucial both for ensuring that the legitimate interests of rights-holders are adequately protected and for enabling the continued development of a data-intensive technological ecosystem in the EU that is not built on systematically unlawful or uncompensated uses of intellectual property, nor constrained by disproportionate legal and procedural barriers.

Aim of the thesis

The main aim of this thesis is to determine whether using CMOs, combined with an opt-out option, can function as an effective model for authorising and remunerating the scraping of databases. In addition, this thesis seeks to propose practical ways in which consent and compensation mechanisms could be improved, both within the current legal framework and, where necessary, through targeted regulatory changes.

Tasks of the thesis

1. To examine the current legal framework on database rights, and to critically assess how the DSM Directive regulates data mining.
2. To analyse the system of CMOs introduced and developed in the DSM Directive, and to evaluate its practical impact on both rights-holders and those engaging in web scraping.
3. To analyse the opt-out possibility provided in the DSM Directive, including its theoretical basis and its practical consequences for rights-holders and scrapers.
4. To compare the opt-out mechanism and the CMO model in order to identify gaps and weaknesses in the existing legislation from the perspective of both rights-holders and scrapers.
5. To suggest practical and/or legislative improvements to address the identified problems in the current regulatory framework.

Scope and object of the thesis

This thesis focuses on the legal and practical solutions that currently exist in EU law for dealing with web scraping in relation to copyright and *sui generis* protection of databases. In particular, it examines how the current framework relies on CMOs to collect and distribute remuneration to rights-holders. The work looks at whether such organisations exist and function effectively in the area of database rights, and whether they in fact provide a fair and balanced system of compensation. It also considers what practical consequences the use of CMOs has for those who scrape databases for data mining or AI training purposes.

In addition, this thesis examines the option available to rights-holders to opt out of certain legal mechanisms, meaning that they may exclude their databases from specific exceptions or licensing arrangements, or explicitly reserve their rights in order to pursue compensation on an individual basis. The research examines how this opt-out mechanism is supposed to work in theory, and why, in practice, it may create serious disadvantages for rights-holders. At the same time, it looks at how widespread use of opt-outs may limit or slow down technological development and innovation in the EU, especially in sectors that depend on large-scale access to data.

Finally, this thesis considers the CMO-based solution and the opt-out mechanism together, as part of a single system. It explores how these two approaches interact with each other and how they influence the current EU rules on scraping, copyright and database protection. On this basis, the thesis seeks to assess their overall impact on the functioning and growth of the data-driven technological market in the EU.

Research methodology

To carry out the research tasks and to achieve the aim of the thesis, the following methods are applied:

i) Method of systematic analysis. This method is used to identify, structure and assess the relevant provisions of EU law that define the position of database rights-holders and of those engaging in scraping. By examining the legal acts as a coherent system, it becomes possible to understand how the different instruments interact and what overall effect they have on the rights and obligations of the parties involved.

ii) Linguistic method. The linguistic method is applied when interpreting the wording of the relevant legislation and case law. Particular attention is paid to the exact language used in the legal provisions and in judicial reasoning in order to clarify how these norms should be understood in practice and how they affect the situation of rights-holders and scrapers. This method helps to avoid overly broad or restrictive interpretations.

iii) Logical analysis method. This method is used to examine whether the opt-out mechanism and the collective management system are coherent and effective in practice. It helps to evaluate whether the legal rules lead to reasonable and consistent outcomes when applied to real-life situations. Logical analysis is also used to balance the interests and needs of rights-holders with those of scrapers and the wider data-driven technology market.

iv) Method of synthesis. This method serves to combine the conclusions drawn from the previous methods and to present a general picture of the problems that arise under the current legal framework. By bringing together the different areas of analysis, this method makes it possible to identify common issues, overlapping gaps and those areas of regulation where targeted improvements could offer benefits to both rights-holders and users of data.

Originality of the thesis

While a number of academic works examine the opt-out clause introduced by the DSM Directive and its impact on rights-holders, there is still a clear gap in the literature. Existing research usually focuses either on the position of rights-holders or on the general functioning of copyright exceptions, but does not analyse, in a systematic way, how the opt-out clause and the system of CMO interact and what this means for both sides of the market.

This thesis seeks to fill that gap by looking at these two key mechanisms (i.e., the opt-out option and the CMO-based model) from the perspective of database and copyright holders, as well as from the point of view of scrapers and other users of data, including AI model developers. In doing so, the work offers a more market-oriented and economic perspective on web scraping. It aims to show how a better balance could be struck between ensuring fair remuneration for rights-holders and allowing the data-driven technological sector in the EU to grow and innovate without unnecessary legal barriers.

Sources and literature used

The thesis is based on a combination of international, EU and national legal sources, as well as academic literature. At the international level, the key instruments are the WIPO Copyright Treaty 1996, the Berne Convention and the TRIPS Agreement, which together form the broader framework for copyright protection and certain aspects of database protection.

At EU level, the main legislative acts analysed are the Database Directive 96/9/EC, the InfoSoc Directive 2001/29/EC and the DSM Directive 2019/790, which set out the relevant

rules on copyright, *sui generis* database rights and text and data mining. These instruments are examined in detail, both individually and in their interaction.

The thesis also relies on academic books and articles by authors such as C. Bernard, K. Jerzyk, T. Margoni and M. Kretschmer, A. Kelli, A. Tavast, K. Lindén et al, as well as E. Rosati, whose work provides important doctrinal and critical insights into copyright, database rights and data mining.

In addition, relevant case law of the Court of Justice of the European Union is used to illustrate how the legislation has been interpreted and applied in practice. One judgment from a German court is also examined in more detail, as it offers a concrete example of how national courts approach issues related to web scraping and database protection.

Definitions and abbreviations

1988 Green Paper – Commission of the European Communities Green Paper on Copyright and the Challenge of Technology - Copyright Issues Requiring Immediate Action (1988).

1995 Green Paper – Commission of the European Communities Green Paper on Copyright and Related Rights in the Informational Society.

AI – artificial intelligence.

AI model – in this thesis, this term is used in the sense of a 'general-purpose AI model' under Article 3(63) of the AI Act. It refers to an AI model, including models trained on very large datasets, often using self-supervised techniques, which shows a broad, general capability and can perform many different tasks. Such a model can be built into a wide range of downstream systems or applications, regardless of how it is made available on the market. The term does not cover AI models that are used solely for research, development or prototyping before being placed on the market.

AI model provider – this term follows the concept of 'provider' in Article 3(3) of the AI Act. It means any natural or legal person, public body, agency or other organisation that develops an AI system or a general-purpose AI model (or has it developed) and then places it on the market or puts it into service under its own name or trademark, whether for payment or free of charge.

CMO – collective rights management organisation.

Collective Rights Management Directive – Directive 2014/26/EU of the European Parliament and of the Council of 26 February 2014 on collective management of copyright and related rights and multi-territorial licensing of rights in musical works for online use in the internal market.

Database Directive – Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases.

Data mining or **scraping** – for the purposes of this thesis, this follows the concept of 'text and data mining' in Article 2(2) of the DSM Directive. It means any automated analytical technique used to examine text or data in digital form in order to derive information, including, but not limited to, patterns, trends or correlations.

DSM 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, amending Directives 96/9/EC and 2001/29/EC.

EU – European Union.

InfoSoc Directive – Directive 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.

Member State – any of the 27 Member States of the EU, namely: Belgium, Bulgaria, Czechia, Denmark, Germany, Estonia, Ireland, Greece, Spain, France, Croatia, Italy, Cyprus, Latvia, Lithuania, Luxembourg, Hungary, Malta, the Netherlands, Austria, Poland, Portugal, Romania, Slovenia, Slovakia, Finland and Sweden.

1. Existing Legislation

To gain a clear understanding of how databases are protected in law, it is necessary to examine the relevant rules at both the international and EU levels. International instruments provide important background principles for copyright protection, but they do not recognise or regulate *sui generis* database rights. As a result, the analysis of *sui generis* protection must rely entirely on EU legislation, where this unique form of right was first developed and remains exclusively regulated. By looking at both layers of law together, it becomes possible to form a complete picture of the legal framework that applies to databases and to the activities of those who create them, use them or extract data from them.

1.1. International legal framework

At the international level, databases are mainly protected through copyright rules that cover compilations of data. The WIPO Copyright Treaty confirms that copyright protection applies to databases, regardless of the form, medium or way in which they are fixed or made available to the public. Article 5 of that treaty makes it clear that compilations of data or other material can be protected as copyright works, while the data as such - for example, individual facts or figures - are not protected. A similar approach is reflected in Article 10 of the TRIPS Agreement, which also recognises compilations of data as copyright-protectable works, provided that they show the required level of intellectual creation in their selection or arrangement.

The Berne Convention does not expressly refer to databases. However, it protects ‘every production in the literary, scientific and artistic domain’ and this formula is generally understood in a broad way. On this basis, it is widely accepted in legal literature and practice that databases can fall within the scope of Berne Convention protection when they meet the originality threshold, even though they are not named separately.

1.2. EU legal framework applicable to databases

The discussion on harmonising copyright protection within the EU began in 1988 when the Commission of the European Communities (hereinafter, the **Commission**) published a Green Paper. In the paper, the Commission emphasised several key objectives: copyright owners should be able to treat the European Community as a single internal market, the EU legal framework should help maintain competitiveness with non-EU markets, copyright rules should be enforceable in practice, and an appropriate balance had to be achieved between different groups of stakeholders[1]. Following this initiative, the Commission introduced a series of directives addressing specific areas of copyright law, among which the Database Directive became one of the most significant[2].

The Database Directive, adopted in 1996, represented the EU's first attempt to create a unified regulatory framework for databases, partly because the protection of such works was neither adequate nor consistent across the Member States at the time. Its stated scope covers the legal protection of databases 'in any form'[3], and therefore does not extend to the protection of computer programs, to rental and lending rights, or to rules governing the duration of copyright protection[4].

The Directive provides its own definition of a database as 'a collection of independent works, data or other materials arranged in a systematic or methodical way and individually accessible by electronic or other means'[5]. The preamble further clarifies that the Directive

[1] Commission of the European Communities, *Green Paper on Copyright and the Challenge of Technology - Copyright Issues Requiring Immediate Action* (1988) paras 1.3.2-1.3.4.
[2] Eleonora Rosati, *Copyright in the Digital Single Market: Article-by-Article Commentary to the Provisions of Directive 2019/790* (Oxford University Press 2021) 3.
[3] Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases [1996] OJ L77/20, art. 1.
[4] Ibid, art. 2.
[5] Ibid, art. 1(2).

does not apply to individual audiovisual, cinematographic, literary or musical works themselves, but only to the structure and organisation of the database as such[6].

For databases to receive copyright protection, Article 4 of the Database Directive requires that their selection or arrangement of content reflect the author's own intellectual creation[7]. The concept of 'intellectual creation' was later elaborated by the Court of Justice of the EU in *Infopaq International A/S v Danske Dagblades Forening*, where the Court interpreted this term as the EU standard of originality[8]. Since neither international treaties nor earlier EU legislation provided a clear definition of originality[9], the use of this terminology in the Database Directive brought much-needed clarity. By specifying that originality must stem from the author's choices in selecting or arranging the materials, the Directive clearly outlines the circumstances under which copyright protection applies to databases.

Article 5 sets out the acts restricted by copyright, including reproduction, translation, adaptation, arrangement, and any other modification of the database. It also prohibits any form of distribution to the public, as well as communication, display or performance of the database or its altered versions[10]. In this context, large-scale scraping of databases for commercial benefit could reasonably fall within the range of restricted acts, depending on the method and extent of scraping undertaken.

1.3. *Sui generis* database rights and their distinction from copyright

Alongside the broader international and EU-level initiatives aimed at harmonising copyright protection for databases, the EU has also developed a separate and unique layer

[6] Ibid, preamble recital 17.
[7] Ibid, art. 4.
[8] Case C-5/08 *Infopaq International A/S v Danske Dagblades Forening* [2009] ECR I-6569.
[9] Andres Kelli, Andra Tavast and Krister Linden et al., *'The Impact of Copyright and Personal Data Laws on the Creation and Use of Models for Language Technologies'* in Selected Papers from the CLARIN Annual Conference 2019 (CLARIN 2020) 387.
[10] Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases [1996] OJ L77/20, art. 5.

of protection in the form of *sui generis* rights. These rights were introduced to safeguard the substantial economic and organisational investment that often goes into creating and maintaining databases - an aspect that traditional copyright, focused on originality, does not always capture. The EU therefore sought not only to protect the creative choices of database authors through copyright, but also to secure the financial and structural investment made in assembling large collections of data.

The *sui generis* right is set out in Chapter III of the Database Directive. Under Article 7(1), a database maker may claim protection if they can demonstrate that obtaining, verifying or presenting the contents of the database required a substantial investment[11]. Unlike copyright protection, which depends on the author's intellectual creation and therefore on elements of creativity or originality, the *sui generis* right is grounded entirely in the effort and resources invested. This makes it a distinctive legal concept: no other category of protected subject matter in EU copyright law is defined by the criterion of investment alone.

This differentiation is central to understanding why the EU created the *sui generis* regime. Copyright rewards creative expression, whereas the *sui generis* right acknowledges the economic reality that databases can require considerable financial, technical and human investment even when the underlying materials lack originality. By establishing a separate legal protection, the Directive recognises that databases have strategic and economic value, and that without adequate legal safeguards, their makers might be discouraged from investing in the creation or maintenance of such datasets. In this sense, the *sui generis* right elevates the commercial significance of databases and highlights the EU's intention to treat them as assets worthy of dedicated legal protection.

The scope of protection provided by the *sui generis* right is relatively broad. Article 7 prohibits both the extraction and re-utilisation of the whole or of a substantial part of a

[11] Ibid, art. 7(1).

database, unless such acts fall within the permitted exceptions[12]. Extraction refers to the transfer of contents to another medium by any means or in any form, whereas re-utilisation typically refers to making the contents available to the public. At the same time, lawful users of publicly accessible databases are allowed to extract and re-utilise only insubstantial parts of the contents, provided that such use does not conflict with the normal exploitation of the database or cause unreasonable prejudice to the legitimate interests of its maker[13].

The Directive sets out only a narrow set of exceptions that allow the extraction or re-utilisation of a substantial part of a database without the authorisation of the rights-holder. Article 9 lists three circumstances in which this may occur: for private purposes of a non-electronic database, for illustration for teaching or scientific research (provided that the source is indicated and the use is justified by the non-commercial purpose), and for reasons of public security or for an administrative or judicial procedure[14]. These exceptions are interpreted restrictively and cover only very specific situations.

Given this structure, it follows that the large-scale scraping or re-use of databases for commercial or technological purposes (such as AI model training) would generally fall outside the permitted exceptions. Unless authorised by the rights-holder, such practices may amount to prohibited extraction or re-utilisation, particularly when they concern substantial parts of the database or affect its normal commercial exploitation. Thus, the *sui generis* regime is designed to place clear limits on high-volume automated data collection, making it especially relevant to the contemporary debates surrounding web scraping and data mining.

[12] Ibid.
[13] Ibid, art. 8.
[14] Ibid, art. 9.

1.4. Web scraping practices and their legal regulation

The emergence of large-scale web scraping has created a regulatory challenge for the EU, particularly because traditional copyright rules were not designed with automated data extraction in mind. As scraping became an increasingly common method for gathering online information, the EU was required to adapt its legal framework to address the tensions between technological practices and the rights of database makers. The first steps in this direction were taken through the InfoSoc Directive, which laid down general rules on reproduction and communication to the public. Later, more targeted measures were introduced in the DSM Directive, which specifically deals with text and data mining and attempts to reconcile innovation needs with the protection of copyright and database rights.

1.4.1. Directive 2001/29/EC (InfoSoc Directive)

The momentum for developing a harmonised approach to copyright within the EU continued after the Commission released its first Green Paper in 1988. A follow-up document, published in 1995, reiterated that a 'deeper harmonisation of Member States' copyright laws' was required in order to avoid fragmentation of the internal market and to prevent legal obstacles that might hinder the growth of the emerging information society[15]. The 1995 Green Paper therefore became a central driver for the adoption of the InfoSoc Directive several years later, as the Commission refined its priorities in its 1996 follow-up communication on copyright and related rights in the digital environment. In that document, the Commission identified several core areas requiring legislative action: the right of reproduction, the right of communication to the public, the right of distribution, and the need to secure legal protection for anti-copying technologies[16].

[15] Commission of the European Communities, *Green Paper on Copyright and Related Rights in the Information Society* (1995) 4.
[16] Eleonora Rosati, *Copyright in the Digital Single Market: Article-by-Article Commentary to the Provisions of Directive 2019/790* (Oxford University Press 2021) 3.

This last element, i.e., the legal protection of technological measures designed to prevent copying, plays an important role in the context of database scraping. Technological protection measures (hereinafter, the **TPMs**) can be used by database makers to restrict automated extraction, and the Commission recognised as early as the mid-1990s that an effective copyright system in the digital age would require enforceable, harmonised rules against the circumvention of such protections. As a result, the preamble of the InfoSoc Directive, adopted in 2001, emphasises the need to ensure EU-wide protection against the circumvention of TPMs and against the manufacture or distribution of tools, devices or services that enable such circumvention[17]. The preamble explicitly states that this protection applies not only to copyright works, but also to related rights and *sui generis* rights in databases, thereby bringing databases within the practical scope of these rules[18].

However, the Directive also makes clear that this protection must not unduly interfere with the normal operation of digital products or with technological development. Recital 48 therefore specifies that the Directive should not prohibit devices that have a commercially meaningful use other than preventing technical protection[19]. This 'balancing clause' was introduced to ensure that legitimate technological innovation would not be restricted by overly broad anti-circumvention rules. While originally intended to avoid restricting devices such as digital media players or computer tools, the same reasoning also indirectly affects the development of scraping technologies. Although scraping tools are capable of bypassing various forms of access control or anti-copying features, their use is not limited to acts of circumvention. They often serve broader commercial functions, such as enabling data-driven services, research tools and AI systems. By adopting this flexible policy approach, the InfoSoc Directive unintentionally provides more room for scrapers than for

[17] Directive 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, recital 47.
[18] Ibid, recitals 47 and 48.
[19] Ibid, recital 48.

database rights-holders, shifting the balance slightly towards technological development at the expense of more tight protection for databases.

Article 6 of the InfoSoc Directive provides the legal definition of 'technological measures', describing them as any technology, device or component that is designed to prevent or restrict unauthorised acts with respect to copyright-protected works or subject matter protected by related rights or the *sui generis* database right[20]. The provision then adds that technological measures are 'effective' when the use of a protected work is controlled by the rights-holder through mechanisms such as encryption, scrambling, transformation, or copy-control methods[21]. This language confirms that EU law recognises a wide range of techniques for protecting databases, and that bypassing these methods could amount to an unlawful act - unless it falls under an exception or serves a legitimate purpose recognised by law.

At the same time, the InfoSoc Directive clarifies its own boundaries. Article 1 states that it does not alter or interfere with existing Community rules governing the legal protection of databases[22]. With the exception of Article 11, which is not relevant for the discussion herein, the Directive therefore does not directly amend the rights afforded under the Database Directive. Given that the InfoSoc Directive serves as a cornerstone of harmonised copyright law, its choice not to cover database rights in more depth might appear surprising. From a policy perspective, excluding databases (together with computer programs and certain broadcasting activities) means that an important category of copyright-related content remains governed almost entirely by sector-specific legislation. For database makers, this has the unexpected advantage that the more extensive exceptions permitted under Article 5 of the InfoSoc Directive do not reduce or weaken their rights. In this respect,

[20] Ibid, art. 6(1).
[21] Ibid, art. 6(3).
[22] Ibid, art. 1(2).

the Directive indirectly benefits database rights-holders by ensuring that their exclusive rights remain governed solely by the narrower and more specific provisions of the Database Directive.

Another relevant provision is Article 7 of the InfoSoc Directive, which prohibits the removal or alteration of electronic rights-management information and bans the distribution, importation, broadcasting, communication or making available of copyrighted or *sui generis*-protected material where the person involved knows or has reasonable grounds to know that they are facilitating, enabling or concealing infringement[23]. This rule strengthens the position of database rights-holders by offering an additional tool for enforcement, since many scraping practices involve the removal or alteration of identifying metadata. Article 7 therefore provides a complementary layer of protection that may be invoked in cases where technological measures are circumvented or where rights-management information is removed in the course of scraping.

1.4.2. Directive (EU) 2019/790 on Copyright in the Digital Single Market (DSM Directive)

By the end of the 2010s, the EU co-legislators (the European Parliament and the Council) recognised that existing copyright instruments did not adequately address the reality of large-scale automated data extraction. Text and data mining, which had become increasingly widespread with the growth of machine learning and AI, created new patterns of use that fell outside the traditional categories of reproduction or communication to the public. To respond to these developments, and to provide a more coherent framework for rights-holders and users alike, the European Parliament and the Council adopted the DSM Directive in 2019.

[23] Ibid, art. 7(1).

The DSM Directive acknowledges from the outset the significance of text and data mining as a technological phenomenon. Recital 3 notes that new technologies enable the automated computational analysis of large volumes of digital information (text, sound, images and data) and that such analysis allows users to derive new insights and identify trends that would not otherwise be detectable[24]. By referencing these capabilities directly in the preamble, the EU legislature makes clear that automated data consumption now plays a central role in research, data-driven innovation and the broader digital economy. At the same time, this recital implicitly recognises that such rapid and extensive processing of copyright-protected works and *sui generis*-protected databases raises questions that the older copyright directives were not designed to address. Unlike human users, text and data mining tools can process massive quantities of protected material at unprecedented speed, raising potential risks to the economic value of those works.

However, the DSM Directive does not approach text and data mining solely as a risk or as a practice to be restricted. It also expressly recognises the substantial benefits such technologies bring, particularly to the scientific and research sector. Recital 8 underlines that text and data mining can be a valuable tool for universities, cultural heritage institutions and other research organisations, helping them to carry out their public-interest missions and promoting innovation within the EU[25]. This dual recognition (acknowledging both the opportunity and the risk) forms the foundation for the regulatory model adopted in the Directive. Rather than prohibiting mining *per se*, the Directive attempts to strike a balance that ensures lawful access to data while still preserving the economic position of rights-holders.

[24] 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 [2019] OJ L130/92, recital 3.
[25] Ibid, recital 8.

Under Article 3 of the DSM Directive, research organisations and cultural heritage institutions are granted an exception permitting text and data mining for scientific research purposes, provided they have lawful access to the materials[26]. This narrow exception reflects the EU's view that non-commercial research uses deserve broader freedom to extract and analyse data, given the general societal benefit they provide. Since this thesis focuses on commercial uses of databases and the economic balance between rights-holders and scrapers, this exception is acknowledged only 'in passing'. It should be noted, however, that Article 5(1) allows the use of protected works for illustration for teaching, and Article 5(4) instructs Member States to ensure fair compensation for rights-holders when these uses occur[27].

The provision most relevant to commercially motivated scraping is Article 4, which introduces a broader, market-oriented text and data mining exception. Article 4(1) permits reproductions and extractions of lawfully accessible works 'for the purposes of text and data mining', without reference to a non-commercial limitation[28]. In practical terms, this means that any entity with lawful access (such as those with subscriptions, licences or access to publicly available online material) may perform mining activities, provided that these activities genuinely serve text and data mining functions rather than underlying purposes.

The scope of this provision is significantly narrowed, however, by Article 4(3), which allows rights-holders to reserve their rights expressly[29]. This reservation must be made in an 'appropriate manner' - for example, by including a machine-readable opt-out signal in online content. Through this mechanism, database owners are given a clear way to prohibit mining of their works. For the purposes of this thesis, this is a critical element: it creates

[26] Ibid, art. 3.
[27] Ibid, art. 5(1) and 5(4).
[28] Ibid, art. 4(1).
[29] Ibid, art. 4(3).

the basis for an opt-out system that allows rights-holders to block scraping or to negotiate separate compensation arrangements.

Where rights-holders do not opt out, Article 12 of the DSM Directive expands the role of CMOs, allowing Member States to support collective licensing schemes for works that are difficult to license individually[30]. Although the provision is drafted in general terms, it has significant potential implications for databases: if rights-holders choose not to exercise the Article 4 opt-out, a system of CMO-based licensing could theoretically provide a structured way for database owners to be compensated for mining activities. This presupposes, however, that database makers organise themselves into CMOs, something that has historically not occurred in most EU Member States.

Taken together, Articles 3, 4 and 12 present an attempt at a balanced approach to the regulation of scraping. On the surface, the model appears straightforward: rights-holders who wish to allow mining can rely on CMOs to manage compensation, while those who prefer to restrict access can reserve their rights through a simple opt-out. For scrapers, the system offers a legal pathway to mine lawfully accessible databases, provided that no opt-out has been applied. On paper, therefore, the DSM Directive creates a structured compromise between facilitating innovation and preserving the economic interests of rights-holders.

Whether this compromise performs well in practice, and whether it is suitable for the scale and speed of modern AI development, is one of the central questions explored in the later sections of this thesis.

[30] Ibid, art. 12.

1.5. Regulation of artificial intelligence systems and its models

In recent years, both policymakers and stakeholders across the EU have paid increasing attention to the transparency and accountability of AI systems, particularly those relying on large-scale data mining. A number of actors have called for clearer safeguards in this area. For example, a joint statement issued by authors and performers in 2023 urged the EU to adopt stronger protective measures concerning generative AI under the proposed AI Act[31]. That same year, the European Guild for AI Regulation published a manifesto advocating for a comprehensive regulatory framework for AI companies operating in Europe[32]. In addition, the Communia Association released a policy paper highlighting the implications of using copyright-protected works as training material for machine-learning systems[33]. These documents reflect a growing consensus that the rapid progress of AI (especially generative and 'data-hungry' models) requires corresponding legal oversight to ensure transparency, fairness and respect for intellectual property.

Taking this into account, the EU's first major legislative instrument in the field of AI - the Artificial Intelligence Act (**AI Act**) - has now been adopted. Unlike earlier, more fragmented initiatives, the AI Act establishes a comprehensive regulatory structure designed primarily around risk categories. Its central aim is to impose proportionate obligations on AI model providers depending on the level of societal risk their systems may generate. Although much of the AI Act focuses on applications deemed 'unacceptable risk' or 'high risk'[34], it also introduces important transparency standards relevant to the processing of copyright-protected works and databases, which is a central concern for this thesis.

[31] Initiative Urheberrecht, 'Authors and Performers Call for Safeguards around Generative AI in the European AI Act' (2023).
[32] European Guild for Artificial Intelligence Regulation, *Manifesto for AI companies regulation in Europe* (2023).
[33] Communia, *Using copyrighted works for teaching the machine* (Communia Policy Paper 15, 2023).
[34] Artificial Intelligence Act, art. 5.

The AI Act divides AI practices into categories. The highest level, labelled 'unacceptable risk', comprises activities explicitly prohibited under Article 5. These include practices regarded as fundamentally incompatible with EU values, such as systems that manipulate human behaviour or exploit vulnerabilities[35]. Beneath this lies the 'high-risk' category, defined in Article 6, which subjects AI providers to detailed obligations relating to safety, documentation, data governance and oversight mechanisms[36]. Although large general-purpose models used for data mining or training AI do not always fall into these two categories, the AI Act nevertheless subjects them to a specific set of transparency requirements intended to address concerns about the origin and quality of the training data they rely on.

One of the most significant transparency rules appears in recital 106 of the AI Act, which states that providers of general purpose AI models must comply with EU copyright and related rights law, including honouring reservations of rights expressed under Article 4(3) of the DSM Directive[37]. This reinforces the idea that the AI Act does not operate in isolation but rather complements existing copyright structures. The recital further clarifies that these obligations are meant to ensure a level playing field for all providers, so that AI developers established outside the EU cannot benefit from weaker copyright regimes in their home jurisdictions while still marketing their products within the EU. In essence, the AI Act attempts to ensure that compliance with copyright rules becomes a prerequisite for accessing the EU market.

This requirement is incorporated into Article 53(c) of the AI Act, which obliges providers of general-purpose models to ensure compliance with copyright rules and reservations of rights. Article 53(d) adds an additional obligation for providers to 'draw up and make

[35] Ibid.
[36] Ibid, art. 6.
[37] Ibid, recital 106.

publicly available a sufficiently detailed summary about the content used for training' the AI model[38]. The structure and level of detail of this disclosure will be standardised through templates developed by the AI Office, the EU body responsible for overseeing implementation and supervision of general purpose AI models.

In addition to these transparency requirements, the AI Act introduces a registration obligation. Under Article 49, both high-risk AI systems and general purpose AI models must be registered in an EU-wide database before being placed on the market or put into service[39]. Article 71(4) specifies that this database must be machine-readable, easy to navigate and maintained under the control of the European Commission[40]. In practical terms, this creates a single point of reference for authorities, stakeholders, and the public to identify and monitor AI systems and the data sources upon which they are built.

To support implementation, the AI Act establishes two institutional actors: the AI Office (formally introduced through the Commission's decision of 24 January 2024) and the European AI Board[41]. These bodies are entrusted with developing further codes of practice under Article 56. These codes will, among other functions, define what constitutes an 'adequate level of detail' in the summaries of training data disclosed by AI model providers[42]. Thus, while the Act sets out the basic legal obligations, many of the operational aspects of transparency will be elaborated further by these institutions.

Annex XI of the AI Act reinforces the documentation requirements placed on AI model providers. Providers must maintain up-to-date technical documentation that can be supplied upon request to the AI Office or national authorities[43]. This documentation must include detailed information about the data used for training, testing and validation, how the data

[38] Ibid, art. 53(c) and 53(d).
[39] Ibid, art. 49(1) and 49(2).
[40] Ibid, art. 71(4).
[41] Ibid, art. 3(47) and 6(5).
[42] Ibid, art. 56(2)(b).
[43] Ibid, Annex XI, s. 1(2)(c).

was obtained, the criteria used for selecting data sources, and any measures taken to detect unsuitable or biased datasets. This requirement has particular relevance for this thesis, as it directly concerns the transparency of using scraped databases for model training.

Taken together, these provisions demonstrate that the AI Act marks a significant shift in EU legislative field. Whereas previous digital market initiatives often operated indirectly with respect to AI, the AI Act expressly targets general-purpose AI models and introduces obligations that, at least in principle, strengthen the position of copyright and *sui generis* database rights-holders. Mandatory information disclosure, registration requirements and compliance with reservations of rights create new avenues for monitoring and enforcing intellectual property rules. In theory, this should mean that AI providers who previously scraped data without consent are now required to acknowledge and follow copyright and database rights when training their systems.

1.6. Regulatory gaps in the current framework

Although the DSM Directive introduces several mechanisms intended to safeguard rights-holders while still enabling text and data mining, important gaps remain. In practice, both the CMO-based licensing solution and the opt-out mechanism contain significant weaknesses, leaving database rights-holders exposed to the risk that their works will be scraped and re-used without their permission or without fair compensation.

To begin with, the Directive establishes two separate exceptions for data mining: one for research organisations and cultural heritage institutions (Article 3), and a second, much broader exception that applies to all users (Article 4)[44]. The combined effect of these two provisions is that a wide range of actors can plausibly rely on at least one of the exceptions, which means that scrapers are often better positioned than rights-holders. The so-called

[44] 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 [2019] OJ L130/92, art. 3 and 4.

'everyone' exception in Article 4 is particularly expansive, because it does not limit the ability to carry out text and data mining to academic or public-interest institutions. Instead, it obliges Member States to introduce limitations to copyright and *sui generis* rights that permit reproductions and extractions of lawfully accessible works for the purpose of text and data mining[45]. Although Article 4 does not authorise commercial re-use of the extracted content, it still allows anyone to perform the initial extraction simply by asserting that the purpose of the activity is mining. This creates uncertainty for rights-holders: once a database is publicly accessible, they effectively lose the ability to control who mines it, or the downstream purposes for which the extracted data may eventually be used.

These risks become particularly clear in light of the Hamburg District Court judgment, discussed later in section 3.2. of this thesis, where the Court's interpretation highlights how easily the boundary between 'mining' and 'commercial use' can become blurred. The difficulty lies in the fact that the Directive allows mining as long as the stated purpose is mining but does not provide practical enforcement mechanisms to ensure that the extracted data is not subsequently used for commercial activity.

Another issue derives from Article 4(2), which allows miners to retain copies of extracted data for as long as is necessary for the purposes of text and data mining[46]. In theory, this rule is meant to limit retention to what is strictly required. In practice, however, the rapid growth of AI-driven products shows that extracted datasets are frequently incorporated into commercial models or digital tools. Once incorporated into a training dataset, that material can influence the model's outputs indefinitely, blurring the distinction between temporary retention and long-term commercial use. The DSM provides no practical supervision

[45] Ibid, art. 4(1).
[46] Ibid, art. 4(2).

mechanism to ensure that data is deleted after mining, nor does it specify how 'necessary' retention should be interpreted.

More broadly, the DSM Directive overwhelmingly focuses on the needs of research bodies, universities, cultural heritage institutions and other non-commercial users[47]. While these bodies certainly benefit from legal certainty, the commercial market for data (especially for AI) has expanded far beyond what the DSM Directive initially envisioned. As a result, Article 4 appears somewhat like an 'afterthought', offering a minimal framework for commercial actors without adequately addressing the economic interests of rights-holders or the practical realities of commercial-scale scraping. The DSM Directive therefore does not fully respond to the competing pressures of two rapidly growing sectors: rights-holders who depend on licensing income to monetise their databases, and data miners whose efficiency depends on large-scale, rapid and legally secure access to data.

This imbalance places both sides in a difficult position. Rights-holders are left with limited control over who mines their databases, and with little recourse to prevent commercial use of their content once it has been extracted. Meanwhile, commercial data miners face their own obstacles: if they attempt to obtain individual licences from every relevant rights-holder, the transaction costs and delays would undermine the speed that makes scraping commercially viable in the first place. Where data must be gathered from thousands of sources, individual licensing becomes practically impossible, threatening the very business model of data-driven companies.

The introduction of transparency requirements under the AI Act adds a further layer of complexity. Many AI developers currently lack the technical ability to trace with precision which datasets were used to train their models or how those datasets were acquired. The AI Act now obliges them to describe their training data in detail and to comply with

[47] Ibid, recitals 3 and 8.

reservations of rights made under Article 4(3) of the DSM Directive[48]. This places model providers in a vulnerable position: if they accurately disclose their training data, they may expose themselves to claims for past unlicensed use. If they fail to comply, they face potential sanctions from EU supervisory bodies.

In addition, once the sources of training data are revealed, rights-holders may require their works to be removed or may demand new licensing agreements. Engaging in large numbers of individual negotiations would cause significant delays and could, in some cases, render the continued operation of the AI model economically unfeasible. The combined effect of these regulatory obligations therefore places pressure both on rights-holders, who struggle to prevent misuse of their content, and on data miners and AI developers, who face rising compliance burdens without a clear, efficient framework for obtaining the necessary permissions.

2. The Collective Management Organisation (CMO) model

The DSM Directive introduces collective licensing as one of the possible mechanisms for addressing the legal and economic challenges associated with text and data mining. This approach is reflected in Article 12, which allows Member States to rely on collective management organisations at national level when implementing systems for licensing certain uses of copyright-protected works and other subject matter, including databases. Any such system would operate within the framework established by the Collective Rights Management Directive, which sets out the rules governing the functioning, transparency and accountability of CMOs[49].

[48] Artificial Intelligence Act, recital 106. See also Directive (EU) 2019/790, art. 4(3).
[49] 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 [2019] OJ L130/92, art. 12; Directive 2014/26/EU of the European Parliament and of the Council of 26 February 2014 on collective management of copyright and related rights [2014] OJ L84/72.

Under the Collective Rights Management Directive, rights-holders may be represented by a CMO even in situations where they have not explicitly authorised that organisation to act on their behalf, provided that the organisation has a legal mandate or is presumed to represent the relevant category of rights-holders[50]. This form of extended collective licensing is not new in EU copyright law. Similar models have long been used for the licensing of musical works, audiovisual content and performances, where individual licensing would be inefficient or practically impossible. In many Member States, these systems have proven effective in ensuring remuneration for rights-holders while allowing users to obtain licences through a single negotiating partner. Against this background, it may appear reasonable to suggest that a comparable framework could be applied to databases and to data mining activities, particularly where the scale of use makes individual negotiations unrealistic.

The DSM Directive itself supports this logic. Recital 73 explicitly states that Member States should have flexibility in how they implement remuneration mechanisms, including through collective bargaining or other collective arrangements[51]. The emphasis on collective bargaining is significant, as it reflects the underlying rationale of CMOs: the collective strength of many rights-holders negotiating together is assumed to result in more balanced outcomes than individual negotiations between a single rights-holder and a large commercial user. In the context of data mining, where powerful technology companies may extract value from thousands of databases simultaneously, collective representation could theoretically restore some bargaining power to database makers.

This rationale is reinforced by Article 12(2) of the DSM Directive, which acknowledges that individual licensing may be ‘onerous and impractical’ in certain situations, to such an

[50] Directive 2014/26/EU of the European Parliament and of the Council of 26 February 2014 on collective management of copyright and related rights [2014] OJ L84/72, arts. 3-7.
[51] 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 [2019] OJ L130/92, recital 73.

extent that licensing transactions are unlikely to occur at all[52]. In cases of large-scale scraping, this observation is particularly relevant. Data mining often involves the automated extraction of content from number of sources, sometimes across different jurisdictions. Requiring miners to identify and negotiate separately with each database rights-holder would undermine the efficiency that makes data mining economically viable in the first place. From this perspective, collective licensing appears to offer a practical solution capable of connecting the gap between legal compliance and technological reality.

Nevertheless, the application of collective management to database rights also raises serious concerns. One of the most controversial aspects of extended collective licensing is that it may significantly reduce the autonomy of individual rights-holders. Subjecting database makers to collective management means that they may lose direct control over whether, and on what terms, their databases are scraped and used. For some rights-holders, especially those whose databases represent a core commercial asset, the idea that remuneration levels would be negotiated collectively rather than individually may be perceived as disempowering. There is a risk that such a system could discourage investment in database creation if rights-holders feel that they no longer have meaningful influence over how their works are exploited.

Recital 73 of the DSM Directive attempts to address these concerns by stating that remuneration should be appropriate and proportionate to the actual or potential economic value of the licensed rights, taking into account factors such as market practices and the real exploitation of the work[53]. However, the DSM Directive does not provide concrete criteria or enforceable standards for assessing whether remuneration is in fact fair. As has been pointed out in academic commentary, the DSM stops short of imposing a clear obligation to ensure fair remuneration in areas where rights cannot be contractually

[52] Ibid, art. 12(2).
[53] Ibid, recital 73.

excluded[54]. This creates a tension between the stated objective of fairness and the practical regulatory outcome.

The combination of vague language on remuneration and the introduction of collective management mechanisms therefore risks undermining confidence among database rights-holders. While CMOs may offer efficiency and legal certainty for data miners, the absence of clear guarantees regarding compensation levels means that rights-holders cannot be certain that collective licensing will reflect the true economic value of their databases. In this sense, the CMO model under the DSM Directive appears to prioritise market functionality and transactional efficiency over the effective protection of individual economic interests. This discrepancy raises doubts as to whether collective management, at least in its current form, can provide an adequately balanced solution for data mining in the EU.

2.1. CMOs in the field of databases

The concerns outlined above become even more noticeable when collective management is considered specifically in relation to database rights-holders. Traditionally, CMOs have played an important role in managing and enforcing the rights of performers, composers, authors and other creators of more conventional copyright-protected works. In those fields, collective licensing has developed over decades and is supported by relatively stable valuation models. Databases, however, differ substantially from traditional copyright subject matter, both in their legal structure and in their economic characteristics. As a result, errors or inefficiencies in how CMOs would calculate and distribute remuneration in the database context could have far-reaching consequences for both rights-holders and users.

[54] Séverine Dusollier, 'The 2019 Directive on Copyright in the Digital Single Market: Some Progress, a Few Bad Choices, and an Overall Failed Ambition' (2019) 57 *Common Market Law Review* 1023.

The idea of adapting CMOs to fulfil altered or expanded functions in the data economy brings them conceptually close to so-called Alternative Compensation Systems (hereinafter, the **ACS**), as discussed by Quintais in his analysis of copyright in the digital environment[55]. The ACS models aim to address mass uses of protected works through mechanisms that differ from typical licensing. However, such systems are not currently recognised or implemented under EU copyright law. For this reason, focusing on existing and legally plausible compensation mechanisms (rather than hypothetical alternatives) provides a more realistic basis for assessing how rights-holders might be remunerated for database scraping under the current legislative framework.

Although Quintais' analysis does not specifically focus on databases, the structural weaknesses he identifies in collective management systems are highly relevant in this context. These include CMOs' frequent inability to provide effective sorting and categorisation functions, high operational costs, significant market power, institutional inertia and slow decision-making, as well as revenue distribution methods that rely on indirect proxies rather than actual use[56]. Each of these shortcomings is likely to be amplified when applied to databases.

The lack of effective sorting mechanisms is particularly problematic for databases because their economic value is highly context-dependent. Unlike many traditional copyright works, databases benefit not only from copyright protection but also from *sui generis* rights, which are directly linked to the investment made in obtaining, verifying or presenting their contents. Their value therefore varies significantly depending on sector, purpose and timing. A database that is extremely valuable for a specific industry - such as finance, healthcare or language technology - may be of little or no value outside that context. Unlike

[55] João Pedro Quintais, *Copyright in the Age of Online Access: Alternative Compensation Systems in EU Copyright Law* (Kluwer Law International 2017).
[56] Ibid.

artworks or musical compositions, which are often valued through established cultural or market benchmarks, databases do not lend themselves to uniform valuation models. Moreover, size alone is not a reliable indicator of value: a small, highly specialised dataset may be far more economically significant than a large but generic collection of data.

This problem is further complicated by the realities of AI model training. In many cases, the underlying value lies not in the database as a coherent structure, but in the individual data points it contains. Databases may be scraped incidentally as part of a broader data collection process, which calls into question whether a broad, uniform remuneration model based on database-level valuation would adequately reflect actual economic use. In such cases, a broad collective compensation mechanism risks disconnecting remuneration from real value creation.

Operational costs pose another important barrier. Managing databases through CMOs would require far more complex infrastructure than the one used for traditional copyright works. For example, musical works or audiovisual content require relatively ordinary storage and tracking capabilities. By contrast, maintaining up-to-date records of databases (many of which change frequently) would demand substantial analytical resources, advanced security measures and constant technical oversight. Even establishing a national-level registry of databases would involve significant investment, and an EU-wide system would multiply these costs substantially. These technical and financial burdens alone could deter Member States from supporting database-focused CMOs, especially where funding would have to be sourced from already limited public or private resources.

Quintais also highlights the tendency of CMOs towards bureaucratic resistance to change - a view supported by research from Handke, Quintais and Bodó[57]. Their work shows that

[57] Christian Handke, João Pedro Quintais and Balázs Bodó, 'The Economics of Copyright Compensation Systems for Digital Use' (SERCI Annual Congress 2013).

large collective organisations often struggle to adapt quickly to evolving markets due to internal bureaucracy and rigid governance structures. While economies of scale can reduce transaction costs for both rights-holders and users, they also slow responsiveness. In the rapidly evolving data and AI markets, where technological and commercial practices change quickly, such delays could render collective licensing frameworks outdated almost as soon as they are implemented. These problems would likely be even more pronounced in the context of an EU-wide CMO, where organisational complexity and political coordination requirements would be even more significant.

One possible way to address some of these concerns would be to limit the role of CMOs to representation and enforcement, rather than valuation. CMOs have frequently been compared to trade unions, acting collectively on behalf of rights-holders to negotiate and enforce rights[58]. If database owners were allowed to self-assess the value of their databases, CMOs could operate primarily as agents that collect licensing fees, monitor compliance and identify infringements. This approach could reduce reliance on raw data for valuation and improve enforcement outcomes. However, it would also increase administrative complexity, as each database would require individual treatment, thereby driving up operational costs even further.

It should also be acknowledged that the existence of CMOs could benefit AI model providers and other data miners. A functioning collective licensing system would provide a single access point for obtaining permissions and paying remuneration, reducing legal uncertainty and encouraging greater transparency. In theory, this could facilitate lawful data use and promote more responsible industry practices.

[58] Richard Caves, 'Creative Industries - Contracts Between Art and Commerce' (2000) 17(2) *Journal of Economic Perspective*s 73; Ruth Towse, *Creativity, Incentive and Reward: An Economic Analysis of Copyright and Culture in the Information Age* (Edward Elgar 2001); Martin Kretschmer, 'Copyright Societies Do Not Administer Individual Rights: The Incoherence of Institutional Traditions in Germany and the UK' in *Copyright in the Cultural Industries* (Edward Elgar 2002) 140.

Despite this potential, the absence of database-focused CMOs in practice is evident. Even prior to the rise of large-scale scraping and AI training, database rights-holders did not organise themselves into collective management structures. This suggests that the perceived benefits of collective administration have not outweighed the costs and risks involved. Although current EU legislation provides sufficient legal clarity to allow database owners to establish CMOs at national level, the fact that this has not occurred indicates a lack of economic incentive. In its current form, collective management therefore appears to be a theoretically plausible but practically underdeveloped solution for compensating database rights-holders in the context of data mining.

3. The opt-out solution

The same economic considerations that complicate the application of collective management to databases are equally relevant when analysing the opt-out mechanism available to rights-holders. In practice, the opt-out option is not merely a theoretical alternative to collective licensing but reflects the current market reality for database owners in the EU. Since no Member State has established CMOs that specifically administer remuneration for database rights-holders, those rights-holders effectively operate as if they have already opted out of any collective system. This has direct consequences for their bargaining power and legal position *vis-à-vis* scrapers and AI model providers, both in terms of compensation and in relation to the deployment of technological protection measures.

In the absence of database-focused CMOs, rights-holders are left to enforce their copyright and *sui generis* rights on an individual basis. This means that, from a market perspective, database owners stand alone when negotiating with large-scale data miners or AI developers. While in theory this preserves full autonomy over their works, in practice it significantly weakens their position. Individual rights-holders rarely possess the economic

leverage, legal resources or negotiating capacity necessary to engage effectively with bodies whose business models depend on scraping data at scale. As a result, the opt-out reality (where rights-holders retain exclusive control but lack collective support) often leads not to effective licensing, but to widespread uncompensated use of databases.

This dynamic has been critically examined by Senftleben, who argues that imposing individual rights clearance obligations on AI trainers is likely to undermine any remuneration framework altogether. According to Senftleben, once AI developers are required to verify rights ownership, comply with specific payment conditions and obtain permissions for individual works or databases, the increasing burden of rights clearance risks making remuneration schemes economically unworkable[59]. This observation is particularly relevant in the context of databases, where data mining typically involves thousands of sources, often combined automatically and continuously.

The current market situation appears to confirm this assessment. Because database rights-holders are not organised through CMOs, they lack a practical mechanism to claim remuneration, even where their rights are clearly infringed. At the same time, scrapers and AI model providers have no realistic way to distribute compensation on an individual basis. Negotiating licences separately with each database owner would be time-consuming, legally complex and financially inefficient, undermining the very scalability that makes data-driven innovation viable. The result is a structural imbalance: rights-holders remain formally protected under EU law, but in reality are unable to monetise their databases, while scrapers continue to extract and use data without meaningful compensation structures in place.

[59] Martin Senftleben, 'Generative AI and Author Remuneration Accepted' (2023) 54 *International Review of Intellectual Property and Competition Law* 1564.

Despite these shortcomings, the opt-out mechanism remains a crucial component of any potential CMO-based solution. Even if a functioning collective management system for databases were to be established, the ability of rights-holders to opt out should not be viewed as unnecessary. On the contrary, opt-out rights serve as an essential safeguard of autonomy. In a well-designed system, collective licensing would operate as the default mechanism for efficient remuneration, while opt-out would remain available for rights-holders who wish to retain individual control over their databases.

From a normative perspective, preserving the opt-out option aligns with the fundamental principles of copyright and database protection. Rights-holders should ultimately retain the authority to decide whether their works are licensed, under what conditions, and for what price. This includes decisions about who may scrape their databases, for which purposes, and on what terms. Removing or weakening this ability could discourage investment in database creation, particularly in cases where databases represent a core commercial asset rather than a by-product of another activity.

Accordingly, while the opt-out mechanism on its own does not offer a viable solution to the problem of compensating database rights-holders in a data-driven economy, it remains an indispensable element of a broader regulatory framework. Its value lies not in replacing collective management, but in complementing it - ensuring that efficiency does not come at the cost of rights-holder autonomy, and that database owners retain the final say over the exploitation of their works.

3.1. Theoretical basis of opting out and non-consent clauses

A core condition for the effectiveness and legality of any extended collective licensing system is that rights-holders retain adequately and informed control over the use of their works. As Jerzyk correctly emphasises, an opt-out mechanism can only function in a manner compatible with EU law if authors and rights-holders are individually informed

about the scope and nature of the uses carried out under a collective licence[60]. This requirement becomes particularly important in the context of databases, where no established CMOs currently operate. In such circumstances, database rights-holders cannot be expected to rely on implied knowledge of collective schemes. Instead, they must be clearly informed about their rights and the practical consequences of remaining within or opting out of a licensing mechanism.

Article 12(3) of the DSM Directive seeks to address this concern by putting an obligation to the Member States to ensure that rights-holders who have not authorised a CMO can easily and effectively exclude their works or other subject matter from the licensing mechanism at any time[61]. As Jerzyk points out, this represents a significant shift from the traditional opt-out model found in the Collective Rights Management Directive. Under the older framework, opt-out rights were typically available only after licences had already been concluded. By contrast, the DSM allows rights-holders to opt out both before and during the licensing period[62].

From the perspective of database rights-holders, this enhanced opt-out flexibility appears beneficial. Given the nature of contemporary data use (especially the training of AI models on large-scale datasets) rights-holders may wish to prevent any form of automated extraction from the outset before their databases are incorporated into training pipelines. The DSM opt-out mechanism therefore acknowledges the increased risks associated with modern data exploitation and strengthens individual control accordingly.

[60] Karolina Jerzyk, 'Balance of Rights in Directive 2019/790 on Copyright in the Digital Single Market - Is the Opt-out Clause Sufficient for the Protection of Author's Moral Rights?' (2021) 7(2) *Santander Art and Culture Law Review* 239.
[61] 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 [2019] OJ L130/92, art. 12(3).
[62] Karolina Jerzyk, 'Balance of Rights in Directive 2019/790 on Copyright in the Digital Single Market - Is the Opt-out Clause Sufficient for the Protection of Author's Moral Rights?' (2021) 7(2) *Santander Art and Culture Law Review* 239.

At the same time, this very flexibility undermines the effectiveness of collective licensing. If rights-holders can withdraw their works at any point, CMOs face reduced bargaining power when negotiating licences on behalf of remaining members. The uncertainty created by the possibility of mid-term opt-outs weakens the stability of licensing agreements and reduces predictability for licensees. In practice, CMOs may be unable to guarantee access to a defined pool of databases over time, which limits their usefulness as intermediaries in large-scale data mining arrangements.

The opt-out mechanism also creates significant difficulties for data miners and AI model providers. Where a licence has been concluded through a CMO, a subsequent opt-out by an individual rights-holder may require the licensee to remove the relevant database from its products or datasets. This presents not only commercial risks (such as reduced functionality or quality of data-driven services) but also technical challenges. In the context of AI model training, it remains unclear whether data that has already influenced a model can be effectively removed. At present, even AI developers themselves often lack full transparency into how training data shapes model behaviour, raising doubts as to whether 'unlearning' specific datasets is technically feasible at all.

This problem is acknowledged indirectly in the AI Act, which introduces obligations for AI model providers to improve transparency not only regarding training data, but also regarding the functioning of AI systems themselves. Article 53(1) requires providers of general-purpose AI models to comply with detailed transparency and documentation duties. However, achieving the level of technical understanding necessary to meet these requirements will take time. While compliance may be more realistic for newly developed models, providers of existing, widely deployed AI systems may be reluctant to invest the resources required to retrospectively analyse training datasets and disclose their sources. The risk of legal exposure, reputational damage and potential liability further discourages voluntary disclosure.

Beyond legal design, the effectiveness of the opt-out mechanism ultimately depends on its enforceability. If rights-holders cannot practically prevent scraping or cannot detect violations, their formal right to opt out remains largely symbolic. In fast-moving data markets, legal protection must be supported by technological measures that operate at machine speed. Without such safeguards, rights-holders may theoretically retain exclusive rights while, in reality, being unable to enforce them.

The InfoSoc Directive provides a general framework for understanding technological protection measures, describing them as access-control or protection processes such as encryption, scrambling, transformation of content, or copy-control mechanisms. These measures are intended to prevent unauthorised acts before they occur. However, their effectiveness varies considerably depending on the technical capacity and resources of individual rights-holders. For publicly available databases, complete technical exclusion is often neither feasible nor desirable.

As a result, an additional layer of protection has emerged in the form of non-consent declarations. To be effective, such declarations must be expressed in a machine-readable format that automated systems can recognise. However, the concept of machine readability has been criticised for its lack of precision, particularly in earlier legislative debates[63]. Nonetheless, Recital 18 of the DSM Directive suggests that machine-readable formats generally refer to structured data that can be automatically processed by computers[64]. In the online environment, this may include terms and conditions or technical signals such as robots.txt files, which indicate whether automated access is permitted[65].

[63] Roberto Ducato and Alain Strowel, 'Limitations to Text and Data Mining and Consumer Empowerment: Making the Case for a Right to "Machine Legibility"' (2018) 50 *International Review of Intellectual Property and Competition Law* 649.
[64] Roberto Ducato and Alain Strowel, 'Ensuring Text and Data Mining: Remaining Issues With the EU Copyright Exceptions and Possible Ways Out' (2021) 43 *European Intellectual Property Review* 324.
[65] Bernt Hugenholtz, 'The New Copyright Directive: Text and Data Mining (Articles 3 and 4)' (2019) *Kluwer Copyright Blog*.

However, the practical value of non-consent clauses is limited. They are effective only against scrapers willing to respect legal and technical norms. Those who decide to ignore such signals can bypass them with relative ease. For this reason, technical barriers remain a more robust, albeit more expensive, form of protection, especially for rights-holders who opt out of collective licensing and seek to negotiate individual agreements.

Taken together, these considerations demonstrate that opt-out rights and non-consent mechanisms are deeply interconnected. Without reliable enforcement of non-consent signals, the opt-out mechanism cannot function effectively as a standalone solution for remuneration outside CMOs. Even with the enhanced transparency obligations introduced by the AI Act, rights-holders who opt out would still bear the burden of identifying whether their databases have been used in AI training. In this respect, collective management may offer advantages in terms of monitoring and enforcement. Nevertheless, the current technological tools for expressing and enforcing non-consent in a machine-readable way remain insufficient, leaving database rights-holders exposed despite the formal legal safeguards available to them.

3.2. The Hamburg District Court decision

The practical limitations of both technological protection measures and machine-readable scraping prohibitions are clearly illustrated by a recent decision of the Hamburg District Court. This judgment provides one of the first detailed judicial assessments in the EU of how database scraping, non-consent clauses and the DSM text and data mining exception operate in practice, particularly in the context of AI model training.

In September 2024, the Hamburg District Court delivered its judgment in case No. 310 O 227/23[66]. The case concerned the scraping of a database and the subsequent use of the

[66] Hamburg District Court, judgment of 27 September 2024, Case No 310 O 227/23.

extracted data as training material for an AI model. Although the dispute did not directly involve the database owner as a claimant, the factual background and legal reasoning are highly relevant for understanding the real-world effectiveness of opt-out mechanisms and non-consent clauses.

The claimant in the proceedings was a professional photographer whose work was included in a database operated by company A. The photographer had voluntarily licensed his photographs to company A for the purpose of commercial licensing to third parties. Company A maintained an online database that incorporated technological measures and a clause indicating that scraping was not permitted. Despite this, company B, a non-profit organisation, carried out automated data mining of the database. The scraped dataset was subsequently accessed by a separate AI development company, which used it to train a for-profit AI model. As a result, the claimant's photograph became part of the model's training data, together with other content from company A's database.

Notably, company A itself was not a party to the proceedings. Instead, the photographer brought the claim independently, asserting that his copyright had been infringed as a result of the scraping and downstream use of the database. This procedural posture already highlights a structural weakness in enforcement: database owners and individual contributors may face fragmented or overlapping interests when attempting to assert their rights against large-scale data mining practices.

The case raised two central legal questions. First, the Court examined whether the technological measures and non-consent clause implemented by company A were sufficient to prohibit the scraping carried out by company B. Second, it assessed whether the extraction and use of data from the database (specifically the claimant's photograph) fell within the scope of the text and data mining exception under Article 4 of the DSM

Directive, despite the fact that the data ultimately contributed to the training of a commercial AI model.

With regard to the first issue, the Court held that the non-consent clause used by company A was likely to satisfy the requirement of machine readability. In reaching this conclusion, the Court interpreted the relevant provisions of German law in a manner that associates 'machine readability' with 'machine comprehensibility'. This interpretation suggests that a clause expressed in natural language can, in principle, be understood by automated systems, provided that the systems are capable of processing such language. The Court further relied on Recital 18 of the DSM Directive, emphasising that rights reservations must be expressed in an 'appropriate' manner, rather than in the simplest or most technically standardised format possible. On the facts of the case, the Court considered it plausible that, at the time of scraping in 2021, the defendant's systems were capable of recognising and processing the natural-language prohibition against scraping.

Despite acknowledging the effectiveness of the non-consent clause, the Court nevertheless found that the scraping activity itself fell within the scope of the Article 4 DSM exception. The decisive factor was the nature and purpose of company B's activities. The Court accepted that company B was a non-profit organisation engaged in scientific research and that the scraping was carried out solely for the purpose of text and data mining. On this basis, the Court concluded that company B was entitled to rely on the Article 4 exception and was therefore not liable for copyright infringement, notwithstanding the explicit objection expressed by the database owner.

Taken together, these findings lead to an interesting outcome. On the one hand, the Court effectively confirmed that company B was aware of company A's refusal to allow scraping. On the other hand, it held that this refusal did not prevent company B from relying on the DSM exception, given its non-profit and research-oriented status. While this reasoning may

appear coherent when viewed narrowly, it raises significant concerns when the broader context is taken into account.

In particular, company B bore no liability for the subsequent commercial use of the scraped data by the AI development company. The Court justified this by stating that there was insufficient evidence of a legal or factual link between company B and the third-party AI developer. As a result, no responsibility was attributed for the fact that data initially extracted under a research exception ultimately contributed to the training of a commercial AI model. This aspect of the judgment exposes a critical gap in the regulatory framework: where data passes through multiple actors, liability may vanish entirely, leaving rights-holders without compensation or effective remedies.

The implications of this reasoning are potentially far-reaching. If scrapers can rely on the Article 4 exception while avoiding commercial exploitation by third parties, the opt-out mechanism loses much of its practical value. Rights-holders may clearly express non-consent, yet still find their works incorporated into commercial AI systems without remuneration. In such scenarios, neither technological measures nor legal reservations offer meaningful protection.

It should be emphasised that the Hamburg District Court is a lower court, and its judgments are primarily authoritative within the German legal system. Nevertheless, in the absence of higher-court or CJEU case law on the issue, this decision provides an influential interpretation of the DSM Directive as applied to database scraping and AI training. Until more definitive guidance emerges at EU level, the judgment serves as a primary example of how current legislation may fail to resolve the tension between innovation, research exceptions and the effective protection of database and copyright holders.

3.3. Outcomes and implications of the opt-out and non-consent system

Hamburg District Court decision illustrates particularly well the structural weaknesses of the opt-out and non-consent system when applied in practice. Even where a database rights-holder clearly expresses an intention not to allow scraping and implements a non-consent clause, the combined effect of the DSM exceptions (especially those benefiting non-profit research organisations) may result in that objection being overridden without any legal consequences for the scraper or downstream users. While such outcomes may align with the DSM's objective of facilitating research and innovation, they nonetheless create an imbalance that disproportionately disadvantages rights-holders.

A key problem lies in the fact that data scraped under a lawful exception does not necessarily remain restricted to the original purpose for which it was collected. As demonstrated by the Hamburg case, once a database has been scraped and stored in the archives of a research organisation, the data may become accessible to third parties. Even if a direct link between the original scraper and a subsequent AI model provider cannot be established in a specific case, the mere possibility that scraped databases may later be used for commercial AI training represents a persistent and unresolved risk for rights-holders. In this respect, the opt-out mechanism fails to provide meaningful protection, as the initial lawful extraction effectively opens the door to downstream uses that escape both control and remuneration.

Another significant issue concerns the obligations placed on institutions that benefit from the DSM Directive exceptions. Research organisations, cultural heritage institutions and other non-profit bodies are granted privileged access to protected works for text and data mining, yet the legislation does not impose correspondingly strict duties on them to prevent secondary exploitation. Where such institutions possess the technical capability to scrape databases, it would be reasonable to expect them to also maintain robust safeguards against

unauthorised re-use, particularly for commercial purposes. At present, however, the DSM Directive does not require these actors to implement protection measures equivalent to those adopted by rights-holders themselves.

This imbalance places database owners in a vulnerable position. Although institutions benefiting from DSM exceptions do not acquire ownership or other proprietary rights over the databases they scrape, rights-holders receive no guarantee that their data will be protected once it has been extracted. Ideally, scraped databases should be subject to at least the same level of technological protection (both in scope and quality) as that initially implemented by the rights-holder. The absence of such obligations means that rights-holders may face repeated infringements of their copyright and *sui generis* rights without any practical means of intervening once the initial scraping has occurred.

The Hamburg case also highlights the crucial role of technological identifiers in enforcing rights. In that case, the Court was able to determine that the photograph in question originated from company A's database because the image contained a visible watermark applied to content accessed by unpaid users. This watermark served as a clear evidentiary link between the scraped content and the original database. However, the judgment implicitly raises a troubling question: how would the case have been resolved if no such watermark had existed? Without a visible or traceable mark, identifying the origin of individual data items within large training datasets would be significantly more difficult, if not impossible.

This observation underscores the importance of technological measures as a practical enforcement tool for database rights-holders. While legal rights and opt-out declarations exist in theory, their effectiveness in practice often depends on the ability to demonstrate that specific content has been scraped and re-used. The Hamburg decision therefore

suggests that, in the absence of robust identification mechanisms, courts may struggle to assess claims relating to database misuse.

Finally, this raises broader concerns regarding cases which occurred before the transparency obligations introduced by the AI Act. Where AI model providers are not required to disclose detailed information about training data, rights-holders may have little chance of identifying whether their databases have been used at all. This creates a temporal gap in protection: databases scraped before the AI Act's transparency regime may remain rooted in AI systems without any realistic prospect of detection or compensation. As a result, the combined opt-out and non-consent framework, while formally protective, appears insufficient to safeguard rights-holders' interests in the current technological and regulatory landscape.

4. Comparative analysis of the two approaches

Both the collective management model and the opt-out approach, which aim to ensure remuneration for database rights-holders, face the same structural difficulty: neither system enables rights-holders to reliably identify the use of their databases within AI models trained on scraped data. Although the AI Act introduces transparency obligations for AI model providers, enforcing these rights is likely to be costly and complex for individual rights-holders. At the same time, delegating this administrative burden to a CMO risks creating a system that is slow, bureaucratic and expensive to operate.

4.1. Perspective of rights-holders

From the perspective of database rights-holders, the transparency obligations introduced by the AI Act can generally be viewed as a positive development. These requirements improve the possibility for individual rights-holders to obtain information about whether their databases have been used in the training of AI models, which in turn makes operating

without the support of CMOS somewhat more realistic. However, when the entire process is considered in practice (from the conclusion of licensing agreements, to the act of scraping itself, to identifying such use and potentially engaging in litigation against AI model providers) the CMO model still appears to offer a more rights-holder-oriented and protective solution overall.

Although database owners are able to implement technological measures in order to identify the contents of their databases, such as watermarks similar to those used in the Hamburg District Court case, there is currently no enforceable obligation under EU law that would require bodies benefiting from the DSM exceptions to apply the same level of protection to the databases they scrape. This regulatory imbalance operates in favour of scrapers and AI model providers in the short term, as it reduces their compliance obligations and limits immediate accountability. At the same time, this lack of harmonised protection creates legal uncertainty for those same bodies, since they may unknowingly use databases that are protected by copyright or *sui generis* rights. In the long term, this uncertainty exposes scrapers and AI model providers to the risk of litigation, as well as to potential requirements to modify or withdraw products that rely on unlawfully sourced data.

Currently, individual rights-holders face substantial difficulties in tracking the use of their databases and in collecting compensation from data scrapers and AI model providers. In many cases, compensation is only discussed once a rights-holder initiates a legal claim. This reflects the current market reality, in which major technology companies (such as OpenAI, Google and Meta) have argued that they are not required to compensate rights-holders for data mining activities carried out over extended periods of time. As long as there are no immediate legal or financial consequences for such practices, scrapers benefit from the lack of effective scrutiny. While AI model providers will eventually be subject to accountability mechanisms under the AI Act, the present lack of clarity places both scrapers and AI developers in a potentially disadvantageous position in the long run. Their products

may ultimately rely on databases whose rights-holders are legally entitled to remuneration and may demand that their works be removed from AI models entirely.

When assessed from an efficiency and enforcement perspective, CMOs operating on the basis of economies of scale appear better equipped to detect infringements of database rights than individual rights-holders, even in light of the transparency requirements introduced by the AI Act. Once AI model providers are required to disclose information about their training data, CMOs would still be more effective than individual rights-holders in systematically analysing such disclosures and identifying uncompensated uses. This approach would likely be less costly than requiring CMOs to store and technically monitor databases themselves. In addition, where AI model providers fail to comply with transparency obligations, CMOs are in a stronger position to draw regulatory attention to such failures than individual rights-holders acting independently.

Nevertheless, Senftleben expresses doubts as to whether transparency alone can ensure effective remuneration, noting that even with machine-readable remuneration systems, it remains difficult to control compliance and to ensure that payments accurately reflect all works used for AI training purposes[67]. This limitation affects both the opt-out and CMO-based approaches and effectively leaves substantial control over scraped databases in the hands of scrapers and AI model providers. Despite this shared weakness, the CMO model retains greater institutional authority and collective leverage, which increases the likelihood of achieving fair compensation. Individual rights-holders, by contrast, possess far less bargaining power when dealing with large-scale commercial actors.

The imbalance between individual rights-holders and multinational technology companies can be illustrated by the lawsuit brought by The New York Times Company against

[67] Martin Senftleben, 'Generative AI and Author Remuneration Accepted' (2023) 54 *International Review of Intellectual Property and Competition Law* 1564.

Microsoft and OpenAI, which was later joined by the Authors' Guild and other authors as a class action[68]. Although the dispute primarily concerns individual journalistic works, the complaint also refers to the New York Times Article Archive and the TimesMachine, which are described as a unique and highly valuable database[69]. While the case is still pending (which is evident by the April 2025 decision[70] of U.S. District Court for the Southern District of New York, which did not resolve the dispute on the merits) and falls outside the EU legal framework, it may nevertheless provide insight into whether litigation is a realistic path for individual rights-holders seeking compensation for the use of their works in AI training. Even though The New York Times is itself a powerful and well-resourced rights-holder, the case still illustrates the structural challenges of enforcing rights against major technology companies.

The New York Times litigation also draws attention to class action lawsuits as a potential alternative for database rights-holders in the absence of CMOs. While class actions allow rights-holders to pool resources and reduce inconsistencies in bargaining power, they remain an imperfect and reactive solution. Class actions do not offer the preventive advantages associated with CMOs, such as concluding licensing agreements before scraping occurs, representing rights-holders on a continuous basis, or systematically identifying infringements. They also raise complex questions regarding class membership and the distribution of compensation. In the context of databases, it would be particularly difficult to justify differentiated compensation based on the self-assessed value of individual databases within a single class. If class actions were to remain the primary

[68] *The New York Times Company v Microsoft Corporation et all*, US District Court for the Southern District of New York, No 1:23-cv-11195 (filed 27 December 2023).
[69] Ibid.
[70] *The New York Times Company v Microsoft Corporation et al* (SDNY, 4 April 2025) Opinion of District Judge Hon. Sidney H. Stein, para 106.

enforcement mechanism, scrapers and AI model providers would continue to hold a significant advantage in terms of valuation and negotiation power.

Despite the disadvantages associated with CMOs (such as bureaucratic structures and high operational costs) they provide a more structured and proactive framework for protecting the interests of database rights-holders than reliance on opt-out mechanisms alone. While the AI Act offers rights-holders greater opportunities to enforce their rights independently, CMOs remain better positioned to facilitate licensing, monitor infringements and engage in collective advocacy. If rights-holders are left to rely solely on individual enforcement, protection risks becoming limited to *ex post* litigation, where non-consent clauses and technological measures may prove insufficient. From this perspective, the financial and administrative costs of maintaining CMOs may be justified by the stronger and more consistent representation they offer to database rights-holders.

4.2. **Implications for the AI model market**

The existing lack of legal clarity and collective organisation among database rights-holders in the EU has already led to a situation in which scrapers and AI model providers operating in other jurisdictions, particularly the United States, are able to exploit databases protected under EU copyright and *sui generis* law without providing appropriate remuneration. In practice, rights-holders whose databases are protected within the EU are unable to effectively prevent their works from being scraped and subsequently used by actors based outside the EU, including those operating in one of the most advanced AI development markets globally.

While it is true that the advantages created by this regulatory gap also benefit AI developers within the EU, the overall development of the EU AI model market remains significantly behind that of the United States. In this context, the absence of effective remuneration mechanisms appears even more inequitable. EU-based rights-holders not only see their

databases scraped without compensation, but those databases are often used specifically to support AI development in third countries. As a result, the economic value generated by EU databases is effectively exported outside the EU, while the rights-holders themselves receive no corresponding financial benefit.

The AI Act seeks to address certain cross-border situations by providing that its rules apply to AI model providers located outside the EU, insofar as their AI systems are made available to users within the EU[71]. However, this territorial extension does not fully resolve the underlying inequity. The continued use of EU-protected databases outside the EU, particularly where such use occurs in violation of non-consent declarations or without any form of remuneration, remains problematic from the perspective of rights-holders.

In addition, the limited enforceability of non-consent clauses beyond EU borders raises further concerns. The inability of rights-holders to prevent scraping does not apply only to bodies in the United States, but also to AI model developers operating in other growing markets, including China. While it is in the interests of rights-holders to prevent unlawful scraping in general, the use of EU databases by AI developers in non-EU jurisdictions raises broader issues. Beyond copyright infringement, such practices may result in the large-scale transfer of valuable data to entities operating outside the EU's economic and regulatory framework. In this sense, the issue extends beyond private economic harm and may also raise questions related to strategic interests and national security.

These challenges are particularly relevant for individual rights-holders who rely solely on the opt-out mechanism. In the absence of collective management structures, their ability to negotiate licensing agreements or pursue litigation against foreign companies is extremely limited. Cross-border enforcement is complex, costly and uncertain, especially where

[71] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) [2024] OJ L, art. 2(1)(a).

defendants are located in jurisdictions with different legal standards and enforcement cultures. As a result, individual rights-holders are placed at a clear disadvantage and the lack of collective representation further reduces the likelihood that their rights will be effectively enforced in the global AI model market.

4.3. Considerations for web scrapers and AI model providers

With the adoption of EU legislation imposing substantial obligations on AI model providers, an important question arises as to whether scrapers and AI developers would, in practice, prefer a collective management framework or the currently fragmented system of individual enforcement. Until now, the short-term incentive structure has clearly favoured scrapers and AI model providers, as the absence of effective accountability mechanisms has allowed data to be mined and used without the obligation to compensate rights-holders. However, this situation is set to change with the entry into force of the AI Act, which requires AI model providers to disclose information about the data used to train their models, including databases obtained through scraping.

From a longer-term perspective, a collective management approach appears more advantageous for AI model providers. Under a CMO system, AI developers would be able to negotiate licensing conditions with a single representative entity, rather than engaging separately with numerous individual rights-holders. This would result in considerable time and cost savings and would offer greater legal certainty. By contrast, negotiating individual licences for each database prior to scraping would be highly inefficient and could significantly delay development processes. From this standpoint, the CMO model aligns more closely with the operational needs of AI model providers.

At the same time, the current structure of the EU AI market must be taken into account. The market remains heavily dependent on AI models developed outside the EU, particularly in the United States, and is dominated by a small number of highly influential

companies. In this context, there is a real risk that such companies may, at least initially, fail to comply fully with EU regulatory requirements. While AI model providers that place their products on the EU market have a clear interest in maintaining access to EU users, the possibility cannot be excluded that stricter enforcement of rights-holder claims (such as demands to remove specific databases from training data) could prompt some providers to limit or withdraw their services from the EU market.

It must also be acknowledged that, given the current state of the industry, most data used to train AI models has not been remunerated and AI technology represents a highly strategic and future-oriented sector. As a result, AI model providers may currently hold greater leverage than rights-holders in shaping market outcomes. The EU market itself is in a relatively weaker bargaining position, as it depends on continued access to advanced AI systems.

Nevertheless, enforcement under the AI Act will primarily take place through EU institutions rather than through individual actions by rights-holders. In practice, AI model providers are therefore more likely to face regulatory scrutiny than private litigation. If compliance with EU law becomes too costly - whether due to strict regulatory requirements or the inefficiency of negotiating individual licences before scraping - there is a risk that the EU market could be deprived of access to certain AI technologies. In this context, negotiating licensing conditions through CMOs would be significantly more convenient for AI model providers than dealing with individual rights-holders.

Accordingly, when the interests of scrapers and AI model providers are considered alongside those of rights-holders, the existence of CMOs appears beneficial to all parties involved. CMOs offer a structured, predictable and efficient mechanism for licensing and remuneration, reducing legal uncertainty and facilitating continued participation in the EU AI market.

Conclusions

Taking into account the persistent difficulties in identifying whether databases have been scraped, the limited effectiveness of technological protection measures, the evolving and sometimes uncertain legislative framework, and the necessity to pursue compensation across multiple jurisdictions (often outside the EU) it becomes apparent that the CMO model offers a more favourable solution than reliance on the opt-out mechanism alone. From a practical standpoint, CMOs provide a clearer and more structured framework for addressing the widespread use of databases in data mining and AI model training.

At the same time, it remains essential that rights-holders retain the ability to opt out of collective licensing arrangements. Particularly in light of the current situation faced by EU database rights-holders, collective organisation cannot be viewed as a complete solution in itself. The establishment and effective operation of CMOs presents considerable challenges, especially in the context of databases. These challenges include the need for substantial technological investment to monitor the use of databases, the difficulty of fairly valuing databases that differ significantly in scope and economic relevance, and the administrative burden associated with managing such systems. Moreover, the creation of a single EU-wide CMO for databases appears impractical given these constraints, while CMOs operating solely at the level of individual Member States are likely to lack effectiveness due to high costs, limited scale and coordination difficulties.

It is equally evident that the current mechanisms for remunerating database rights-holders, as well as the existing practices surrounding the scraping and use of databases for AI training, are insufficient to ensure the effective protection of copyright and *sui generis* rights. The existing framework does not adequately reflect the economic value of databases, nor does it provide rights-holders with realistic means of enforcing their rights. As a result,

additional approaches - whether regulatory, organisational or market-based - must be explored to address these shortcomings.

The present situation also creates uncertainty for web scrapers and AI model providers. A lack of legal clarity, combined with the absence of organised remuneration demands from rights-holders, exposes these bodies to the risk of sudden regulatory or legal shifts. Such changes could impose unexpected obligations, threaten the scalability of existing business models or require significant modifications to AI systems. Furthermore, the fragmented nature of rights ownership makes it impractical for scrapers and AI developers to negotiate licences with individual database owners prior to data mining or model training, as the sheer number of rights-holders renders such an approach commercially unviable.

Overall, the current regulatory ambiguity (while temporarily benefiting scrapers and AI model providers by allowing the use of databases without compensation) represents a short-term imbalance that disadvantages rights-holders. In the longer term, particularly with the enforcement of the AI Act, this imbalance is likely to result in increased liability for AI developers, changes in remuneration practices, and potential adjustments to training datasets. For this reason, both scrapers and AI model providers would ultimately benefit from a structured and predictable system that enables them to obtain consent and provide compensation to rights-holders in a lawful and efficient manner.

Future research opportunities

The analysis of both the CMO model and the opt-out alternative demonstrates that the regulation of access to databases remains significantly underdeveloped when compared to the EU's approach to the protection of personal data. In the field of personal data protection, EU legislation and enforcement mechanisms have resulted in a legal framework where the personal data of EU citizens is processed only in ways that are necessary, proportionate, and subject to strict safeguards. Importantly, this protection extends beyond the EU, as entities established in third countries are also required to comply with EU rules when they handle the personal data of EU citizens.

A similar regulatory approach could be explored in relation to the extraction and use of non-personal data, including databases protected by copyright and *sui generis* rights. Applying comparable principles could assist in addressing unresolved issues surrounding the technical measures that rights-holders are expected to implement in order to protect their works. It could also contribute to greater transparency regarding the origin of data used to train AI models. If data originating from the EU were subject to enhanced handling obligations, users of scraped databases would be required to treat such data with a higher level of care, thereby introducing additional safeguards for databases and other protected works.

At the same time, the broader consequences of extending data protection-style mechanisms to non-personal data require careful examination. The legal objectives and underlying rationales of personal data protection and copyright protection are not identical, and the two regimes are not directly interchangeable. Nevertheless, there is clear potential for further research into whether elements of the EU's personal data protection model could inform a more effective framework for safeguarding the rights of database makers. A shift in perspective that considers database protection through a different regulatory lens may

offer valuable insights and contribute to the development of more coherent and balanced solutions in this area.

Bibliography

Treaties and International Instruments:

1. Berne Convention for the Protection of Literary and Artistic Works (adopted 9 September 1886, as amended) 828 UNTS 221
2. Agreement on Trade-Related Aspects of Intellectual Property Rights (adopted 15 April 1994, entered into force 1 January 1995) 1869 UNTS 299
3. WIPO Copyright Treaty (adopted 20 December 1996, entered into force 6 March 2002) 2186 UNTS 38542

European Union Legislation:

1. Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases [1996] OJ L77/20
2. Directive 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
3. Directive 2014/26/EU of the European Parliament and of the Council of 26 February 2014 on collective management of copyright and related rights and multi-territorial licensing of rights in musical works for online use in the internal market [2014] OJ L84/72
4. 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 [2019] OJ L130/92
5. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) [2024] OJ L

European Union Policy Document:

1. Commission of the European Communities, *Green Paper on Copyright and the Challenge of Technology - Copyright Issues Requiring Immediate Action* COM(88) 172 final
2. European Commission, *Green Paper on Copyright and Related Rights in the Information Society* COM(95) 382 final
3. European Commission, *Follow-up to the Green Paper on Copyright and Related Rights in the Information Society* COM(96) 568 final

Books:

1. Quintais JP, *Copyright in the Age of Online Access: Alternative Compensation Systems in EU Copyright Law* (Kluwer Law International 2017). Retrieved on 1 November 2025: https://books.google.lt/books/about/Copyright_in_the_Age_of_Online_Access.html?id=3oyWDwAAQBAJ&redir_esc=y
2. Towse R, *Creativity, Incentive and Reward: An Economic Analysis of Copyright and Culture in the Information Age* (Edward Elgar 2001). Retrieved on 1 November, 2025: https://www.e-elgar.com/shop/gbp/creativity-incentive-and-reward-9781840642544.html?srsltid=AfmBOopw2SxjcWmuEwy2C5lIzVQ0DmwEuwOt42qdS2W4ycBa7YwkL9cS
3. Rosati E, *Copyright in the Digital Single Market: Article-by-Article Commentary to Directive 2019/790* (Oxford University Press 2021). Retrieved on 5 November, 2025: https://www.cambridge.org/core/journals/international-and-comparative-law-quarterly/article/abs/copyright-in-the-digital-single-market-by-eleonora-rosati-oxford-university-press-oxford-2021-491pp-isbn-9780198858591-145-hbk-and-ebk/7C0DD3105F72B33CFCC5DF3C43721170

Contributions to Edited Volumes:

1. Kretschmer M, 'Copyright Societies Do Not Administer Individual Rights: The Incoherence of Institutional Traditions in Germany and the UK' in Ruth Towse (ed), *Copyright in the Cultural Industries* (Edward Elgar 2002). Retrieved on 4 November, 2025: https://ideas.repec.org/h/elg/eechap/2378_9.html

Journal Articles:

1. Caves R, 'Creative Industries - Contracts Between Art and Commerce' (2000) 17(2) *Journal of Economic Perspectives* 73. Retrieved on 11 November, 2025: https://www.jstor.org/stable/3216857
2. Dusollier S, 'The 2019 Directive on Copyright in the Digital Single Market: Some Progress, a Few Bad Choices, and an Overall Failed Ambition' (2020) 57 *Common Market Law Review* 979. Retrieved on 2 November, 2025: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3695839
3. Ducato R and Strowel A, 'Limitations to Text and Data Mining and Consumer Empowerment: Making the Case for a Right to "Machine Legibility"' (2019) 50

IIC 649. Retrieved on 14 November, 2025: https://link.springer.com/article/10.1007/s40319-019-00833-w#citeas

4. Ducato R and Strowel A, 'Ensuring Text and Data Mining: Remaining Issues with the EU Copyright Exceptions and Possible Ways Out' (2021) 43 *European Intellectual Property Review* 322. Retrieved on 14 November, 2025: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3829858
5. Handke C, Quintais JP and Bodó B, 'The Economics of Copyright Compensation Systems for Digital Use' (2013) SERCI Annual Congress. Retrieved on 9 November 2025: https://www.serci.org/congress_documents/2013/Handke.pdf
6. Jerzyk K, 'Balance of Rights in Directive 2019/790 on Copyright in the Digital Single Market – Is the Opt-out Clause Sufficient for the Protection of Authors' Moral Rights?' (2021) 7 *Santander Art and Culture Law Review* 229. Retrieved on 22 November, 2025: https://ejournals.eu/czasopismo/saaclr/artykul/balance-of-rights-in-directive-2019-790-on-copyright-in-the-digital-single-market-is-the-opt-out-clause-sufficient-for-the-protection-of-authors-moral-rights
7. Kelli A, Tavast A and Lindén K, 'The Impact of Copyright and Personal Data Laws on the Creation and Use of Models for Language Technologies' (2020) CLARIN Annual Conference Proceedings. Retrieved on 8 November, 2025: https://www.researchgate.net/publication/342820173_The_Impact_of_Copyright_and_Personal_Data_Laws_on_the_Creation_and_Use_of_Models_for_Language_Technologies
8. Meys R, 'Data Mining under the Directive on Copyright in the Digital Single Market: Are European Database Protection Rules Still Threatening the Development of Artificial Intelligence?' (2020) 69 *GRUR International* 457. Retrieved on 25 Novermber, 2025: https://www.researchgate.net/publication/341088955_Data_Mining_Under_the_Directive_on_Copyright_and_Related_Rights_in_the_Digital_Single_Market_Are_European_Database_Protection_Rules_Still_Threatening_the_Development_of_Artificial_Intelligence
9. Senftleben M, 'Generative AI and Author Remuneration Accepted' (2023) 54 *IIC* 1535. Retrieved on 12 November, 2025: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4478370

Blogs and Policy Papers:

1. Hugenholtz PB, 'The New Copyright Directive: Text and Data Mining (Articles 3 and 4)' (Kluwer Copyright Blog, 2019). Retrieved on 29 November, 2025: https://legalblogs.wolterskluwer.com/copyright-blog/the-new-copyright-directive-text-and-data-mining-articles-3-and-4/
2. Communia, *Using Copyrighted Works for Teaching the Machine*. Policy Paper No 15 (2023). Retrieved on 28 November, 2025: https://communia-association.org/policy-paper/policy-paper-15-on-using-copyrighted-works-for-teaching-the-machine/
3. European Guild for Artificial Intelligence Regulation, *Manifesto for AI Companies Regulation in Europe* (2023). Retrieved on 19 November, 2025: https://www.egair.eu/
4. Initiative Urheberrecht, *Authors and Performers Call for Safeguards around Generative AI in the European AI Act* (2023). Retrieved on 1 December, 2025: https://urheber.info/diskurs/call-for-safeguards-around-generative-ai

Case Law:

European Union:

1. Case C-5/08 *Infopaq International A/S v Danske Dagblades Forening* EU:C:2009:465

National Courts:

1. Hamburg District Court, 27 September 2024, Case No 310 O 227/23

United States:

1. *The New York Times Company v Microsoft Corporation et al* (SDNY, 4 April 2025) Opinion of District Judge Hon. Sidney H. Stein, para 106

Other Sources:

1. The New York Times Company, *Complaint against Microsoft Corporation et al* (US District Court for the Southern District of New York, filed 27 December 2023, No 1:23-cv-11195)