EUIPO
EUROPEAN UNION
INTELLECTUAL PROPERTY OFFICE

www.euipo.europa.eu

EXECUTIVE BRIEFING

The Development of Generative Artificial Intelligence from a Copyright Perspective

May 2025

1. Foreword

In an era marked by rapid technological transformation, copyright remains a cornerstone of Europe's cultural diversity and economic strength. The European Union's creative industries—firmly supported by a robust copyright framework – play a vital role in sustaining employment, fostering innovation, and preserving cultural heritage. Copyright-intensive sectors alone account for more than 17 million jobs and nearly 7% of the EU's GDP, underlining the central role of intellectual property in driving Europe's prosperity and global competitiveness.

Over the past three decades, successive waves of digital innovation have reshaped the way content is created, distributed and accessed. Throughout these transformations, copyright law has adapted to ensure that creators receive recognition and remuneration for their work, thereby sustaining the creative sectors that enrich our societies. However, the emergence of Generative Artificial Intelligence (GenAI) presents unprecedented challenges and opportunities, necessitating a re-evaluation of existing legal frameworks and support mechanisms to address the complexities introduced by this technology.

GenAI is already transforming the way we create, communicate, and innovate. While it offers immense potential as a source of growth and competitiveness in the future, it blurs the existing lines of content creation and introduces a new paradigm where not all content is created by humans. It therefore raises profound questions about how copyright can continue to serve its purpose while supporting innovation. It is essential to find a balance between these two objectives.

GenAI is often described as a black box, with little transparency around its input, functioning and outputs. This makes understanding its impact on copyright even more complex. This evolution prompts critical questions: How does GenAI use copyright-protected content? What is the European Union (EU) legal framework applicable to such use, and how can copyright holders reserve their rights and opt-out content from GenAI training? What are the developing technologies to mark or identify AI-generated content? And finally, what are the opportunities for copyright holders to license the use of their content by GenAI? All questions that need answers if we are to fully understand the development of GenAI from a copyright perspective.

This study is designed to clarify how GenAI systems interact with copyright – technically, legally, and economically. It examines how copyright-protected content is used in training models, what the applicable EU legal framework is, how creators can reserve their rights through opt-out mechanisms, and what technologies exist to mark or identify AI-generated outputs. It also explores licensing opportunities and the potential emergence of a functioning market for AI training data. Although the study is intended for experts in the field, it lays the groundwork for developing clear and accessible informational resources for a broader audience.

Furthermore, this report will provide insights for policymakers to maximise the innovative potential of the EU in light of these new technologies. As the Draghi report on the future of EU competitiveness recently underlined, and as highlighted in the European Commission AI Continent Action Plan, Europe must lead in the digital and AI transformation, not only by investing in infrastructure and skills, but also by shaping the regulatory frameworks that govern emerging technologies. Copyright is a key component of such a framework. It is central to maintaining Europe's capacity to innovate on its own terms—grounded in values of fairness, transparency, and respect for intellectual property.

The EUIPO Strategic Plan 2030 reinforces this vision. It calls on the office to support the strengthening of the IP ecosystem in line with technological developments, such as the rise of GenAI, demonstrating the need for action and new solutions to support both innovation and copyright protection. This study represents an early and important step in meeting that strategic commitment. But it is also a starting point. Much more is needed to guide and support rights holders, AI developers, and policymakers through this fast-changing environment, if we are to realise the full potential of EU digital markets for creators and businesses.

To that end, the EUIPO will launch the Copyright Knowledge Centre by the end of 2025. With regard to GenAI, this new Centre will equip copyright holders with clear, practical information on how their works may be used in the development of GenAI – and how they can effectively manage and protect their intellectual assets. It will also provide a platform for stakeholders, enabling creators, developers, and institutions to share needs, identify gaps, and explore opportunities for collaboration. Drawing on the insights of this study, the Centre will provide a foundation for discussions among experts on how copyright can effectively support content creation and innovation in the GenAI landscape.

It is essential to make copyright rules work in a way that keep human creators in control and ensure their proper remuneration, while allowing AI developers of all sizes to have competitive access to high-quality data. Balancing both interests can be facilitated by simple and effective mechanisms for copyright holders to reserve their rights and the use of their content, as well as licensing and mediation mechanisms to facilitate the conclusion of license agreements with AI developers. As GenAI applications and markets mature, further reflections might also be needed on whether content generated by AI deserves protection through existing or new intellectual property rights.

At the EUIPO, we stand ready to play our part. By working in close cooperation with European and international institutions to contribute our expertise on IP protection and awareness, and in the development of technical solutions and mediation services to help ensure that, as with earlier digital innovation cycles, copyright keep supporting creators and technological progress.

2. Executive Overview

2.1 Background on the development of Generative Artificial Intelligence

Over the past several years Artificial Intelligence (AI) technologies have experienced major advances, with the release of Large Language Models (LLMs) and GenAI systems, allowing end-users to generate synthetic text, code, image, audio and video content. This has led policymakers and regulators to examine how existing legal frameworks should evolve to address the implications of large-scale AI adoption, and to balance innovation with intellectual property (IP) protection.

In this context, the study explores the developments in GenAI from the perspective of EU copyright law. It analyses the (1) **technical, legal and economic** aspects of GenAI development, as well as copyright-related issues regarding the (2) **use of copyright-protected content to develop GenAI services** and the subsequent (3) **generation of content**.

The study offers an in-depth technical and legal analysis on solutions underlying the effective implementation of EU laws, focusing on existing and developing solutions for:

- Copyright holders to reserve their rights from use by AI developers,
- AI developers to ensure that the content generated through their services is detectable in a machine-readable format.

2.1.1 Technical background

GenAI systems draw insights from large quantities of training data, used as **input** to develop algorithmic processes which can generate new content with similar characteristics as **output**.

GenAI input: The data collection process is just the first stage in GenAI training. Collected data must then be cleaned, annotated, and processed before it is used in AI training, which consist of multiple stages from **model pre-training** to **model fine-tuning,** to optimise the execution of specific tasks. Additional input data is often needed for the final phase of the training process, named **reinforcement learning**.

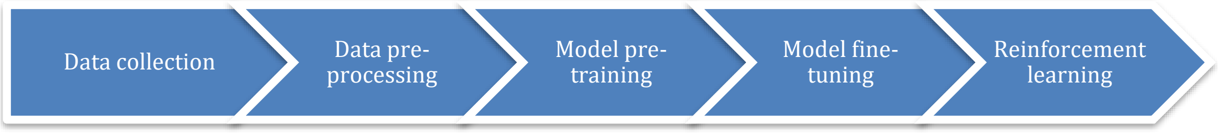

Figure 1: Main components of the GenAI training process

Early developments of AI systems were generally based on carefully sourced, curated, and labelled datasets. However, the evolution of AI technologies has given rise to **demands for increasingly large training datasets**. Massive quantities of data available online are now collected for the development of AI training datasets through a process known as **web scraping.** It consists of extracting data from websites by retrieving their content and analysing it, to obtain the desired information.

GenAI output: The technical process of content generation depends on the type of GenAI model, and the types of content they generate. Given the high costs of training AI models and the limitations of constantly (re)training models on new content, **Real-time Augmented Generation (RAG) technologies** are increasingly deployed. They consist of combining GenAI with information retrieval. For example, **answer engines** generate answers by searching, identifying and synthesising up-to-date information available online. With RAG, copyright-protected content is not only used for training, but also for content generation purposes.

2.1.2 Regulatory and legal background

In the EU, two legal instruments are particularly relevant for framing the implications of GenAI developments from a copyright perspective: the **Copyright in the Single Market Directive (CDSM)** and the **EU Artificial Intelligence Act (AI Act).**

The CDSM establishes a legal framework for **Text and Data Mining** (TDM), by providing exceptions to the copyright holders exclusive right of reproduction for TDM activities for scientific research (Article 3), and any others (Article 4) purposes.

Most importantly, Article 4 allows copyright holders to **reserve their exclusive reproduction rights**, which is commonly referred to as **'opting-out'** of the TDM exception. For content that is publicly available online, such reservation must be **made expressly**, **by the right holder** in an **appropriate manner**, including by **‘machine-readable means’**.

When an opt-out reservation has been expressed, AI developers need an authorisation by the right holder to use their content, for example through licensing agreement. This means that **effective solutions for TDM opt-out are needed for a market for content licensing to develop**.

The application of Article 4 is central to GenAI compliance with copyright in the EU, as the collection and analysis of online content for the development of AI training datasets without prior authorisation of copyright holders has become common practice.

The **AI Act** sets out a regulatory framework for AI technologies in the EU, with specific obligations on the providers of general–purpose AI (GPAI) models. These obligations refer to **compliance with the TDM opt-outs expressed by copyright holders**. GPAI system providers are also required to **publish ‘sufficiently’ detailed summaries of the training data** they utilise, to facilitate copyright holders enforcing their rights. The AI Act also places obligations on the deployers of GenAI systems to **ensure that generative output is detectable** in a machine-readable format.

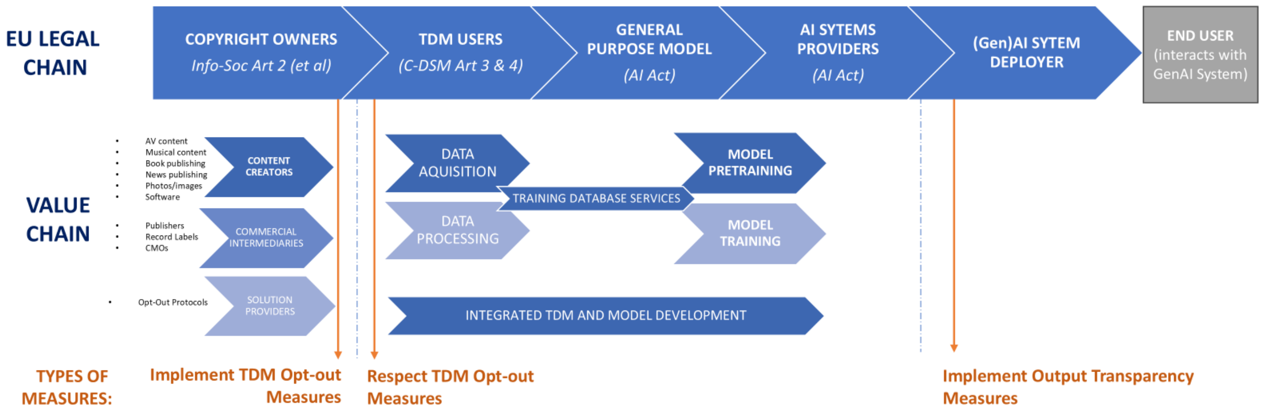

Figure 2: Simplified mapping of GenAI value and legal chains in the EU

The development of GenAI has given rise to several legal disputes between rights holders and GenAI system providers. These lawsuits have mainly arisen in the USA so far, but cases have also been brought forward in China, Canada, UK, India and the EU. They **generally focus on whether copyright limitations and exceptions apply in the context of AI training**, and the circumstances under which restrictions expressed by rights holders must be respected. In the EU, **three cases have been filed in Germany and one in France.**

2.1.3 Economic background

In parallel to litigations, several agreements on the use of copyright protected content for AI training have been agreed between rights holders and GenAI developers. The study highlights several factors driving such agreements. These include GenAI developers' concerns about a **potential shortage of data** to train their models, their **need for high-quality content** with good metadata and annotation, or their level of **risk tolerance** and relative negotiating power.

If the opportunities for direct licensing markets differ between content sectors, the press and scientific publishing sectors seem to be uniquely positioned to take advantage of **licensing opportunities associated with RAG** applications that are central to the development of some GenAI services (e.g. answer engines). The emergence of content aggregation services as commercial intermediaries developing training datasets also presents opportunities for smaller rights holders to license their content for AI training.

The study identifies several factors that may affect the evolution of **licensing terms** as the direct licensing market develops. These include the establishment of **benchmark market rates,** the **basis for remuneration** of copyright holders, and whether they are linked to the value of their content for training purposes and/or to the revenues of the GenAI services.

The direct licensing market has the potential to bring new revenues streams to creators and the creative sectors. However, a pre-requisite for such market to develop is the capacity for copyright holders to effectively express their TDM rights reservations

2.2 Generative Artificial Intelligence Input

The practice of collecting large amounts of online data for AI training through web scraping has led many copyright holders to focus on opt-out measures for this practice.

The **Robots Exclusion Protocol** (REP) currently serves as a *de facto* standard for managing web crawling and scraping activities and has been deployed as a primary strategy for TDM rights reservations. However, there is a prevailing consensus amongst stakeholders that it is not optimal as a TDM opt-out mechanism and serves more as a temporary solution. This is mainly due to REP's

limited granularity and use-specificity, its need for intermediation by website managers, unenforceability, and the voluntary disclosure of web-scrapers identities. In that respect, REP is also sometimes complemented by traffic management strategies for restricting web-scrapers access to online content in the first place.

Given the complexity of the AI ecosystem, and the specific needs and business models of different content sectors, **no single opt-out mechanism** has emerged as the sole standard used by rights holders. Instead, a number of **legally-driven measures** and **technical measures** are used by rights holders to express their TDM rights reservations.

The legally-driven measures for rights reservations include unilateral declarations by copyright holders, licensing constraints, and website terms and conditions. Beyond REP, the technical measures for rights reservations include solutions specifically developed to address TDM opt-out (e.g. TDM reservation protocols), solutions being adapted to serve such purposes (e.g. C2PA Content Authenticity Initiatives), or solutions that are still under development and aim at creating an infrastructure to manage copyright online more broadly (e.g. Liccium Trust Engine Infrastructure or Valuenode's Open Rights Data Exchange platform).

The study compares such measures in relation to different criteria to highlight their **respective advantages and limitations** in supporting TDM reservations and direct licensing.

In general, the reservation measures analysed are just informative and none of them support the actual enforcement of TDM reservation expressed. Legally-driven measures are typically applied to specific copyright-protected work, but also entire repertoire of works. Technically-driven measures are categorised as either **'location-based'** (i.e. associated to the location of a piece of content online) or **'asset-based'** (i.e. associated with the actual content irrespective of where it is available online). Both approaches have their distinct advantages and limitations.

The diversity of measures is evidenced by the feedback from stakeholder interviews, which indicate that their content management and rights reservation strategies frequently employ a combination of legally-driven and technical measures. Interviewed stakeholders supported increased efforts for **standardisation of rights reservation measures**, as well as the **flexibility to incorporate multiple measures** to adapt to different use cases. As the GenAI ecosystem keeps evolving, standard practices are also expected to emerge to adapt to the specific needs of different content sectors and GenAI developers.

The current situation regarding rights reservation measures suggests a **role for public authorities**, such as national IP offices and the EUIPO. **Technical support** may consist of implementing and administering federated databases of TDM reservations expressed by right holders. **Non-technical support** may consist of increasing public awareness of the copyright issues surrounding the deployment and use of GenAI technologies, providing information on various rights reservation measures, and analysing industry trends in terms of technical developments and commercial licensing terms.

2.3 Generative Artificial Intelligence Output

The AI Act requires transparency on the content produced by GenAI systems, and the study analyses a number of existing and developing solutions to **identify and disclose the nature of synthetic content**.

These **generative transparency measures** include **provenance tracking, detection measures** for AI-Generated content, as well as **content processing solutions** that allow the re-identification of content. The study compares a selection of these **generative transparency measures** in relation to different criteria to highlight their **respective advantages and limitations** to identify or detect synthetic content.

GenAI system providers are also developing solutions to mitigate the risk of their systems generating copyright infringing content. These include tools to **compare generated content** with potential input sources, **filters for preventing duplicative output**, and different approaches to **prompt rewriting or filtering**. **'Model unlearning'** and **'model editing'** is used to erase, adjust or update the information coded into the model's parameters, enabling AI developers to solve issues detected after the model's deployment. Several GenAI system providers also offer some form of **legal indemnification** to mitigate the risk for their customers.

The issues surrounding GenAI outputs and copyright also suggests a potential **role for public institutions** active in the field of IP. **On information for GenAI developers and policy makers,** they could openly share information on measures available to mitigate potential infringing output, detect synthetic content, and good practices developing in that field. **On information for the general public**, they could provide information on ethical prompt usage and cooperate with other relevant bodies to increase the public's capacity to identify generative output. Public institutions could also serve as **forums for technical information sharing** and collaboration supporting the interoperability of output transparency measures across platforms and GenAI systems.

3. Concluding observations

The study shows that as GenAI models and services evolve, access to high-quality and recent content is becoming more important not only for fine-tuning models, but also for content generation using RAG techniques. There are indications that while most GenAI developers' source and use content available online without prior authorisation of copyright holders, a market for direct licensing is slowly emerging. In the EU, developing this market requires **effective opt-out solutions for copyright holders to reserve their rights**, with all the solutions analysed having their respective advantages and limitations.

The study also highlights the importance of accurate information, firstly about a work's origin to identify its right(s) holder(s), secondly about permissible uses to see if copyright protected works can be used by GenAI services, and thirdly to identify content that has been created by AI. This transparency has an impact on effective application and enforcement of copyright from the side of rights holders and AI developers alike.

Most importantly, the study demonstrates the complexity of the technical and legal challenges raised by the development of GenAI from a copyright perspective for the creative and cultural sectors, but also for AI developers. The launch of the **EUIPO Copyright Knowledge Centre** by the end of 2025, will be an opportunity to address this complexity, and develop comprehensive information resources for copyright holders to understand how their content may be used by GenAI and the solutions at their disposal to reserve their rights.

The study also demonstrates the need for AI developers to engage more actively with the creative and cultural sectors on effective ways to account for the TDM opt-out expressed. In that respect, the EUIPO Copyright Knowledge Centre should also provide a platform for discussions on ways to support simple and effective mechanisms for copyright holders to reserve their rights, and for AI developers to comply with such reservations. It should also support discussions on **licensing and mediation mechanisms to facilitate the conclusion of license agreements with AI developers**.

The EUIPO Copyright Knowledge Centre should also provide a platform to raise awareness on measures and good practices to mitigate the risks of GenAI output infringing copyright. Finally, as the study shows that there may be an economic interest in protecting AI-generated output in certain cases, this should also be a platform to **discuss the ways such content may be protected by intellectual property rights**.

EXECUTIVE BRIEFING –

The development of Generative Artificial Intelligence from a Copyright Perspective