Council of the European Union
General Secretariat

Brussels, 08 March 2024

WK 3727/2024 INIT

LIMITE

PI

This is a paper intended for a specific community of recipients. Handling and further distribution are under the sole responsibility of community members.

INFORMATION

From: To:	the Finnish Delegation Working Party on Intellectual Property (Copyright)
N° prev. doc.:	WK 1805/2024
Subject:	Copyright Infrastructure Task Force – Defining the AI & Copyright use case (Draft)

Delegations will find attached the draft document on the above-mentioned subject, as presented by the Finnish delegation at the Working Party on Intellectual Property (Copyright) on 8 March 2024, under AOB (CM 1928/24).

DEFINING THE AI & COPYRIGHT USE CASE

1. Background

The purpose of this document is to describe legal aspects of the use case to at hand of the Open Rights Data Framework (ODRF) by a Use Case Group (UCG) proposed by the Copyright Infrastructure Task Force (the CI task Force) to Europeum – a European Digital Infrastructure Consortium (EDIC) in formation.

The latest draft implementation strategy published by the CI Task Force on 9 January 2024 proposed to conduct a use case on AI & Copyright. The document states that –

"Considering –

- the need and urgency to equip the creative industries with adequate tools to face challenges arising from Large Language Models and Generative AI, and
- the current EBSI work, whereby Intellectual Property is a use case for the Pre-Commercial Procurement and an Open Rights Data Exchange is an application scenario for TRACE4EU, the EBSI traceability project,

the CI Task Force suggest leveraging an AI & Copyright Use Case to –

- launch its operations in 2024,
- conduct a Large-Scale Pilot, and
- work closely with the EDIC Europeum while becoming one of its Use Case Group".

The proposition states that in specifying the details of the use case (piloting the ORDF on a large scale), legal experts of the CI Task Force and the Commission will analyse the key issues and recommend wordings in early 2024. These wordings are needed to be subsequently implemented by engineers. This document has been discussed at three CI Task Force meetings
30 January and 27 February.

The progress has been reported at the Council Working Party on Copyright and will be discussed with stakeholders during the Belgian Presidency of the European Union (EU), among others at the

Copyright Conference in Namur. To be effective and complete the ORDF needs to cater for various models in a global context.

2. Legal considerations

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

The use of copyright protected works and other subject matter for reproduction or communication to the public requires the authorization of the rightsholder concerned unless relevant copyright exceptions and limitations apply.

Articles 3 and 4 of the Directive (EU) 2019/790 (the DSM Directive) provide for two exceptions to the exclusive right of reproduction under copyright and the right to prevent extraction under the database right when those acts are committed for the purpose of text and data mining (TDM) as defined in Article 2(2) DSM Directive. The copyright on protected works and the related rights on other subject matter (including databases and press publications) are subject to those exceptions. The directive has been implemented into national law by almost all EU Member States.

The definition in Article 2 of DSM on text and data mining is *"text and data mining' means any automated analytical technique aimed at analysing text and data in digital form in order to generate information which includes but is not limited to patterns, trends and correlations"*.

Under Article 3, TDM is allowed for scientific research by research organisations and cultural heritage institutions under the condition of lawful access to the protected work or subject matter.

Under Article 4, TDM for any other purpose, including for commercial purpose, is allowed when the works and other subject matter are lawfully accessible. According to Article 4(3) the exception applies on condition that the use "*has not been expressly reserved by their rightsholders in an appropriate manner, such as machine-readable means in the case of content made publicly available online*". This is referred to as a possibility to make an "opt-out" from the application of the exception or limitation.

Especially since the flush of generative AI tools made available online, rightsholders feel the need to reserve their rights and to opt-out of the exception in certain situations. The DSM Directive requires the reservation to be "expressly reserved by their rightsholders in an appropriate manner". Other than the reference to machine-readable means for online content, the provision does not detail how the opt-out should be implemented. Recital 18 of directive (EU) 2019/790 states: "In the case of content that has been made publicly available online, it should only be considered appropriate to reserve those rights by the use of machine-readable means, including metadata and terms and conditions of a website or a service. [...] .

The December 2023 political agreement on the Artificial Intelligence Act (AI Act) provides for directly applicable provisions in relation to copyright protected works. The Act was approved by

the Council on 2 February, currently waiting for approval by the European Parliament. The Act normally comes into force after a 12 month-period of its adoption, *i.e.* sometime in 2025.

According to the AI Act providers of general purpose AI (GPAI) models shall "*put in place a policy to respect Union copyright, in particular to identify and respect, including through state of the art technologies, the reservations of rights expressed pursuant to Article 4(3)*" of the DSM Directive. "State of the art technologies" could be considered to mean blockchain or distributed ledger technology, which the EBSI – European Blockchain Service Infrastructure provides for.

Furthermore, providers of GPAI models must "*draw up and make publicly available a sufficiently detailed summary about the content used for training of the GPAI model, according to a template to be provided by the AI Office*". The summary obligation should be generally comprehensive in its scope (instead of technically detailed), and thus include the general, main content sources/sets for training the GPAI in order to facilitate parties with legitimate interests (including copyright holders) in exercising and enforcing their rights. This could for instance be achieved by listing the main private or public databases/archives used and by providing an account of the other sources used. The AI Office is expected to provide a template for the summary, which should be simple, effective and allow to provide the required summary in a narrative form. The AI Office was established by a Commission decision of 24 January 2024.

2.2. Material scope of the AI & copyright use case

Text and data mining techniques may be used extensively for the retrieval and analysis of various types of content, which may be protected by copyright and related rights. Training of AI models require vast amounts of data. Most of the data composing works and other content is protected by copyright. The TDM exception is one important way to ensure that AI models can be trained with copyright protected content. The permission to use the content is provided through a provision in the law, based on certain caveats.

The AI & copyright use case (later "AI use case") will focus on the first of the obligations for the GPAI *i.e.* the identification and respect of the opt-out made by the rightsholders with regard to the and data mining, unless this is done for the purposes of scientific research. Ideally, the ORDF could be developed and piloted at a later stage with regard to the summary requirement or any other use case in relation to the proper functioning of the copyright system in the digital environment such as Art 17(4) or Art 8(4) of the DSM Directive.

The AI Act does not specifically provide that a new standard should be developed, but rather "a policy to respect Union law on copyright and related rights, in particular to identify and respect the reservations of rights expressed by rightholders pursuant to Article 4(3) of Directive (EU) 2019/790". However, in practice, in order to comply with the opt-out mechanism a standard needs to be set on EU level. This is proposed to be developed through an open rights data framework. This would also be very important from legal certainty as well as growth and competitiveness from point of view of the European companies in training of AI.

Piloting the opt-out mechanism is a complex matter.

For example, in order to know if copyright applies in the first place and whether there is a need to seek permission, the mechanism would need to establish the "copyright status". In implementing these rules (for instance under the TDM.AI-model) this would be "true": "in copyright" or "false" out of copyright or in public domain). Status should, ideally, be based on the date of death, if applicable, linked with the term of expiration of the rights.

Only then some rights apply and can be managed by the rightsholder in accordance with Article 7 of Directive 2001/29/EC[1].

Terms and/or exceptions must also relate to a specific right, in this case the right to reproduction. This could translate to "reproduction" true: yes/false: no. Then an exception to the exclusive right could be determined; "TDM" true: yes/false: no TDM; The same would go for context "commercial" (true: yes and false: no) and finally the opt-out use based on exception "AI training" true: yes and false: no.

Taking all these elements into consideration shows how the use case is complex.

While legal aspects should not make the piloting of the opt-out overly complex, neither should the use case be conducted merely as technical endeavor. In developing the requirements for the use case certain key aspects are presented below as basis for discussion at the CI Task Force meeting on 26 March (placeholder). Please find below some considerations on the scope of the use case.

First of all, the use case should focus on the opt-out that is made in a machine-readable way. For online content, the opt-out must be expressed in a machine-readable way. This means that the opt-out is a statement included in the metadata of the work that can be read by the search robot when searching for content to train an AI model.

Secondly, the use case will pilot the right to reserve TDM for other purposes than scientific purposes.

The initial purpose of the TDM exception in the proposal of the Commission was to allow the analysis of large masses of data for the purpose of scientific research. Only later it was extended to also cover other purposes, such as TDM for for commercial purposes, including training of AI. Although the exception could cover instances of training of GPAI, it is not well suited for this purpose. This is particularly true with regard to generative AI systems, as these systems produce new content (and revenue) based on the content they were trained with.

Thirdly, the opt-out should be piloted keeping in mind the scope of the TDM provision itself, taking into account the so-called three-step-test.

[1] Article 7. Obligations concerning electronic rights management information. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32001L0029

Therefore, the opt-out use case should be limited to cases that are likely to be in line with the requirements under Art 7(2) of the DSM directive[2] when they were implemented[3]. In accordance with Article 5(5) of the Information society directive 2001/29/EC, use based on a limitation is allowed only in "*certain special cases which do not conflict with a normal exploitation of the work or other subject-matter and do not unreasonably prejudice the legitimate interests of the rightsholder*".

Hence, there are two cases in which an opt-out would not be relevant, even if it is expressed in a machine-readable way.

One would be the case where the conditions for the exception for scientific research are met and no opt-outs are allowed. This is confirmed also under Art 4(4), which states that the exception in Art 4 must not impact the application of TDM under Art 3. Another case is when a) the training of the AI could not be considered text and data mining (under Art 4) or b) when the use is otherwise considered not to pass the three-step test.

Several court cases are pending that could clarify the application of copyright, including one in Germany[4] regarding the application of the TDM exception for training of AI models that are used for offering generative AI systems. Such systems normally require the consent of the authors.

Consequently, the use case would normally explore how the opt-out could technically be made in a way that takes into consideration not only the opt-out but also the purpose of the use and its lawfulness under the exception.

The user does not need to be identified, only the purpose. Anyone can apply the TDM provision under Art. 4(3), subject to the opt-out by the rightsholders and in the case of GPAI models, subject to the identification and respect of the reservations of rights. However, it is a key element of the data economy and the use of new technologies that each entity participating in the sharing of data, is also identified with an open and interoperable identifier.

2.3. Geographical scope of the pilot

The use case will focus on applying the opt-out on uses of works for training an AI model in the EEA. The 4 to 6 EU Member States (MS) which would participate in the pilot must have transposed the DSM Directive in their national copyright act. The copyright law of a Member State applies to the reproduction of a work, taking place in its national territory.

[2] Art 7.2 (2): Article 5(5) of Directive 2001/29/EC shall apply to the exceptions and limitations provided for under this Title. The first, third and fifth subparagraphs of Article 6(4) of Directive 2001/29/EC shall apply to Articles 3 to 6 of this Directive.

[3] In the doctrine interpreting the three step the conformity with the three steps test is assessed at adoption of the provision providing legal certainty for the user of the exception. If the threshold would be assessed in individual cases, the application of it would not be serving the very purpose of the exception itself.

[4] Some development in this regard is taking place; see https://cepic.org/news/an-up-date-on-the-robert-kneschke-v-laion-e-v

MS could include Finland, Estonia, and 2 members of the Europeum EDIC, and possibly also Latvia and Lithuania, subject to confirmation by each MS that they want to participate in the use case. Relevant provisions in the respective copyright acts are referred to below.

For the obligation to respect the opt-out to be effective, however, it would be important to apply it regardless of the jurisdiction in which the copyright-relevant acts take place. Training of AI *i.e.* the reproductions for that purpose, are likely to first take place outside the EU (at least for the models developed by the leading US actors).

There are no harmonised rules on training of AI at international level and the issue remains open in many jurisdictions.

2.4. Creative sectors best suited for the pilot

Selecting the creative sector to be piloted with the ORDF is also important. Keeping in mind the previously expressed requirements, there are several good candidates within the creative sectors for an AI & copyright use case.

The **book-publishing sector** has been active in developing the copyright infrastructure and there is a high percentage of open and interoperable identifiers (ISBN, ISSN, DOI, ISNI, ISCC) in use in this sector. The book industry represents European stories and is of immense value for the European Cultural heritage. The new mechanism to assist accessibility of out of commerce works online (out of commerce works portal) was promoted through the decision on the Cultural Heritage Data Space. According to preamble 17, Europeana has been key in strengthening cooperation and standardisation activities across borders, in the EU and beyond. Its standardised frameworks for sharing digital content and metadata online, in particular, the Europeana Data Model, Rights Statements and the Europeana Publishing Framework should be used to ensure access and to improve interoperability of works made available on Europeana. The open availability of works online could on the other hand make European cultural heritage still in-copyright vulnerable for reproduction for training of AI without a possibility to opt out. The ORDF is key in ensuring that these works would be safe from TDM for commercial purposes, still allowing TDM for scientific purposes.

We could also envisage a specific sub-sector of the publishing industry, like the scientific publishing sector. The **scientific journals** may have several identifiers. Some are identified with the ISNI, some with the ORCiD identifiers. The newly introduced, DLT enabled ISCC identifier could work very well for identifying works that have several versions such as articles and journals. These journals are valuable training material to LLM's focusing on the production of new content based on that information, such as ChatGPT more focused on re-use of data than on the content itself. This would be an interesting area as scientific articles would be largely used for research purposes but also for training of GPAI. On the other hand, scientific authors are under obligations to go for Open Access publications and make their papers available online (on university repositories for example that are used for TDM). Most uses of the repositories will be for scientific research (not commercial). The opt-out in this sector is not likely to happen by

authors. It would also not be consistent with the purpose for open access to promote making opt-outs by the publishers in this field.

The **photographic industry** has been very vulnerable in the digital environment as to infringing use of images made available online and also stripping of metadata indicating the intent to attest copyright. Photographs have normally only one author. Photographers have also chosen to use commercial image banks or open source models to promote the use of their works. AI models like the ones providing new images based on existing images for instance the purpose of advertising cause significant harm to the industry. What kind of AI model would use photographs to train it but would also be justifiable under the TDM exception from the perspective of the rightsholders? The DSM Directive clarifies the status of works of visual art in the public domain with the objective to increase legal certainty. This means that for photographs, a point to consider is whether it would be necessary to ensure that metadata also captures the possibility of providing status of copyright (author is alive, or death date) including if it is in public domain. One interesting factor to consider is the benefit of the use case would have on the emergence of fake news and dis- and misinformation.

Another possibility for a specific sector is **video games.** There the rights are kept in one place, with the developers. On the other hand, the video games are not identified with an open an interoperable identifier.

The **music sector** would also be an interesting area for the use case. Text and data mining is made also on music. AI applications are used by a broad array of users, including for commercial purposes. However, considering the complexity of the use case on an opt-out as such, the additional complexities could risk slowing down the conducting of the use case. The sheer number of rightholders concerned (at least one author, publisher, performer, producer) and the availability to various partly overlapping identifiers that are not fully interoperable, may also limit the benefit of the use case in the larger context.

Let's continue to discuss! Written comments are welcome! Please send them to anna.vuopala@gov.fi by 22 March 2024.