Matt Rogerson
Director of Global Public Policy & Platform Strategy
Financial Times
Bracken House
1 Friday Street
London EC4M 9BT

Baroness Stowell
Chair
Communications and Digital Committee
House of Lords
London SW1A 0PW

18th October 2024

Dear Baroness Stowell,

Follow up to future of news inquiry oral evidence session on 15th October

I write to thank you and the Committee again for your interest and inquiries into the future of news. I write specifically to address a point that was raised specifically in the evidence session on the 15th October. During that session, the Minister for Sport, Media, Civil Society and Youth suggested that the Financial Times (FT) had a commercial licensing agreement in place with Google for the use of its intellectual property (IP) Google's suite of LLMs and related AI products. I write to be clear for the record that this is not the case.

The FT has announced a licensing agreement with OpenAI in April 2024, and has forged a number of other agreements for innovative AI related partnerships since then. For example, in August 2024, the FT agreed to with the ProrataAI on a private beta of an innovative IP marketplace that enables AI developers to accurately attribute and share revenues on a per-use basis with content owners.

Both of these agreements begin to align the incentives of AI platforms and publishers in the interests of quality journalism, the reader and respect for IP. We strongly believe that sharing revenues between technology companies that use IP and the publishers that create it – can help develop a healthier and fairer information ecosystem that encourages accurate and authoritative journalism and rightly rewards those who produce it.

The goal of aligning incentives between publishers is, however, being undermined by the scraping practices of incumbent technology companies. While Google has activated a second Google-extended scraper that allows website owners to opt-out of IP being used to train a number of Google LLMs, the presence of the existing Google search scraper on almost the vast majority of websites on the open web enables Google unparalleled access to IP published online. This IP has traditionally been used to appear within Search Engine Results Pages (SERPs) from where users click to source websites to engage directly with the source website. That IP is now being used to enable Google's LLMs, and those of third party companies such as Meta, to respond accurately to user queries in real-time.

This capability, known as Retrieval Augmentation Generation (RAG) when used by Google and sold to companies like Meta, means that those companies extract commercial value from the source material, without a user ever engaging with the source of that information.

From Wikipedia to the Watford Observer, websites rely on engagement with users: engagement that is generated by the content invested in and generated by those sites. Without such engagement the ability to generate any of those revenue streams disappears. *This* was the social contract of the open web, that value would be shared between search and social gateways and the investors in intellectual property.

As recently as 2020, Google suggested its role was *"Sending people to publishers' news sites—not keeping them "walled" up on Google products, as some claim—is a key way we provide value to the news industry. Every month we send Google users to news sites 24 billion times, providing an opportunity for publishers to grow their audiences and show Google's users ads or offers for subscriptions."* In the age of generative AI responses, both within Google's environment and the third-party LLMs to whom Google sells data generated via the search crawler, this social contract has been broken.

I understand there is presently no option to prevent the use and sale of IP for RAG purposes from occurring, other than to opt-out of Google search entirely. This leaves website owners with an unenviable choice. To opt-out of the Google Search crawler entirely, and become invisible to the 90%+ of the UK population that currently uses Google Search, or allow scraping to continue in ways that both extract value without compensation, and undermine nascent commercial licensing markets for the use of high quality IP to build and enable the AI models of the future.

As the United States government's Department of Justice outlines in its proposed remedies framework submission to the court case UNITED STATES OF AMERICA v. GOOGLE LLC, the *"harms of Google's conduct also extend to the generation and display of new and developing features of general search, such as generative artificial intelligence (including on- device artificial-intelligence functionality) and retrieval-augmented-generation-based tools. These results and features often rely on websites and other content created by third parties, who have little-to-no bargaining power against Google's monopoly and who cannot risk retaliation or exclusion from Google. Google's ability to leverage its monopoly power to feed artificial intelligence features is an emerging barrier to competition and risks further entrenching Google's dominance."*

It is easy to be blinded by the light of new forms of AI. But taking a cold hard look at how these models are brought to life should make one realise that rather than weakening IP owners' position in relation to the enforcement of IP rights, it is essential that UK regulators and the UK government empower IP owners to make choices that create new markets that deliver much needed competition, innovation and growth. Individual creators of IP, local news organisations documenting life in their communities, or global news organisations producing the highest quality financial journalism, should be in a position to exercise genuine choice.

In addition to the points of clarification above, I attach to this letter by way of Appendix, a note prepared by the European Publishers Council for the European Commission on the topic of the enforceability of a text and data mining opt-out regime. The note contains the experience of publishers from across the EU who have sought to enforce the TDM opt-out, which I thought may be useful for the Committee to consider.

Yours sincerely

Matt Rogerson
Director of Global Public Policy and Platform Strategy
Financial Times

EPC European Publishers Council

4 December 2023

EPC feedback on TDM and Article 4 of the DSM Directive

Introduction:

It is well known that AI development relies on access to large quantities of training data, the vast majority of which has been scraped from the internet, and much of which is protected by copyright. Access to high-quality, well-structured content is necessary for the creation of high-quality AI. Many AI developers do not disclose information about the data and content used to train AI and this untransparent situation creates problems for the rightsholders, such as publishers, but also exposes much wider questions of trust in AI for society more generally.

Publishers typically own large libraries of content and data, produced by them or licensed in, which is made available to the public online or via apps: either free at the point of access to the public, financed by advertising, or via various transactional or subscription models where the content is behind a partial or full paywall. **In all cases, regardless of the business model, all rights are reserved and benefit from exclusive rights of protection**.

In addition to making content available to read, content owners also allow their content to be indexed, especially by very large search engines, but all rights for additional or separate uses remain reserved, meaning that separate authorisation is required for each different use. Just because Google search robots have access to publishers' content for indexing does not give them lawful access to use that same content to train their AI. It is important that this is well understood as content made available to read for free by the public, is not freely available for reuse by third parties for their own commercial purposes without authorisation, unless covered by an exception, and in which case it must not breach the 'three step test' as per the InfoSoc Directive. Here I would refer you to Eleonora Rosati's recently published paper, as well as our own legal advice which was previously shared with you.

Under article 4 of the DSM copyright directive rightholders can reserve their rights by means of an opt-out, by any means, including in machine readable form. Once the opt-out is exercised, the exception is no longer valid. However, even though publishers are reserving their rights, there are three clear problems when it comes to relying on the TDM exception for training AI:

1. Much of the training already took place before the Directive came into force during which there was no applicable exception and was in clear violation of publishers' exclusive rights under the InfoSoc directive.

www.epceurope.eu

EPC
Chairman > Christian Van Thillo
chairman@epceurope.eu

Executive Director > Angela C. Mills Wade
angela.mills-wade@epceurope.eu

EPC > Avenue Des Arts 43, 5th floor, 1040, Brussels, Belgium | Tel +32 (0) 2231 12 99

2. There is no standard, enforceable way to reserve rights (full details below). There is no technical solution whatsoever that technically limits TDM. The technological measures that are available are based on specific user agents, and not purposes. Robots.txt works as a statement of intention ("code of conduct") for robots and crawlers, mostly applicable to SEO (search engine optimisation). Robots.txt is used by publishers, but robots.txt was not created to fulfill such a legal purpose of a TDM opt-out. Other measures, such as meta tags in code, have the same weaknesses. AI companies and their agents should be required to be transparent about the bots they use and for which purpose. Then it would be easier for publishers to accept or reject them.

3. The Directive was drafted back in 2016 as part of delivering the EU's 2015 Digital Single Market Strategy, when the aim of providing for a TDM exception was to strengthen the position of the EU by creating legal certainty. At the time our understanding of TDM, and we assume that of the European Commission, was the computational process of discovering and extracting knowledge from structured or unstructured data which is quite different from our understanding of the full extent of the capabilities of generative AI and the nature and scale of unauthorised acquisition of publishers' content to develop it.

Therefore, regardless of whether or not publishers have reserved all their rights, use robots.txt or metatags, the use of TDM techniques for the purpose of training AI, the creation of Large Language Models and generative tools cannot not be covered by the exception due to conflict with the conditions of 'three-step test' which must be met for the exception to apply. Consequently, there is a sound legal basis for requiring AI developers wishing to use news media content for AI training purposes to secure a licence.

Some practical questions have emerged recently due to the on-going debate in the framework of the AI Act. We have asked our members to comment on these and we believe the answers below are revealing of the current situation that publishers are confronted with. We can update these and add more questions in future

1. TDM / AI - identifying those who access publisher content for research or commercial purposes

"*Research TDM*" is use of data in order to generate new knowledge by recognising patterns or identifying trends, cf. also Recital 8 of the DSM Directive.

"*Commercial TDM*" is where data are used as the foundation for the development of AI systems that can be used to create images, books, films, news, etc. that compete with the rightholders (generative AI). Such use is clearly covered by rightholders' copyright. **The opposite would disrupt the normal exploitation of the works and unreasonably impair the legitimate interest of the rightholders**. This would also be contrary to Recital 6 of the DSM Directive, which reads: "*The exceptions and limitations provided for in this Directive seek to achieve a fair balance between the rights and interests of authors and other rightholders, on the one hand, and of users on the other. They can be applied only in certain special cases that do not conflict with the normal exploitation of the works or other subject matter and do not unreasonably prejudice the legitimate interests of the rightholders.*" and the Article 5, Section 5 of the Infosoc Directive.

We asked our members these questions:

a) How do you distinguish between commercial and research TDM?

It is impossible from a technical perspective to distinguish between commercial and research TDM which means publishers cannot identify the purpose. Publishers can make assumptions based on IP ranges and user agents, but it is impossible to be confident that a request is truly what it claims to be. In practice, it is almost impossible for the publisher to ensure that all possible IP ranges and user agents worldwide are known and then possible to make any limitations against them.

b) What happens if the one who uses your content for TDM it is not a recognised organisation?

For commercial organisations, permission needs to be obtained, and the organisation would need to contact the publisher. For academic research, especially if we are refereeing to an academic publisher, these institutions are known globally. Usually, publishers are part of the academic community and TDM rights are normally confirmed and an API set up. There are of course limitations in making opt-outs for entities and user agents that are unknown. As publishers are not in a position to constantly monitor, they are taking action when they notice that their content appears somewhere it should not.

c) Determining applicable law

It must be up to the third party to adhere to applicable law, and follow the provisions that are applicable to them based on whether they are doing research or non-research related TDM.

Often **what publishers try to do** in order to identify the purpose of TDM is to try – if possible – to identify the organisation carrying out the data mining and determine their activities and data transfer policies. However, in practice, distinguishing the two is increasingly hard, as datasets are transferred between organisations and sometimes used for multiple purposes. For example, CommonCrawl.org has for many years carried out TDM (CCbot) to collect vast amounts of data from the internet that "can be used by anyone". For a long time, the general impression has been that datasets collected by CommonCrawl were mainly used for reasonable research purposes. However, in recent years those same datasets have become a core resource for training f LLMs by OpenAI, Google and Anthropic.

Another example can come from National Libraries. Many of them for a long time have been doing TDM for research purposes with authority from the National Legal Deposit Acts. However, some actors intend to use these datasets to train foundation models that will be open sourced and can be used for commercial purposes. However, the legality of this use of data collected with authority from the National Legal Deposit Acts is questionable.

A general comment is that this distinction between research and commercial TDM is highly problematic. TDM might, as a starting point, be for research purposes, but if the result of the research is of commercial value, so it should be assumed that the result will be exploited commercially. This is definitely the case for e.g. LLM. This raises questions such as:

- Was the original TDM covered by the research related provisions, or should it have been assessed as commercial?
- Is the legal assessment that the transition from research to commercial de jure illegal unless it was assessed as non-research to start with?
- How this can be enforced in reality?

d) Do you apply any technological measures to help differentiate?

The technological measures that are available are based on specific user agents, and not purposes. Robots.txt works as a statement of intention ("code of conduct") for robots and crawlers, mostly applicable to SEO (search engine optimisation) so far. There is no technical solution whatsoever that technically limits TDM.

A general rule for all bots could be included in robots.txt, but this will have overly broad effects, also removing content from indexing. For news media, removal of content from being indexed is commercially not possible at all, as many publishers get traffic from web search services especially the dominant one. It would also have problematic societal effects, as editorial content would be less available for individuals.

A critical point: robots.txt files are (with the caveat above) not machine-readable for general purpose limitations/opt-outs, but for concrete known user agent prohibitions. This is definitely the case for AI TDM, for which **currently there is no standard for a general machine-readable, AI TDM purpose limited opt-out for unknown user agents/bots.** If you e.g. know that OpenAI's user agent is "X", you can disallow that user agent. But if a rogue actor starts using a different user agent, that user agent will not be disallowed for TDM / AI. Further, user agent is generally a voluntary item of information, varies extensively and is in practice very problematic for the purpose of identifying the recipient of an http request.

Several EPC members are examining whether tools such as botscorner (which maintain a mapping of crawlers) can be used to monitor TDM activity (see botscorner.fr).

e) How do you identify the bots of any AI company

It is still complicated especially as there is as yet no legal requirement for them to identify themselves, or their purpose. As mentioned above, there are some tools like botscorner and others that could help identify and block some bots. See above also comments about robots.txt.

The W3C is working on a TDM Rep standard. When (and if) finalised it will also tell bots whether they can access specific sites. Of course, this does not actually block them. If they wish to ignore the instructions, they are able to do so.

<u>For publishers it is critical that:</u>

- companies are required to be transparent about the bots they were using and for which purpose. Then it would be easier for publishers to accept or reject them.
- companies separated their bots depending upon what they were being used for, so if a publisher wishes a company to simply index content and do nothing else, they could indicate this.
- companies committed to abiding by the "instructions" within robots.txt and other instruments.

f) Opt-out is a measure which came into force with the DSM copyright directive. What happens about the content already used for training?

EPC members have reserved all rights to their content already before the DSM Directive was proposed let alone implemented. Therefore, the training of AI that took place before the implementation of the DSM directive could not be legal. DSM article 4 waives the exclusive right, and therefore the exclusive right must have been the clear starting point prior to the DSM. An important challenge to any legal action on this is of course the lack of transparency into which models actually have violated publishers IPR rights. In addition, in any case, the use of works for training AI involves copyright-relevant actions that are not covered by Article 4 DSM.

The opt-out solution is a poor legal solution for several reasons:

- Firstly, because there is currently no general protocol or technical solution for opting out towards all known and unknown (non-research) TDM crawlers. Current law requires opt-out, without this being possible in practice to any full extent.
- Robots.txt is used by some publishers, but robots.txt was not created to fulfill such a legal purpose. Other measures, such as meta tags in code, have the same weaknesses. From a technical perspective, all measures that can be made are anyway "soft" as there is no technical solution whatsoever that technically limits TDM.
- Publishers will look out for possibilities to bring a class action against this type of violations, but the lack of transparency is a deterring factor.

g) We understand that the EC believes that any content scraped from the open web which didn't include machine readable rights reservations after the Directive came into force was 'fair game', and in such cases the AI systems have been legally built. What is your view on their "lawful access" to build LLMs - i.e. if there are no machine-readable rights reservations then the exception kicks in?

Machine-readable means cannot be seen as an absolute requirement, taking into account that there are no technical means a publisher can take to ensure that our content is not used. This is the case today, and has been the case up until now as well.

The reproduction and analysis of copyright protected works for the creation of Foundation Models and Generative AI regardless of a machine readable or other reservation constitutes an obvious copyright infringement. The opposite would disrupt the normal exploitation of the works and unreasonably impair the legitimate interest of the rightholders. This would also be contrary to Recital 6 of the DSM Directive: "*The exceptions and limitations provided for in this Directive seek to achieve a fair balance between the rights and interests of authors and other rightholders, on the one hand, and of users on the other. They can be applied only in certain special cases that do not conflict with the normal exploitation of the works or other subject matter and do not unreasonably prejudice the legitimate interests of the rightholders.*" and the Article 5, Section 5 of the Infosoc Directive.

As stated in the introduction above, lawful access means permission is required from the rightsholder e.g. by means of a subscription or a data extraction agreement. The fact that copyrighted content is on the open web, supported by advertising to fund the investment in the production of that content, does not mean that it is available to scrape and use. There is also the distinction of lawful access for private use and for commercial use. Content which is available on the

web for individual users, does not imply that anyone can take it and use as they please. Such limitations exist in terms of use, and the lack of adhering to terms of use must result in lack of lawful access.

In the transposition of the DSM Directive into national law, in many cases it was emphasised in the legal comments that the rules on Copyright also apply to TDM. Therefore, it should not be sufficient for the AI developers to only rely on the TDM opt-out. Furthermore, they must also ensure that the content is legally available on the internet. The Books3 case illustrates this very well, as the approx. 200,000 e-books were freely available on the internet, but they were made illegally available, which is why the AI developers should not be able to absolve themselves of responsibility for their choice to train AI on the illegal copies by referring to the TDM exception.

Furthermore, as mentioned in previous answers, there are no technical machine-readable ways of opting out to unknown entities and user agents. The solutions that are used in practice today were not designed with this purpose in mind (e.g. robots.txt). Therefore, any other provided mechanism and wording on refusing reuse of publisher content is as relevant as limitations. Any other view is in practice highly problematic, as the EU law would in essence have expropriated all publisher content.

h) When reserving all rights in your terms and conditions is this done in machine readable form? If not, why?

The question is raised on incorrect assumptions. Terms and conditions are readable for humans, and even though digitally published terms and conditions can be digitally processed, TDM crawlers cannot read terms and conditions and assess whether they involve an opt-out. And if such technology can be applied in the context of TDM crawling, it is certainly not applied.

Other technical means must therefore be applied, and, as stated several times, there are no sufficient technical means to ensure a full machine-readable opt-out. Even if you involve free text in files such as robots.txt or meta tags that are meant to be machine-readable, the free text as such is as a starting point not machine-readable. Nevertheless, all publishers are using robot.txt, and have included the right reservations in there even though the standard does not support it.

i) Do you have any evidence that content behind paywalls has been scraped by (AI) crawlers?

This would depend on the type of paywall. Currently there are “hard paywalls”, which means that those sections of content that are behind the paywall, are behind paywall no matter what client you use. There are other types of paywalls that allow for the scraping of articles, even if the user runs into a paywall in their web browser. LLM companies often claim that they remove any paywalled content from scraped datasets before doing actual training. However, it is known that the initial (big) work by OpenAI did not respect paywalls[1].

[1] https://www.windowscentral.com/software-apps/chatgpt-pauses-bing-integration-to-stop-people-from-bypassing-paywalls

2. Territoriality

a) Publishers can opt out from their content being used for commercial TDM in the EU based on Art. 4 of the DSM, but what happens in other parts of the world with different systems, like the US? What is the territorial reach of the opt-out? Is your rights reservations respected there as well?

Global copyright protection has always been a complicated issue. Some publishers have a global approach, without distinctions per territory, while other consider that they have IPR protection according to the jurisdictions in which they operate.

Third parties need to respect these rights, and adhere to the laws in the publisher respective jurisdictions. Publishers already know that this is not happening in reality. Any reservation by a rightsholder to the use of works and other subject matter, be it in machine-readable format or in contract, is applicable regardless of the establishment of the entity conducting TDM. Thus, the use of the opt-out by any rightsholder established in the EU to the use of works etc. made available from/on websites in the EU must be respected regardless of the establishment of the entity conducting TDM.

For many publishers it is unclear where they can establish legal proceedings against IPR violations performed by entities performing TDM outside of their jurisdiction. Our view is that publishers must be able to do this within the courts of their jurisdiction, as long as their services are produced in their jurisdictions and aimed at an audience within their jurisdiction.

b) Is the US publishers position similar to that of the European opt-out?

Please refer to the US NMA White Paper and full submission to their Copyright Office.

Also see the Global Principles for AI which are aimed at ensuring publishers' continued ability to create and disseminate quality content, while facilitating innovation and the responsible development of trustworthy AI systems, in full respect of copyright.