NEWS MEDIA
ALLIANCE

January 10, 2020

Andrei Iancu
Under Secretary of Commerce for Intellectual Property
and Director of U.S. Patent and Trademark Office
P.O. Box 1450
Alexandria, Virginia 22313-1450

Re: Request for Comments on Intellectual Property Protection for Artificial Intelligence Innovation

Dear Mr. Iancu:

The News Media Alliance (the "Alliance") respectfully submits this comment to Question No. 3 in the Request for Comments published by the Patent & Trademark Office at 84 Fed. Reg. 58141 (Oct. 30, 2019):

> To the extent an [Artificial Intelligence] algorithm or process learns its function(s) by ingesting large volumes of copyrighted material, does the existing statutory language (e.g., the fair use doctrine) and related case law adequately address the legality of making such use? Should authors be recognized for this type of use of their works? If so, how?

The members of the Alliance are deeply concerned about the unlicensed use of their news reporting for machine learning purposes by technology companies that do not share the cost of reporting the news but commercially benefit from the work product of the news media by using the news in a manner that does not qualify as fair use. While current copyright doctrine should compel the conclusion that this constitutes infringement of copyright, a variety of obstacles to enforcement of the media's IP rights has diminished the value of those rights. The modes of distribution and consumption of news content are rapidly changing in the digital age, and the failure to properly compensate the media for the use of their content has already become an existential threat to the business of journalism. Moreover, the continued unlicensed use of reporting by technology companies portends injury, not just to the news industry, but to the public interest that it serves: in a world in which everyone is a republisher, there will be no original reporting to republish. Accordingly, the Alliance believes that stronger enforcement of existing laws is needed to reset the balance between the originators of news and those who consume it for their own commercial advantage.

The News Media Alliance

The Alliance is a nonprofit organization that represents the interests of more than 2,000 news media organizations in the United States and internationally. The Alliance diligently advocates for newspapers before the federal government on issues that affect today's media organizations, including protecting newspapers' intellectual property.

News organizations play an important role in the U.S. economy and democracy. Every day, news publishers invest in high-quality journalism that keeps our communities informed, holds those in power accountable, and supports the free flow of information and ideas in society. Without free and flourishing news media, our society would be less well-off and less informed.

The newspaper industry generates over $25 billion in total revenue and employs a total of approximately 152,000 people in the United States.[1] These journalists and others who rely on newspapers for their living create content that reaches 136 million adults in the United States each week, representing 54 percent of the country's adult population.[2] Online, news organizations receive over 200 million unique visits and 6.7 billion page views per month, while 44 percent of the news media audience relies exclusively on print publications.[3] News publishers also ensure the health of our local communities, with most local news media companies reaching more adults in their local markets than any other local media.[4]

Notwithstanding their vital societal role and the public's reliance on accurate and current information, news publishers are struggling to sustain investments in high-quality journalism. Despite an increase in digital audience and subscriptions, both overall and print circulation dropped by approximately 8 and 12 percent between 2017 and 2018, respectively.[5] In total, news publisher revenues have decreased by 58 percent since 2005, and newsroom employment has dropped from over 72,000 to an estimated 37,900 in the same period.[6] While the share of digital advertising revenues has grown in recent years,[7] such revenues are often not enough to offset the reduced print advertising and subscription revenues. News publishers struggle in large

[1] Pew Research Center, Newspaper Fact Sheet (July 9, 2019), http://www.journalism.org/fact-sheet/newspapers/; United States Department of Labor, Bureau of Labor Statistics, Occupational Employment Statistics (May 2018), https://www.bls.gov/oes/current/naics5_511110.htm#00-0000.

[2] News Media Alliance, News Advertising Panorama: A wide-ranging look at the value of the news audience at 64 (2018), https://www.newsmediaalliance.org/wp-content/uploads/2018/10/FINAL_NMA_PANORAMAbook_WEB_10-19-18.pdf.

[3] *Id*.

[4] *Id*.

[5] Pew Research Center, Newspaper Fact Sheet (July 9, 2019), http://www.journalism.org/fact-sheet/newspapers/.

[6] *Id*.

[7] *Id*.

part because the online marketplace is dominated by a few online platforms, referred to in this comment as "tech platforms," that control the digital advertising ecosystem and determine the reach and audience for news content online, thereby reducing the ability of news publishers to benefit from digital advertising and to develop their relationships with their readers.

The Problem

The news media rely on robust legal protection of their intellectual property, typically in the form of copyrights, for their very existence. The remuneration made possible in the form of subscriptions and various licensing fees for distribution of their content generates the revenue necessary to finance the cost of reporting the news, such as the global network of news bureaus and journalists that major media enterprises must maintain to carry on their work. The ability of those who do not foot the bill to free ride on those efforts would quickly extinguish the practice of journalism as we know it. And the stakes are not merely commercial. The words of district judge Denise Cote in *Associated Press v. Meltwater U.S. Holdings, Inc.*, 931 F. Supp. 2d 537, 553 (S.D.N.Y. 2013), are apt:

> [T]he world is indebted to the press for triumphs which have been gained by reason and humanity over error and oppression . . . . Permitting [Meltwater] to take the fruit of [AP's] labor for its own profit, without compensating [AP], injures [AP's] ability to perform [its] essential function of democracy.

Artificial Intelligence ("AI") is increasingly involved in various ways in the practice of journalism. As relevant here, tech platforms scrape news websites and ingest copyright-protected news content. The content appropriated by these technologies, typically in massive quantities, may be used to train the AI to perform a variety of functions, which increasingly includes learning from news media reports how to write a story, and drawing on content from multiple sources to create a rendition of the news that is not identical to that of any one contributing source while being completely dependent on all of those sources in combination. For simplicity, we will refer to this broad type of use as "training AI." It is a form of "machine learning": "Instead of requiring people to manually encode hundreds of thousands of rules, this approach programs machines to extract those rules automatically from a pile of data."[8] While the deployment of this process by tech giants such as Google and Amazon may be the most remarked upon, they are not alone in engaging in the sort of machine learning of greatest concern to the Alliance and its members.[9]

[8] Karen Hao, We Analyzed 16,625 Papers to Figure Out Where AI is Headed Next, *MIT Technology Review* (Jan. 25, 2019), https://www.technologyreview.com/s/612768/we-analyzed-16625-papers-to-figure-out-where-ai-is-headed-next/.

[9] *See, e.g.*, Knowhere Launches with $1.8M in Funding to Deliver Unbiased News Coverage with Machine Learning (Apr. 4, 2018), https://s3.us-west-2.amazonaws.com/cruncher-images/static/press-release/knowhere-launch-press-release.pdf ("Knowhere's technology scours the internet, evaluating narratives, factual claims and bias in reporting, by outlets as varied as the New York Times and Breitbart, to inform three 'spins' of every controversial story: left, impartial, and right, or positive, impartial and negative. The technology can write stories in

The use of copyright-protected news content by tech platforms to train their AI represents an increasing threat to the practice of journalism. The challenges are multiple. First, current technology makes the replication and manipulation of vast amounts of content inexpensive and easy. Just as the printing press and later the photocopier exponentially increased the capacity of humans to reproduce content that initially had to be manually copied, another giant leap has occurred in the digital age and text can be replicated, processed and disseminated with a few keystrokes. Second, this use is not readily detected. Whereas copying and re-dissemination of all or a substantial portion of intact text to the public is generally detectible and provable as an infringement, programs running in the background at unlicensed tech platforms that make use of ingested news content as the raw material input for machine learning cannot so easily be policed by the content proprietor. Third, the dominance of the principal tech platforms renders enforcement of IP rights extremely difficult when the publishers are dependent on the same tech platforms to mediate between the publishers and their audience by locating and linking to reports of interest to users. Fourth, the training of AI is increasingly being used to support news products that cause the audience to remain inside the tech platform's ecosystem, rather than simply as a search tool that links users to the original information provider. This evolution can be seen in contrasting public statements by Google. In 1998, in a publication entitled "Ten Things We Know to be True," Google maintained, "We may be the only people in the world who can say our goal is to have people leave our website as quickly as possible."[10] By 2011, however, the chief executive of Google would testify to the Senate Judiciary Committee, "if we know the answer … it is better for the consumer for us to answer that question so that they don't have to click anywhere."[11] The phenomenon of tech platforms "answering the question" can be seen in such relatively recent developments as "featured snippets" in Google search responses, which obviate the need for users to click through to a source for the requested information, and voice assistant products such as Amazon's Alexa and Google Assistant, which will directly answer user queries without providing a ready path for users to consult the source of the information. These devices are powered by AI systems that have been trained on the information ingested from the websites of content originators, many of which are traditional news sources that have expended time and labor to collect and report it.

The recent implementation by Google of its machine learning tool nicknamed "BERT" shows how tech platforms can make use of expressive textual content to train their search engines. A recent article by a Google officer explains:

anywhere from 60 seconds to 15 minutes, depending on the amount of controversy among the sources. Once article drafts are complete, human journalists review the piece, which in turn trains the machine learning algorithm.").

[10] Ten things we know to be true, Google, https://www.google.com/about/philosophy.html (last visited Jan. 7, 2020).

[11] The Power of Google: Serving Consumers or Threatening Competition?, Hearing before the Subcommittee on Antitrust, Competition Policy and Consumer Rights of the Committee on the Judiciary, 112th Cong., Sept. 21, 2011, *available at* https://www.govinfo.gov/content/pkg/CHRG-112shrg71471/html/CHRG-112shrg71471.htm.

> With the latest advancements from our research team in the science of language understanding--made possible by machine learning--we're making a significant improvement to how we understand queries, representing the biggest leap forward in the past five years, and one of the biggest leaps forward in the history of Search.
>
> **Applying BERT models to Search**
>
> Last year, we introduced and open-sourced a neural network-based technique for natural language processing (NLP) pre-training called Bidirectional Encoder Representations from Transformers, or as we call it--BERT, for short. This technology enables anyone to train their own state-of-the-art question answering system.
>
> This breakthrough was the result of Google research on transformers: *models that process words in relation to all the other words in a sentence, rather than one-by-one in order. BERT models can therefore consider the full context of a word by looking at the words that come before and after it*—particularly useful for understanding the intent behind search queries.[12]

Even before the advent of sophisticated AI tools to enhance the efficiency of search engines, enterprises such as Google scraped the full text of news reports from media websites and ingested that material into the Google search database. The description of the BERT program as used by Google makes it all the more clear that this process requires copying and analyzing the full text of third party content, because the AI is exquisitely reliant on the precise grammar and word selection of the text to teach itself how to interpret queries, carry out searches and deliver responsive content. The tech platforms' AI can also learn how to write news articles by analyzing the text of news reports provided by human sources. For all these reasons, the tech platforms make use of the precise expression in the news articles they ingest and do more than just extract facts from those reports.

The Alliance believes that, in the words of Question 3, "ingesting large volumes of copyrighted material" for this purpose constitutes copyright infringement if undertaken without a license from the proprietor of the material. This is true whether or not the platform goes on to disseminate any of that material in a form substantially similar to the ingested original. Tech platforms that appropriate vast quantities of news content for this purpose should pay for the privilege of doing so, no less than they should pay for the electricity that powers their computers or motorists for the fuel that powers their cars.

[12] Pandu Nayak (Google Fellow and Vice President, Search), Understanding Searches Better than Ever Before, (Oct. 25, 2019), https://www.blog.google/products/search/search-language-understanding-bert/ (emphasis supplied)

Analysis

1. Textual and visual news content is fully protected by copyright. While facts by themselves cannot be protected by copyright, the expression of facts is so protected. *Feist Publ'ns, Inc. v. Rural Tel. Serv. Co.*, 499 U.S. 340, 348 (1991). The case law is replete with examples of infringers of news reporting in diverse media being held liable for infringement. *See, e.g.*, *Fox News Network, LLC v. TVEyes, Inc.*, 883 F.3d 169, 182 (2d Cir. 2018) ("*TV Eyes*") (video news clips); *Nihon Keizai Shimbun, Inc. v. Comline Business Data, Inc.*, 166 F.3d 65 (2d Cir. 1999) (newswire text articles); *H.C. Wainwright & Co. v. Wall St. Transcript Corp.*, 418 F. Supp. 620 (S.D.N.Y. 1976), *aff'd*, 558 F.2d 91 (2d Cir. 1977), *cert. denied*, 434 U.S. 1014 (1978) (reports concerning stocks and bonds); *Agence France Presse v. Morel*, 934 F. Supp. 2d 584, 592 (S.D.N.Y. 2013) (news photographs).

2. The ingestion of substantial volumes of news content is, at minimum, a prima facie infringement of the reproduction right. The Copyright Act enumerates the exclusive rights of copyright in section 106. First among those rights is the exclusive right "to reproduce the copyrighted work in copies . . . ." 17 U.S.C. § 106(1). Because the reproduction must be verbatim to lend itself to training the AI, the question whether the copy is sufficiently similar to the original does not arise. When a tech platform ingests a volume of news reporting and fixes it in a database in the memory of its computer system, it has made an infringing copy. *E.g.*, *Stern Elecs., Inc. v. Kaufman*, 669 F.2d 852, 855 (2d Cir. 1982) ("[T]he memory devices of the game satisfy the statutory requirement of a 'copy' in which the work is 'fixed.'").[13] If, as is common practice, the dataset is then manipulated in the course of carrying out machine learning, infringement of the exclusive right to create derivative works likely also occurs. 17 U.S.C. § 106(2).

Those who would immunize these activities often describe the ingestion of content by tech platforms as "non-expressive" use, because the AI learning assertedly depends on use of the content as "data" rather than as a communicative work. The Copyright Act, however, does not make this distinction. A reproduction of a work in copies—which are "material objects . . . in which a work is fixed by any method now known or later developed, and from which the work can be perceived, reproduced, or otherwise communicated, either directly or with the aid of a machine or device[,]" 17 U.S.C. § 101—without more, is a prima facie exercise of the section 106(1) right irrespective of the intended use. Similarly, the fact that the ingestion of a work for machine learning can be described as a "non-display use" because it does not involve further dissemination of the work is itself not a defense. The reproduction right exists separately from the further rights to "distribute copies . . . of the copyrighted work to the public . . . [,]" *id.* §106(3), and "to display the copyrighted work publicly[,]" *id.* § 106(5). The making of the copy is prima facie infringement *whether or not* it is then distributed or displayed, for example, in

[13] We are not here concerned with "intermediate" copying of a computer program, *see* 17 U.S.C. § 117(a)(1), or with copies whose existence is so brief as to not constitute a "fixation." *See Cartoon Network LP v. CSC Holdings, Inc.*, 536 F.3d 121, 129-30 (2d Cir. 2008).

response to a search query.[14]

3. It would be a mistake to assume that the ingestion of substantial volumes of news content for machine learning is fair use. While understood to be a part of copyright law for centuries, the fair use doctrine was not codified until the 1976 Act in section 107. The preamble of that section mentions "news reporting" as an example of the type of use that could attract a fair use defense, but the Supreme Court has made clear that these examples are "illustrative and not limitative" and "provide only general guidance about the sorts of copying that courts and Congress most commonly ha[ve] found to be fair uses." *Campbell v. Acuff-Rose Music, Inc.*, 510 U.S. 569, 577-78 (1994). Thus, as the Second Circuit observed in the *Google Books* decision, "[t]hose who report the news undoubtedly create factual works. It cannot seriously be argued that, for that reason, others may freely copy and re-disseminate news reports." *Authors Guild v. Google, Inc.*, 804 F.3d 202, 220 (2d Cir. 2015) ("*Google Books*"). Of course, inquiries into fair use are necessarily fact-specific and do not lend themselves to bright-light generalizations. Nevertheless, any consideration of whether ingestion of news content for machine learning purposes is fair use would proceed to consideration of the four factors set out in section 107, which themselves are not exhaustive, but merely indicative of whether a use should be deemed fair.

(1) **The purpose and character of the use**. As *Campbell* and numerous other authorities agree, the "transformative" use of a copyrighted work is a distinctive feature of a fair use. Secondary works that are productive in that they put the original to a new use, such as quotation for the purpose of parody or criticism, in and of themselves provide something new and different to the public and are less likely to merely supersede the purposes of the original. They therefore have a greater call on protection from infringement. As Judge Leval explained in *Google Books*, "transformative uses tend to favor a fair use finding because a transformative use is one that communicates something new and different from the original or expands its utility, thus serving copyright's overall objective of contributing to public knowledge." 804 F.3d at 214. Yet, as Judge Leval—the originator of the "transformative use" concept[15]—also cautioned in the same decision, the term "transformative" must be applied with discretion. It would be overly simplistic to suggest that any and all repurposing or recasting of a copyrighted work is transformative in a sense meaningful to fair use analysis. Among other things, "transformation" of the work into a derivative work is a right expressly reserved to the copyright proprietor, 17 U.S.C. § 106(2).

The ingestion of volumes of news content to obtain material for machine learning is a pure act of consumption, not of transformation. Whereas *Google Books* and like decisions turn on the value of "communicating" to the audience something new and socially desirable, in the case of machine learning the relevant value of the news content is extracted within the computer system of the tech platform before any new work is created and is not defensible unless, arguably, all possible resulting uses by the ingesting party are fair use. For example, the linchpin of the fair use determination in *Google Books* was the provision of information "about" the

[14] Whether or not such a use qualifies as fair use is a separate issue, discussed below.

[15] Pierre Leval, Toward a Fair use Standard, 103 Harv. L. Rev. 1105 (1990).

original books to users interested in knowing, for example, whether particular words were used in the original in order to help users find what they were looking for in a book. No such socially redeeming value to the public can be discerned in the ingestion of content for broad applications of machine learning. Here, the originals are used as the fuel to run the tech platform's engine.

The possibility that machine learning may in turn be applied to *some* purpose that would not infringe copyright does not excuse all antecedent ingestion. As commentator Benjamin Sobel has cogently argued:

> Making gigabytes upon gigabytes of copies of copyrighted art, in order to teach a machine to mimic that art, is indeed a remarkable technological achievement. An artificially intelligent painter or writer may yield social benefits and enrich the lives of many beholders and users. However, this view of productivity is overbroad. No human can rebut an infringement claim merely by showing that he has learned by consuming the works he copied, even if he puts this new knowledge to productive use later on.[16]

That observation is surely correct. When tech platforms ingest published news content and set their AI programs upon that text in order to, in the words of Google, "process words in relation to all the other words in a sentence," (*supra*, note 12), they are appropriating the expressive content of the original work and do not enjoy blanket immunity merely because some downstream activities facilitated by that appropriation may be deemed productive or socially desirable. For example, the fact that some recipients of unlicensed copies of broadcast news clips may have wished to use them for research or other salutary purposes does not render the pervasive copying and distribution of such clips a fair use. *See TVEyes*, 883 F.3d at 178 n.4 ("That a secondary use can facilitate research does not itself support a finding that the secondary use is transformative.") (citing *American Geophysical Union v. Te*xaco, 60 F.3d 913 (2d Cir. 1994)).

(2) **The nature of the copyrighted work**. This factor is generally recognized to be the least significant in the fair use calculus. *E.g*., *Google Books*, 804 F.3d at 220 (citing William F. Patry, Patry On Fair Use § 4.1 (2015)). To be sure, there is dictum in *Harper & Row* that "[t]he law generally recognizes a greater need to disseminate factual works than works of fiction or fantasy." *Harper & Row, Publrs., Inc. v. Nation Enters*., 471 U.S. 539, 563 (1985). *Google Books* questions how far that dictum should impact fair use analysis. 804 F.3d at 220. But to the extent it suggests a possibly greater scope of fair use in connection with factual reports, it would only be because of the public interest in "dissemination" in order to afford public access to the facts reported—a need not satisfied when tech platforms ingest the news for their own commercial purposes.

[16] B. L. W. Sobel, Artificial Intelligence's Fair Use Crisis, 41 Colum. J. L. & Arts 45, 73 (2017) (citing *Sony Corp. of Am. v. Universal City Studios, Inc*., 464 U.S. 417, 455 n.40 (1984), which suggests that a "constituent who copies a news program to help make a decision on how to vote" would not be protected by the fair use doctrine despite the salutary purpose).

(3) **The amount and substantiality of the portion used in relation to the copyrighted work as a whole**. This factor supports the view that ingestion of substantial, indeed vast, volumes of text without the permission of the copyright owner for the purpose of machine learning is not a fair use. Although a compelling fair use purpose on rare occasions can justify taking the entirety of a work when taking less will not suffice, *Swatch Grp. Mgmt. Servs. v. Bloomberg L.P.*, 756 F.3d 73, 90 (2d Cir. 2014), this is not such a case. *Swatch*, which involved an unlicensed dissemination of a corporate earnings call, turned on the public interest in having access to the contents of the call and the risk that paraphrasing or excerpting would not accurately render the nuance of what was discussed. Again, this provides no justification for a use that does not enhance public knowledge but represents pure consumption by the tech platform.[17]

(4) **The effect of the use upon the potential market for or value of the copyrighted work**. This is often stated to be the most important fair use factor. *E.g.*, *Harper & Row Publ. v. Nation Enters.*, 471 U.S. 539, 566 (1985). Members of the Alliance are particularly concerned about the predictable, and already occurring, commercial harm inflicted on them by the increasing use of their intellectual property without their permission to train the AI employed by tech platforms. While the diversion of audience that occurs when a person in horizontal competition with the content proprietor appropriates an original work and markets it in competition with the originator is an obvious example of Factor 4 harm, that is not the only cognizable type of harm. Factor 4 has a vertical aspect as well: depriving content creators of natural markets wherein they can sell or license their works for such consumption is also—literally as well as a matter of commercial reality—a pernicious "effect on the potential market for or value of" the copyrighted material. As the Second Circuit noted in *Castle Rock Entm't, Inc. v. Carol Pub. Group, Inc.*, 150 F.3d 132, 145 (2d Cir. 1998), "[t]he fourth factor must also 'take account ... of harm to the market for derivative works,' defined as those markets 'that creators of original works would in general develop or license others to develop[.]'" (citation omitted). "It is indisputable that, as a general matter, a copyright holder is entitled to demand a royalty for licensing others to use its copyrighted work, and that the impact on potential licensing revenues is a proper subject for consideration in assessing the fourth factor[.]" *Texaco*, 60 F.3d at 929 (citations omitted). In doing so, the courts look to the use's impact on "traditional, reasonable, or likely to be developed markets." *Id.* at 930. Thus, in *Texaco*, the bulk photocopying by a commercial enterprise's research arm of scientific articles published by plaintiff was deemed not a fair use where the licensing of such articles was a natural, and to some extent already exploited, market for such scientific articles through the development of clearinghouses established to license such photocopying.

When a consumer of copyrighted material exploits that material without permission,

[17] We do not view the advent of "federated learning", *see generally* B. McMahan & D. Ramage, Federated Learning: Collaborative Machine Learning without Centralized Training Data, Google Research Blog (Apr. 6, 2017), https://perma.cc/XVA2-J96J, to change the analysis or result. When a central authority such as Google delegates portions of a large database to individual users to analyze and return results in order to replicate the more conventional model of ingesting the entire database at one site, the delegator should incur liability for inducing infringement, and the individual user delegees for direct infringement of the portions copied to their devices.

Factor 4 is triggered even where the use is for a purpose collateral to the main or original purpose of creating the material. That is one of the important lessons of *TVEyes*, where the defendant ingested vast amounts of broadcast television news programming and enabled its subscribers to watch, download and save actual news clips of up to ten minutes duration without license from the source broadcasters. *Id.* at 175. The court found that "Fox itself might wish to exploit the market for such a service . . . . [and that] TVEyes deprives Fox of revenues to which Fox is entitled as the copyright holder." *Id.* at 180.[18]

Here, as in *Texaco* and *TVEyes*, the licensing of the copyrighted content for the use made of it by the tech platforms is, if not "traditional," certainly a "reasonable, or likely to be developed market[]." *Texaco*, 60 F.3d at 930. Media entities for some time have identified this market as one in which their proprietary content has particular value and have curated and made available annotated *corpora* of their published news reporting for the specific purpose of training AI. This has been done by, for example, the copyright holders of *The Wall Street Journal*,[19] *The New York Times*,[20] and the Reuters News Service[21]. The Linguistic Data Consortium catalogue

[18] Another lesson of *TVEyes* is that limiting the unlicensed use to "internal purposes only" confers no talismanic immunity from infringement. The defendant TVEyes purported to contractually require its subscribers to make only such "internal" use of downloaded clips, *id.* at 175, but that did not privilege its unlicensed distribution of copyrighted video clips. So too here, when tech platforms ingest news content to train their AI, they are making an "internal" use of that content, but that should not protect them from infringement.

[19] Linguistic Data Consortium – BLLIP 1987-89 WSJ Corpus Release 1, https://catalog.ldc.upenn.edu/LDC2000T43 (last visited Jan. 6, 2020).

[20] Linguistic Data Consortium – The New York Times Annotated Corpus, https://catalog.ldc.upenn.edu/LDC2008T19 (last visited Jan. 6, 2020). For a description of the corpus, see this blog post announcing and explaining the corpus: Jacob Harris, Fatten Up Your Corpus, NYT Open (Jan. 12, 2009), https://open.blogs.nytimes.com/2009/01/12/fatten-up-your-corpus/ ("Available for noncommercial research license from The Linguistic Data Consortium (LDC), the corpus spans 20 years of newspapers between 1987 and 2007 (that's 7,475 issues, to be exact). This collection includes the text of 1.8 million articles written at The Times (for wire service articles, you'll have to look elsewhere). Of these, more than 1.5 million have been manually annotated by The New York Times Index with distinct tags for people, places, topics and organizations drawn from a controlled vocabulary. A further 650,000 articles also include summaries written by indexers from the New York Times Index. The corpus is provided as a collection of XML documents in the News Industry Text Format and includes open source Java tools for parsing documents into memory resident objects.").

Google has acknowledged accessing this corpus for machine learning, specifically, developing "entity salience." *See* Dan Gillick, *et al.*, Teaching machines to read between the lines (and a new corpus with entity salience annotations), Google AI Blog (Aug. 25, 2014), https://ai.googleblog.com/2014/08/teaching-machines-to-read-between-lines.html.

[21] *See, e.g.*, David D. Lewis, Reuters-21578 Text Categorization Test Collection Distribution 1.0 README file (Sep. 26, 1997), https://perma.cc/V7JJ-CNVW. This corpus consists of the contents of the Reuters newswire for 1987.

lists hundreds of such *corpora* available for license.[22] Today, major news media organizations continue to commercially license these rights and have formulated and offer licensable data products for this purpose. The use in question thus easily satisfies the requirement of *Texaco* that it be in a market in which users should reasonably expect to seek permission of the copyright proprietors and to compensate them for the use of their works—even though some major tech platforms do not do so.

This use is also dissimilar from those held to be fair in such decisions as *Google Books*, 804 F.3d 202; *Authors Guild, Inc. v. HathiTrust*, 755 F.3d 87, 95 (2d Cir. 2014); *Kelly v. Arriba Soft Corp*., 336 F.3d 811 (9th Cir. 2003); *Perfect 10, Inc. v. Amazon.com, Inc*., 508 F.3d 1146 (9th Cir. 2007). In those decisions, the defendant's copying of copyrighted material was in the service of creating a searchable index that enabled users to locate desired content and link to it. In theory at least, this was beneficial to content originators and drove traffic to their websites by displaying a snippet that by itself did not substitute for the original full-text work. *See, e.g.*, *Authors Guild, Inc. v. Google, Inc.*, 954 F. Supp. 2d 282, 291 (Google Books "uses snippets of text to act as pointers directing users to a broad selection of books"), *aff'd*, 804 F.3d 202; *Perfect 10, Inc.*, 508 F.3d at 1165 ("a search engine transforms the image into a pointer directing a user to a source of information"). This feature is absent from the case of tech platforms consuming news content to train their AI for broader purposes. The tech platform *may* use the AI to make more efficient the searches that *may* send inquirers to the original website, but they may also be used—and increasingly are used—to create alternative renditions of news and information that disintermediate between the original source and its audience. This results in market harm that far outweighs any consumer good generated by this process. Thus, the recognition in current case law that internal reproduction and indexing of content for some machine-driven purposes may be fair use should not be extended to a blanket immunity for all ingestion of copyright-protected content by tech platforms for any and all commercial purposes.

* * * * *

Today, having access to trusted content of consistently high quality is integral to power machine learning. By compensating news media organizations for their intellectual labor in generating that content, an appropriate balance can be struck between advancing AI-based technology, while preserving the media's "ability to perform [its] essential function of democracy." *Meltwater*, 931 F. Supp. 2d at 553. The Alliance believes that copyright law, properly understood and consistently enforced, should lead to a system where content originators are compensated for their work. Various business and licensing models, with proper legal support, may be employed to achieve this end. Pressing policy concerns—including sustainability of the news media industry and, by extension, benefit to the public that relies on it for accurate and current content—demand this outcome.

[22] Linguistic Data Consortium, LDC Catalog, https://catalog.ldc.upenn.edu/ (last visited Jan. 6, 2020).

To the extent the current legal framework cannot support such a regime, legislative solutions may prove useful or even necessary. But the issues addressed herein require more analysis, dialogue among all stakeholders, and careful attention to detail. The Alliance calls for and stands ready to contribute to continued study and deliberation on this important issue that will help move American copyright law fully into the 21st century.

Respectfully submitted,

David Chavern
President & Chief Executive Officer
NEWS MEDIA ALLIANCE
4401 North Fairfax Drive, Suite 300
Arlington, Virginia 22203

Danielle Coffey
Senior Vice President & General Counsel
NEWS MEDIA ALLIANCE
4401 North Fairfax Drive, Suite 300
Arlington, Virginia 22203

Robert P. LoBue
Julie Simeone
PATTERSON BELKNAP WEBB &
TYLER LLP
1133 Avenue of the Americas
New York, New York 10036
Counsel for the News Media Alliance