IMPLICACIONES LEGALES DEL *WEB SCRAPING* EN EL ENTRENAMIENTO DE MODELOS DE INTELIGENCIA ARTIFICIAL GENERATIVA

JUAN MANUEL PACHECO CHAPARRO*
LAURA BARRERO RAMÍREZ**

RESUMEN

El *web scraping* es una técnica que se usa para recopilar datos en Internet y almacenarlos en una base de datos. Ese proceso se usa, entre otras cosas, para entrenar modelos de inteligencia artificial generativa y ha generado controversia alrededor del mundo debido a sus riesgos legales. En este artículo se analizará la viabilidad legal del uso de técnicas de *web scraping* y se abordarán tensiones relacionadas con asuntos contractuales de los términos de servicio de las páginas web, los riesgos legales que se desprenden de estas técnicas y, en particular, del uso de obras protegidas en el entrenamiento de modelos de inteligencia artificial generativa, de protección de datos personales y de implicaciones penales; las licencias *open source*, *open access* y de *Creative Commons*, así como también los datos de dominio público y en cabeza del Estado colombiano. Este artículo pretende ser un marco teórico inicial para la discusión del *web scraping* en modelos de inteligencia artificial generativa, dado que, a la fecha de elaboración de este artículo, el desarrollo normativo y jurisprudencial sobre este tema es aún incipiente.

Palabras clave: datos, *web scraping*, inteligencia artificial, términos de servicio (ToS), privacidad, propiedad intelectual, dominio público.

* Abogado de la Pontificia Universidad Javeriana, cursando su especialización en Derecho Informático y Nuevas Tecnologías en la Universidad Externado de Colombia. Afiliación institucional: asociado en el equipo de Telecomunicaciones, Medios y Tecnología de Brigard Urrutia. Contacto: bpacheco@bu.com.co. Fecha de recepción: 1 de abril de 2024. Fecha de aceptación: 2 de mayo de 2024.

** Estudiante de Derecho con énfasis en Derecho Comercial y Derecho de Daños de la Pontificia Universidad Javeriana. Afiliación institucional: auxiliar asociada en el grupo de práctica de Propiedad Intelectual de Baker McKenzie. Contacto: laura.barrero@bakermckenzie.com. Fecha de recepción: 1 de abril de 2024. Fecha de aceptación: 2 de mayo de 2024. Para citar el artículo: Pacheco Chaparro, Juan Manuel y Barrero Ramírez, Laura, "Implicaciones legales del web scraping en el entrenamiento de modelos de inteligencia artificial generativa", en *Revista La Propiedad Inmaterial* n.º 38, Universidad Externado de Colombia, julio-diciembre 2024, pp. 167-189. DOI: https://doi.org/10.18601/16571959.n38.07.

LEGAL IMPLICATIONS OF WEB SCRAPING IN THE TRAINING OF GENERATIVE ARTIFICIAL INTELLIGENCE MODELS

ABSTRACT

Web scraping is a technique used to collect data on the Internet and record it in a database. This process is used, among other things, to train generative artificial intelligence models, and has generated worldwide controversy due to its legal risks. This article will analyze the legal viability of the use of web scraping techniques and will address tensions related to contractual issues in the terms of service of web pages; legal risks arising from web scraping, and in particular the use of protected works in the training of generative artificial intelligence models, personal data protection, and criminal implications; open source, open access, and Creative Commons licenses, as well as public domain data and data held by the Colombian State. This article aims to be an initial theoretical framework for the discussion of web scraping in generative artificial intelligence models, given that at the time of writing this article, the regulatory and jurisprudential development on this topic is still incipient.

Key words: Data, Web Scraping, Artificial Intelligence, Terms of Service (ToS), Privacy, Intellectual Property, Public Domain.

INTRODUCCIÓN

Los modelos de inteligencia artificial generativa obtienen sus datos a través de datos públicamente accesibles[1]. Para obtener esos datos se pueden usar distintas técnicas, y una de las más populares es el *web scraping*.

El *web scraping* permite extraer datos en Internet y almacenarlos en una base de datos[2]. El método de extracción puede ser manual (el tradicional *copy-paste*) o automatizado (a través de tecnologías como *crawlers* y *parsers*)[3]. Debido a la cantidad de datos que se requieren para entrenar un modelo de inteligencia artificial generativa, el uso de estas técnicas se ha generalizado en Internet. OpenAI, por ejemplo, utiliza el GPTBot, un *crawler* que se usa para extraer datos y procesarlos con el objetivo de que el modelo sea más preciso y aumente sus funcionalidades[4].

1 OpenAI. "GPT-4 Technical Report". *Computation and Language (cs.CL)* (Nueva York: Cornell University. arXiv:2303.08774 [cs.CL], 2023).

2 Bo Zhao. "Web Scraping". *Encyclopedia of Big Data*, pp. 1 - 3. DOI: 10.1007/978-3-319-32001-4_483-1

3 Chaimaa Lofti, Swetha Srinivasan, Myriam Ertz y Imen Latrous. Web scraping techniques and applications: A literature review. *SCRS Conference Proceedings on Intelligent Systems*, pp. 381 - 394.

4 OpenAI. "GPTBot". [En línea]. 2023. [Consulta: 10 de marzo de 2024].

Sin embargo, se ha generado una discusión sobre la viabilidad legal y técnica de extraer datos a través de *web scraping*. La técnica puede tener un impacto en materia de privacidad, ciberseguridad, propiedad intelectual e incluso en materia contractual. Al usar datos dispuestos en la red sin licencia, es previsible el reclamo de personas que aleguen un mejor derecho.

Para hacer frente a esa discusión, se han propuesto estándares y protocolos voluntarios para restringir el acceso a sitios web de clientes automatizados como los *crawlers*[5]. Según datos de Originality AI, para marzo del 2024, el 32,7 % de 1000 de los sitios web más buscados en Internet están bloqueando el *crawler* de GPTBot[6].

Este artículo pretende abordar la legalidad del *web scraping* para el entrenamiento de modelos de inteligencia artificial generativa mediante un estudio de las discusiones más relevantes a ese respecto. Para esos efectos, el artículo estudiará el contexto internacional del *web scraping* y lo analizará desde la perspectiva legal colombiana. El artículo se divide de la siguiente forma: I. Contexto legal del *web scraping* y los modelos de inteligencia artificial generativa. II. Los términos de servicio y el *web scraping*. III. Riesgos legales del *web scraping*. IV. Licencias en Internet, dominio público y *web scraping*. V. Conclusión.

I. CONTEXTO LEGAL DEL *WEB SCRAPING* Y LOS MODELOS DE INTELIGENCIA ARTIFICIAL GENERATIVA

Desde el lanzamiento de ChatGPT en el 2022, se han visto múltiples demandas y procedimientos legales por motivos relacionados con los datos que se usan para el entrenamiento de modelos de inteligencia artificial generativa. A continuación, se resumen algunos de los casos más representativos:

– *Noviembre de 2022*: Microsoft, GitHub y OpenAI fueron demandados por presuntamente violar derechos de autor al reproducir código para el entrenamiento de Copilot[7].

– *Enero de 2023*: un grupo de artistas visuales demandó a Stability AI, Midjourney y DeviantArt por presuntamente violar derechos de autor al crear imágenes con base en el estilo de esos artistas sin licencia[8].

5 Henner Zeller, Lizzi Sassman y Gary Illyes. "Protocolo para formalizar la especificación del protocolo de exclusión para robots" (*Blog del Centro de la Búsqueda de Google*, 2019). [En línea]. [Consulta: 10 de marzo de 2024].

6 Originality.ai. "Websites that have blocked OpenAI's GPTBot CCBot Anthropic Google Extended - 1000 Website Study". 2024 [En línea]. [Consulta: 15 de marzo de 2024].

7 James Vincent. "The lawsuit that could rewrite the rules of AI copyright". *Artificial Intelligence, The Verge*. 2022. [En línea]. [Consulta: 23 de febrero de 2024]. Caso: Doe et al v. GitHub, Inc. et al, U.S. District Court for the Northern District of California, n.º 4:22-cv-06823 (N.D. Cal. Nov 03, 2022).

8 Blake Brittain. "Lawsuits accuse AI content creators of misusing copyrighted work". *Reuters*. [En línea]. 2023. [Consulta: 23 de febrero de 2024]. Caso: Andersen v. Stability AI Ltd, U.S. District Court for the Northern District of California, n.º 3:23-cv-00201.

– *Febrero de 2023*: Getty Images demandó a Stability AI por presuntamente violar derechos de autor al usar más de doce millones de imágenes de Getty para entrenar al modelo de generación de imagen de Stable Diffusion AI[9].

– *Octubre de 2023*: Universal Music Publishing Group, Concord Music Group y ABKCO demandaron a Anthropic por presuntamente violar derechos de autor al usar letras de canciones para entrenar su modelo de inteligencia artificial[10].

– *Diciembre de 2023*: New York Times demandó a Microsoft y a OpenAI por presuntamente violar derechos de autor al usar artículos para entrenar su modelo de inteligencia artificial[11]. Con argumentos similares, organizaciones de noticias como The Intercept, Raw Story y AlterNet demandaron a OpenAI en febrero del 2024[12].

– *Marzo de 2024*: Nvidia fue demandado por tres autores por presuntamente violar derechos de autor al usar sus libros para entrenar a NeMo[13].

Estos casos guardan una relación importante con el *web scraping*, pues como se mencionó, este es usado en la recopilación de datos de acceso público en Internet (p. ej., código, imágenes, letras de canciones, artículos, etc.) para entrenar a sus modelos de inteligencia artificial.

Al margen de modelos de inteligencia artificial generativa, las Cortes en Estados Unidos han analizado cuestiones relacionadas con *web scraping* en diversas oportunidades. Antes de 2000, existía una interpretación desfavorable, según la cual su uso constituía fraude informático[14]. De 2000 a 2009, se observó una tendencia más permisiva, según la cual no bastaba con que los sitios web prohibieran su uso, sino que era necesario que aplicaran medidas técnicas para restringir el acceso a quienes las utilizaran[15] (p. ej., bloquear *crawlers*).

9 Blake Brittain. "Getty Images lawsuit says Stability AI Misused photos to train AI" (Reuters, 2023). [En línea]. [Consulta: 23 de febrero de 2024]. Caso: Getty Images (US) Inc v. Stability AI Inc, U.S. District Court for the District of Delaware, n.º 1:23-cv-00135.

10 Tatjana Patern y Layna Deneen. "AI Threats Emerge in Music Publishers' Battle with Big Tech". [En línea]. 2024. [Consulta: 23 de febrero de 2024]. Caso: Concord Music Group Inc v. Anthropic PBC, U.S. District Court for the Middle District of Tennessee, n.º 3:23-cv-01092.

11 Emilia David. "Microsoft invokes VCRs in motion to dismiss the New York Times' AI lawsuit". 2024. [En línea]. [Consulta: 10 de marzo de 2024]. Caso: New York Times Co v Microsoft Corp et al, U.S. District Court for the Southern District of New York, n.º 23-11195.

12 Blake Brittain. "OpenAI hit with new lawsuits from news outlets over AI training". 2024. [En línea]. [Consulta: 10 de marzo de 2024]. Caso: Raw Story Media v. OpenAI Inc, U.S. District Court for the Southern District of New York, n.º 1:24-cv-01514 y The Intercept Media Inc v. OpenAI Inc, U.S. District Court for the Southern District of New York, n.º 1:24-cv-01515.

13 Jonathan Stempel. "Nvidia is sued by authors over AI use of copyrighted works". 2024. [En línea]. [Consulta: 15 de marzo de 2024]. Caso: Nazemian et al v Nvidia Corp, U.S. District Court, Northern District of California, n.º 24-01454.

14 Andrew Sellars. "Twenty Years of Web Scraping and the Computer Fraud and Abuse Act" (Boston: Boston Univeristy School of Law, Scholarly Commons at Boston University School of Law, 2018), 379.

15 *Ibid*., 380.

En una sentencia del Tribunal del Noveno Circuito de Apelaciones de Estados Unidos, en el caso de hiQ Labs, Inc. vs. LinkedIn Corp., se definió que el *web scraping* es legal en datos que se pueden acceder públicamente en Internet[16]. Para el Tribunal, "fraude informático" es el acceso no autorizado a un sistema informático, por lo que en estos casos se exige que el operador de sitios web demuestre el haber intentado restringir el acceso al *web scraper* y que este, no obstante, ha persistido en sus acciones.

El avance más reciente en estos temas viene de Meta, la compañía detrás de Instagram, Facebook y WhatsApp. Meta demandó a Bright Data, compañía Israelí de tecnología, por violar los términos de servicio de sus plataformas al utilizar la técnica de recolección. En una solicitud de Meta para sentencia sumaria parcial, la Corte dispuso que Meta no tenía la suficiente evidencia para probar que Bright Data utilizara sus propias cuentas de Facebook e Instagram para el *scraping*[17]. Esto significa que Bright Data no actuaba como "usuario" de los servicios en el momento en que realizaba el *scraping*, sino solo como "visitante", por lo que no le son exigibles los términos de servicio de Meta. En febrero de 2024, Meta retiró la demanda con Bright Data, lo que significó una victoria para la comunidad de *web scraping*, según Or Lenchner, CEO de Bright Data[18].

En todo caso, no hay unanimidad en relación con la posición de la industria de la tecnología sobre este método de reelección de datos. Aun cuando hay demandas abogando en contra del *web scraping*, se pueden encontrar compañías que han licenciado su contenido a modelos de inteligencia artificial. A continuación, se resume a grandes rasgos algunos de los avances más significativos en ese sentido:

– OpenAI licenció datos de Shutterstock, compañía competidora de Getty Images, a través de una alianza que anunciaron en octubre del 2022[19]. Esa alianza se extenderá por seis años, consolidando a Shutterstock como el proveedor líder de datos de imágenes para modelos de inteligencia artificial de OpenAI[20].

16 Zack Whittaker. "Web scraping is legal, US appeals court reaffirms". *TechCrunch*. 2022. [En línea]. [Consulta: 10 de febrero de 2024]. Caso: hiQ Labs, Inc. v. LinkedIn Corporation, U.S. Court of Appeals for the Ninth Circuit.

17 Sarah Perez. "Court rules in favor of a web scraper, Bright Data, which Meta had used and then sued". *TechCrunch*. 2024. [En línea]. [Consulta: 15 de marzo de 2024]. Caso: Meta Platforms, Inc. v. Bright Data Ltd., U.S. District Court for the Northern District of California, n.º 3:23-cv-00077-EMC.

18 Sarah Perez. "Meta drops lawsuit against web-scraping firm Bright Data that sold millions of Instagram records". *TechCrunch*. 2024. [En línea]. [Consulta: 15 de marzo de 2024]. Caso: Meta Platforms, Inc. v. Bright Data Ltd., U.S. District Court for the Northern District of California, n.º 3:23-cv-00077-EMC.

19 Shutterstock. "Shutterstock partners with OpenAI and leads the way to bring AI-Generated content to all". *Comunicado de prensa*. 2022. [En línea]. [Consulta: 15 de marzo de 2024].

20 Shutterstock. "Shutterstock expands partnership with OpenAI, signis new six-year agreement to provide high-quality training data". *Comunicado de prensa*. 2023. [En línea]. [Consulta: 15 de marzo de 2024].

– Associated Press y OpenAI han llegado a un acuerdo para licenciar determinados contenidos de Associated Press, mientras se examinan posibles casos de uso de inteligencia artificial generativa en productos y servicios informativos[21].

– Reddit firmó un acuerdo con Google para hacer disponible su contenido para el entrenamiento de los modelos de inteligencia artificial de Google[22].

En Colombia, la discusión es aún incipiente. No ha habido avances regulatorios o jurisprudenciales relacionados con el *web scraping* y mucho menos con la información que se recopila para entrenar modelos de inteligencia artificial.

La Superintendencia de Industria y Comercio (SIC), sin embargo, en su rol de autoridad nacional de protección de datos personales, investigó a LinkedIn Corp. y a LinkedIn Ireland por asuntos de *web scraping* a finales de 2023. La SIC ordenó que esas compañías reforzaran sus medidas de seguridad para proteger los datos personales y la información de sus usuarios en Colombia.

La SIC entendió el *web scraping* como una brecha de seguridad. Según la Delegatura para la Protección de Datos, dado que la extracción masiva de datos personales se realiza normalmente por medios automatizados, la técnica constituye un riesgo permanente para el debido tratamiento de la información personal en una plataforma como LinkedIn, que a enero de 2023 contaba con doce millones de usuarios activos en Colombia.

En la Resolución 71406 de 2023 (que no está disponible públicamente), la SIC impartió órdenes que conminan a LinkedIn a garantizar la seguridad de los datos personales de los usuarios, evitando su: (1) acceso no autorizado o fraudulento; (2) uso no autorizado o fraudulento; (3) consulta no autorizada o fraudulenta; (4) adulteración o (5) pérdida.

Sin perjuicio de que la Resolución 71406 de 2023 no está disponible públicamente, los argumentos esgrimidos por la SIC en su artículo de prensa no son consecuentes con el desarrollo jurisprudencial y de industria que se ha dado en Estados Unidos sobre el *web scraping*. Según el Tribunal del Noveno Circuito de Apelaciones de Estados Unidos, en un caso que también concierne a LinkedIn, el *web scraping* no configura un "fraude informático" (en los términos del *Computer Fraud and Abuse Act* de Estados Unidos), en la medida en que el acceso al sitio web sea autorizado y los datos sean de acceso público.

II. LOS TÉRMINOS DE SERVICIO Y EL *WEB SCRAPING*

Con base en los avances que se han dado en materia de *web scraping* en el contexto internacional, se entiende que la discusión legal reviste asuntos contractuales. En

21 Associated Press. "AP, OpenAI agree to share select news content and technology in new collaboration". *Comunicado de prensa*. 2023. [En línea]. [Consulta: 15 de marzo de 2024].

22 Anna Tong, Echo Wang y Martin Coulter. "Exclusive: Reddit in AI content licensing deal with Google". *Reuters*. 2024. [En línea]. [Consulta: 15 de marzo de 2024].

efecto, la viabilidad legal del uso de tecnologías de *web scraping* en una determinada página depende, en gran medida, de los términos de servicio del sitio web.

Para abordar este argumento, en primer lugar, es necesario explicar dos tipos de contratos que se generan en Internet: (1) los *click-wrap*, que se refieren a aquellos que se crean cuando un usuario hace clic en un icono dentro de un sitio web para dar su consentimiento al contenido del contrato propuesto por el operador del sitio web, (2) los *browse-wrap*, entendidos como un conjunto de condiciones presentadas en un sitio web que no promueven ninguna manifestación externa y explícita de consentimiento, salvo el que se supone por la simple navegación por el sitio web[23].

En los *click-wrap*, el análisis se torna caso-a-caso. Deben revisarse los términos de servicio de los sitios web para identificar si hay alguna prohibición expresa que restrinja el uso de técnicas de *web scraping*. Legalmente, sin embargo, podría argumentarse que en casos en los que la prohibición es ambigua y no se refiere textualmente a *web scraping* (p. ej., que se prohíban técnicas de ingeniería inversa y tecnologías asociadas), si las cláusulas fueron dictadas por una de las partes (predisponente en este caso, es decir, el operador de la página web), se deben interpretar en contra de ella de conformidad con el artículo 1624 del Código Civil.

En relación con los *browse-wrap*, las cortes sostienen, por regla general, que el usuario no manifiesta su consentimiento de forma expresa[24]. En Colombia, de acuerdo con el artículo 1502 del Código Civil, la existencia de un contrato depende de cuatro requisitos, entre los que está el consentimiento. La cuestión en esta modalidad de contratación está en el consentimiento, por lo que si se acepta que en los *browse-wrap* el usuario no manifiesta su consentimiento, el contrato no existiría y el usuario no estaría sujeto a los términos de servicio de la página web.

El consentimiento tiene dos elementos: (1) uno subjetivo, que consiste en la voluntad consciente de producir consecuencias de derecho y (2) uno objetivo, que consiste en una conducta observable que tiene por propósito expresar dicha voluntad[25]. En los *browse-wrap* no se forma ninguno de los dos elementos, por cuanto la persona que visita un sitio web sin registrarse no pretende, en principio, formar una relación jurídica con el operador del sitio web. Lo anterior excluye el elemento subjetivo y, de contera, el objetivo, debido a que no hay una voluntad que pueda observarse mediante una conducta expresa.

Por razón de lo anterior, en tanto el usuario puede acceder a la información de una página web sin dar su consentimiento expreso (p. ej., al registrarse como usuario de esta y así aceptar los términos y condiciones correspondientes), el *web*

23 Fidel Usma. "El consentimiento en los contratos en línea B2C y su protección bajo la ley colombiana". *Cuadernos de la Maestría en Derecho n.° 5* (Bogotá: Universidad Sergio Arboleda), 306.

24 Aaron Rubin y Jackie Li. "Court discovers rare and elusive "enforceable browsewrap". *JDSupra*. 2020. [En línea]. [Consulta: 10 de febrero de 2024].

25 Fidel Usma. "El consentimiento en los contratos en línea B2C y su protección bajo la ley colombiana". *Cuadernos de la Maestría en Derecho n.° 5* (Bogotá: Universidad Sergio Arboleda), 291. Cita a Brehm en su libro *Allgemeiner Teil des BGB*.

scraping es válido y no se sanciona legalmente, incluso si los términos de servicio incluyen una prohibición expresa de la técnica. En la medida en que no se forma un contrato por el hecho de que el visitante acceda al sitio web, los términos de servicio no le son exigibles y el *web scraping* es legalmente viable.

Esta conclusión se ajusta al entendimiento de los tribunales en Estados Unidos que, en casos como *hiQ Labs v. LinkedIn* y *Meta v. Bright Data*, han dispuesto que cuando no existe un sistema de autenticación en las páginas web, terceros pueden acceder a la información pública a través de *web scraping*. El caso de *Meta v. Bright Data* es particularmente interesante, dado que el Tribunal no accedió a los argumentos de que usar herramientas automatizadas para saltarse las restricciones de acceso, como los CAPTCHA, era lo mismo que acceder a un "sitio web protegido con contraseña". Para el Tribunal, debe haber algún tipo de aceptación expresa a los términos de servicio (p. ej., el registro en la plataforma o el uso de la plataforma a través del inicio de una sesión) para violar a estos. De lo contrario, si la información es de acceso público y no requiere de algún tipo de privilegio de acceso es legal.

III. RIESGOS LEGALES DEL *WEB SCRAPING*

Si bien, como fue mencionado, actualmente no se han presentado avances jurisprudenciales o regulatorios de manera específica relacionados con el *web scraping* como técnica de recolección de datos destinados al entrenamiento de modelos de inteligencia artificial, estas prácticas pueden representar riesgos legales. En efecto, aun si el *web scraping* no está prohibido expresamente en Colombia, su implementación podría tener implicaciones en materia de derechos de autor, derecho marcario, privacidad e incluso derecho penal.

A. El uso de obras protegidas en técnicas de *web scraping*

Los derechos de autor, por ejemplo, protegen los derechos de autores sobre sus obras, entendidas, en los términos de la Decisión 351 de la Comunidad Andina, como toda creación artística o literaria intelectual original materializada de cualquier forma perceptible. El régimen de derechos de autor también cobija a los titulares de derechos conexos, por ejemplo, intérpretes, ejecutantes de obras, productores y organismos de radiodifusión[26]. Por otro lado, frente a los derechos de autor se presenta una ausencia de formalidad, es decir, los derechos se adquieren con la materialización de la obra y no requieren de registro alguno[27]. Es así como una expresión artística o literaria que revista originalidad y pueda ser reproducida o divulgada por cualquier medio adquiere protección con su creación misma[28]. Los objetos de protección de este régimen pueden ir desde fotografías, artículos

26 Congreso de la República de Colombia. Ley 23 de 1982.
27 *Ibid*.
28 Dirección Nacional de Derechos de Autor. Resolución 11 de 2017.

periodísticos, películas, dibujos, pinturas, bases de datos (que contengan elementos de originalidad en su disposición), fonogramas hasta programas de ordenador y obras arquitectónicas.

Con lo anterior en mente, se ha considerado que cuando estamos frente a datos que reúnen los elementos requeridos para ser considerados obras protegidas por derechos de autor, la posibilidad de extraer, almacenar, procesar y usar estos datos publicados en Internet puede constituir una infracción a los derechos de sus titulares[29]. Se ha sostenido que el almacenamiento de los datos protegidos recolectados podría considerarse una reproducción no autorizada de las obras, pues involucra la realización de una copia no autorizada de estas[30]. El *web scraping* extrae datos en Internet (algunos protegidos por derechos de autor) y los almacena en una base de datos, generando una copia que supone una reproducción en los términos de derechos de autor.

Sin perjuicio de lo anterior, el régimen de derechos de autor prevé un sistema de límites y excepciones. Este sistema se concibe como una lista de eventos en los que se permite el uso sin autorización del titular de la obra. La Decisión 351 de 1993 de la Comunidad Andina de Naciones dispone que la aplicación de los límites y excepciones no debe causar un perjuicio injustificado al autor y debe permitir la explotación normal de la obra[31]. Además, dado que representan una restricción a los derechos de autor, son de aplicación excepcional y restrictiva[32]. Estos límites y excepciones se encuentran consagrados en los artículos 31 a 42 de la Ley 23 de 1982 y en el artículo 22 de la Decisión 351 de 1993[33]. Ahora, ni la jurisprudencia ni la legislación colombiana ha dado una aplicación a estos límites y excepciones en lo que respecta al *web scraping* y el uso de datos para entrenamiento de sistemas de inteligencia artificial.

En todo caso, en el contexto internacional se ha revisado estos casos desde otros sistemas legales. En Estados Unidos se aplica la doctrina del *fair use* para definir aquellos eventos en los que para el uso de una obra protegida no se requiere la autorización del titular[34]. Se ha entendido el uso transformativo como uno

29 Vlad Kroto, Leigh Redd, y Leiser Silva. "Tutorial: Legality and Ethics of Web Scraping" (Communications of the Association for Information Systems, 2020).

30 Philipp Hacker. "A legal framework for AI training data—from first principles to the Artificial Intelligence Act". *2020* [En línea] [Consulta: 1 de febrero de 2024].

31 Decisión 351 de la Comunidad Andina de Naciones

32 Rahn De Frutos. "Excepciones y limitaciones al derecho de autor en Colombia: propuestas legislativas". 2014. [En línea]. [Consulta: 10 de marzo de 2024].

33 Entre estos se encuentra, el uso de la obra siempre que se cite adecuadamente, el uso relacionado con publicaciones de noticias, el uso con fines educativos, el uso de obras relacionadas con eventos de actualidad, el uso de discursos pronunciados públicamente, el uso de retratos con fines científicos, didácticos, culturales o de interés público, la reproducción para uso privado, la reproducción de las bibliotecas de obras para uso de sus lectores, fines de conservación o préstamo a otras bibliotecas, la reproducción de obras ubicadas en lugares públicos, la anotación de conferencias por parte de estudiantes, la reproducción de textos legales, el uso en procesos judiciales, modificaciones de proyectos arquitectónicos.

34 US. Copyright Office. "U.S. Copyright Office Fair Use Index" *2023* [En línea]. [Consulta: 10 de marzo de 2024]

de los presupuestos para el *fair use*, por lo que se ha usado como defensa para el entrenamiento de la inteligencia artificial con obras protegidas. En cualquier caso, esta doctrina no tiene la aplicación en Colombia, por lo que no es posible trasladar directamente estos argumentos a nuestro sistema legal.

Un elemento que se debe tener presente de cara a este riesgo es que no todos los datos son protegidos por los derechos de autor, pues para que una obra sea protegida debe ser una creación artística o literaria intelectual original materializada de cualquier forma perceptible. En el caso de no ser un dato que revista las características necesarias para ser considerado una obra, la propiedad intelectual no sería el instrumento de protección idóneo a la hora de determinar la viabilidad de usar técnicas de *web scraping* para la extracción de datos para entrenamiento de herramientas de inteligencia artificial. Por consiguiente, no se presentaría el riesgo de infracción a derechos de autor en este supuesto, pues los datos obtenidos carecerían de dicha protección.

Adicionalmente, existen obras que se encuentran en el dominio público[35], es decir, carecen de las restricciones impuestas por la propiedad intelectual por lo que cualquiera las puede usar sin necesidad de obtener permiso de tu titular[36]. En la medida en que estas obras son de libre acceso, no se concretaría el riesgo de infracción. Como resultado, en cuanto al régimen de propiedad intelectual no existiría un obstáculo para el uso de las técnicas de recolección de datos en obras de dominio público.

En pocas palabras, cuando el *web scraping* extrae obras protegidas por el derecho de autor, existe la posibilidad de que quien emplea la técnica incurra en vulneraciones a los derechos de sus autores. Es claro que esta discusión no ha sido definida en su totalidad y mucho menos ha sido discutida completamente en espacios legislativos o jurisdiccionales en Colombia. No obstante, se podría determinar que cuando la información recolectada hace parte del dominio público o existe alguna licencia de su titular, el *web scraping* sería legítimo en cuanto a derechos de autor respecta.

B. El uso de marcas en técnicas de web scraping

Las páginas web en las que se lleva a cabo el *web scraping* pueden contener marcas y otros signos distintivos. Se entienden como marcas aquellos signos que distinguen productos o servicios en el mercado, susceptibles de representación gráfica, que pueden ser nominativos (conformada por una o varias palabras distintivas), figurativos (conformada por dibujos o imágenes), mixtos (combinaciones de figuras

35 El artículo 187 de la Ley 23 de 1982 dispone que las obras se encuentran en este estado son las obras cuyo periodo de proyección haya culminado, las obras folclóricas y tradicionales de autores desconocidos, las obras cuyos autores hayan renunciado a sus derechos.

36 Christian Schmitz. "Propiedad intelectual, dominio público y equilibrio de intereses". *Revista Chilena de Derecho* 36, n.° 2 (2009).

y palabras) e incluso otros no tradicionales como marcas de color o sonoras. El registro de una marca otorga el derecho de uso exclusivo en el territorio para que se concede y el derecho de impedir a cualquier tercero, entre otras cosas, el uso de la marca. Sobre este último punto recaerá nuestro análisis.

La marca puede ser usada por terceros con distintos propósitos: (1) a título de marca, (2) como otro tipo de signo distintivo y (3) con finalidades distintas a la identificación comercial[37]. En el contexto de *web scraping*, las marcas no son usadas como signos distintivos, es decir, su uso no busca la identificación de un producto o servicio. Por el contrario, se hace para incluir a la marca dentro de la base de datos que entrena al modelo de inteligencia artificial.

El artículo 157 de la Decisión 486 del 2000 de la Comunidad Andina de Naciones establece que terceros pueden referenciar a una marca con propósitos de identificación o de información. Ese uso, además, debe hacerse de buena fe, no puede ser a título de marca y no puede ser apto para inducir al público a confusión sobre la procedencia empresarial de los productos o servicios.

El uso de marcas a través de técnicas de *web scraping* se da únicamente para efectos de entrenar al modelo de inteligencia artificial. Este uso no es público ni se presenta como un uso no autorizado del signo en el comercio y no es susceptible de generar confusión sobre otros productos o servicios, por lo que puede catalogarse como de buena fe. Bajo ese entendido, el uso de marcas en técnicas de *web scraping* no constituye una infracción a derechos marcarios.

C. Los datos personales en técnicas de *web scraping*

Las páginas web en las que se lleva a cabo el *web scraping* pueden contener datos personales, por lo que a continuación se analizan las preocupaciones de privacidad que esta técnica podría generar. Los datos personales, entendidos como "cualquier pieza de información vinculada a una o varias personas determinadas o determinables"[38], se caracterizan por

> i) estar referido a aspectos exclusivos y propios de una persona natural, ii) permitir identificar a la persona, en mayor o menor medida, gracias a la visión de conjunto que se logre con el mismo y con otros datos; iii) su propiedad reside exclusivamente en el titular del mismo, situación que no se altera por su obtención por parte de un tercero de manera lícita o ilícita, y iv) su tratamiento está sometido a reglas especiales (principios) en lo relativo a su captación, administración y divulgación.[39]

37 Ricardo Metke. Lecciones de propiedad industrial (III) (Baker & McKenzie, 2006), 138.

38 Congreso de la República de Colombia. Ley 1266 de 2008.

39 Corte Constitucional. Sentencia SU139 de 2021, 14 de mayo de 2021, M.P. Jorge Enrique Ibáñez Najar.

En Colombia, la SIC ha considerado que el *web scraping* puede constituir un riesgo para el debido tratamiento de información personal, incluso si la información recolectada es de acceso público[40]. En múltiples oportunidades se ha recalcado que el hecho de que los datos personales sean de acceso público no los convierte en información de naturaleza pública. En palabras de la autoridad "recolectar datos personales privados, semiprivados o sensibles en internet no legitima al recolector para apropiarse de dicha información y hacer lo que quiera con la misma[41].

Con miras a la importante distinción anteriormente expuesta, los datos personales pueden ser categorizados como privados, semiprivados o públicos. Así, los datos privados son aquellos que por su naturaleza son relevantes solo para su titular, mientras que los datos semiprivados son relevantes para su titular y un determinado grupo de personas, por ejemplo, datos crediticios o financieros[42].

Los datos públicos han sido definidos con base en un criterio residual, pues son datos no catalogados como semiprivados o privados, esto comprende, pero no se limitan a ello, información sobre el estado civil de las personas, contenida en documentos públicos o en sentencias judiciales[43]. Efectivamente, frente a los datos públicos se presenta una excepción dado que para su tratamiento no se requiere autorización de su titular[44]. Por lo tanto, siempre que se cumpla con los demás principios y normas aplicables al tratamiento de datos personales, se podrán implementar técnicas de *web scraping* para la recolección de datos públicos para el entrenamiento de sistemas de inteligencia artificial.

El régimen general de protección de datos personales en Colombia, contenido en la Ley 1581 de 2012 y el Decreto 1074 de 2015, dispone que cualquier tratamiento (que incluye la recolección, uso, circulación, almacenamiento, supresión, transmisión, transferencia y cualquier otra operación realizada sobre datos personales) debe ser autorizado. Ahora bien, la autorización debe ser concedida de manera previa (anterior al tratamiento), expresa (manifestada por escrito, de manera oral o a través de conductas inequívocas que revelen la intención del titular; el silencio en ningún caso se asimila a una conducta inequívoca) e informada por el respectivo titular.

El titular debe ser informado, al momento de obtener su autorización, de acuerdo con el régimen general de protección de datos personales en Colombia, sobre (1) el tipo de tratamiento al que serán sometidos sus datos personales; (2) las finalidades para las que serán procesados sus datos personales; (3) sus derechos como titular de datos personales y la forma de ejercerlos; (4) datos de identificación, dirección física o electrónica y teléfono del responsable del tratamiento (es decir,

40 Superintendencia de Industria y Comercio. Resolución 58834 de 2023.
41 *Ibid*.
42 Superintendencia de Industria y Comercio. "*Protección de datos personales: aspectos prácticos sobre el derecho de hábeas data*". 2022 [En línea] [Consulta: 10 de marzo de 2024].
43 Congreso de la República de Colombia. Artículo 3, Ley 1266 de 2008.
44 Congreso de la República de Colombia. Artículo 10, Ley 1581 de 2012.

la entidad que decide sobre las bases de datos y el tratamiento de la información ahí contenida; (5) los datos personales que serán recolectados y procesados; (6) el lugar donde podrá consultar la política de tratamiento de la información del responsable, y (7) en caso de procesar datos sensibles (p. ej., datos cuyo tratamiento pueda afectar la intimidad o puede resultar en su discriminación), indicación sobre que el tratamiento de esta información es enteramente facultativo.

Sin embargo, la Ley 1581 de 2012 dispone ciertos eventos en los que no se requiere la autorización del titular para el tratamiento de los datos personales. Las excepciones consagradas en el artículo 10 de dicha ley incluyen (1) información requerida por una entidad pública o judicial en ejercicio de sus funciones, (2) los datos de naturaleza pública, (3) casos de urgencia médica, (4) tratamiento de datos autorizado por la ley para fines históricos, estadísticos o científicos y (5) derechos relacionados con el Registro Civil.

Con base en lo anterior, la SIC podría sancionar conductas en las que se traten datos personales sin la respectiva autorización, el tratamiento exceda la finalidad informada o el tratamiento se realice de forma distinta a la que el titular autoriza[45]. Así, la utilización de datos personales privados o semiprivados en técnicas de *web scraping* requiere de autorización previa, expresa e informada en los casos que dispone la ley.

D. Implicaciones penales en el uso de técnicas de *web scraping*

El Código Penal consagra una serie de conductas tipificadas como delitos informáticos que resultan altamente relevantes en este contexto. De acuerdo con Alberto Suárez Sánchez:

> Mediante este bien jurídico se responde desde el ámbito penal a la puesta en peligro de nuevas logísticas necesitadas y merecedoras de protección, como los datos y la información, y al interés colectivo en la seguridad y fiabilidad en su almacenamiento, tratamiento, procesamiento y transferencia en los sistemas.[46]

Con la promulgación de la Ley 1273 de 2009, se buscaba proteger el bien jurídico de la protección de la información y los datos mediante la armonización del ordenamiento colombiano con políticas globales contra de los delitos en el ciberespacio[47]. Por la gran afectación que pueden producir estos delitos en diversos derechos que trascienden la información y los datos, la doctrina y la jurisprudencia han considerado que son delitos pluriofensivos, pues además de atentar contra

45 Corte Constitucional. Sentencia T-020 de 2014, 27 de enero de 2014. M.P. Luis Guillermo Guerrero Pérez.

46 Alberto Suárez. Delitos informáticos. "Lecciones de derecho penal: parte especial" (Bogotá: Universidad Externado de Colombia, 2014), 13-72.

47 Sala de Casación Penal. Corte Suprema de Justicia. Sentencia SP2699-2023, 3 de agosto de 2023 M.P. Fernando León Bolaños Palacios.

estos bienes jurídicos también arremeten contra la intimidad, la libertad de la información, la honra, la integridad de los documentos y el patrimonio económico, entre otros[48].

El primero de los delitos que podría tener aplicación en el uso de *web scraping* es el de acceso abusivo a un sistema informático. De acuerdo con el artículo 269A del Código Penal, consiste en acceder en todo o en parte de un sistema informático o mantenerse dentro de este sin autorización o fuera de la autorización dada. La Sala de Casación Penal de la Corte Suprema de Justicia ha considerado este tipo penal como un "*hacking* directo o mero intrusismo". Así, la conducta se configura con la mera intrusión ilegítima al sistema[49]. Para la configuración de este delito es necesario acceder a un sistema restringido, en el caso del *web scraping* que se realiza en páginas de acceso público o en cumplimiento de los límites autorizados por el titular del sistema informático en sus términos de uso, no estaríamos frente a un uso delictivo del *web scraping*[50]. En suma, para que el *web scraping* sea relevante para el derecho penal, en este contexto, debe presentarse en un sistema informático para el cual su administrador haya dispuesto restricciones de acceso o permanencia.

Otro tipo penal cuyo estudio es pertinente es el de obstaculización ilegítima de sistema informático o red de telecomunicación, consagrado en el artículo 269B del Código Penal. La conducta proscrita por este tipo es la obstaculización del funcionamiento o acceso normal a un sistema informático, a los datos que contiene o a una red de telecomunicaciones. De acuerdo con las interpretaciones doctrinales de este delito, su conducta requiere imposibilitar o entorpecer el acceso o el funcionamiento del sistema informático. Se ha reconocido que algunos usos del *web scraping* pueden dar lugar a un daño al sistema informático que podrían resultar en la obstrucción de su acceso[51]. Por ejemplo, en Estados Unidos se han analizado casos en que el *web scraping* ha generado una sobrecarga del sistema como en Dryer and Stockton 2013[52]. En el evento en que se genere una obstrucción al funcionamiento o acceso normal al sistema de manera dolosa con el uso de la técnica podría configurarse este tipo penal.

El artículo 269D incorpora el delito de daño informático entendido como la destrucción, daño, eliminación, deterioro, alteración o supresión de datos informáticos o un sistema de tratamiento de información o sus partes lógicas. Este

48 Alberto Suárez. Delitos informáticos. "Lecciones de derecho penal: parte especial" (Bogotá: Universidad Externado de Colombia, 2014), 13-72.

49 Sala de Casación Penal. Corte Suprema de Justicia. Sentencia SP592-2022, 2 de marzo de 2022 M.P. Diego Eugenio Corredor Beltrán.

50 Johan Sanabria. *Sector privado y libre competencia: implicaciones jurídicas del web scraping* (Bogotá: Universidad Externado de Colombia, 2021).

51 Ajay Bale, Naveen Ghorpade, S. Kamalesh, S. Rohith, R. Rohith y S. Rohan. "Web Scraping Approaches and their Performance on Modern Websites" (Coimbatore, India: Proceedings of the Third International Conference on Electronics and Sustainable Communication Systems, 2022).

52 Vlad Kroto, Leigh Redd y Leiser Silva. "Tutorial: Legality and Ethics of Web Scraping" (Communications of the Association for Information Systems, 2020).

delito solo admite la modalidad dolosa por lo que frente a una comisión dolosa de la conducta descrita por medio de las técnicas de *web scraping* podría existir una responsabilidad penal.

Finalmente, el artículo 269F del Código Penal dispone que

> El que, sin estar facultado para ello, con provecho propio o de un tercero, obtenga, compile, sustraiga, ofrezca, venda, intercambie, envíe, compre, intercepte, divulgue, modifique o emplee códigos personales, datos personales contenidos en ficheros, archivos, bases de datos o medios semejantes, incurrirá en pena de prisión de cuarenta y ocho (48) a noventa y seis (96) meses y en multa de 100 a 1.000 salarios mínimos legales mensuales vigentes.

Retomando lo expuesto anteriormente, en caso de vulnerar los derechos de titulares sobre sus datos personales, por ejemplo, realizando un tratamiento no autorizado de datos personales, podría enmarcarse dentro la una conducta ilícita descrita en este tipo penal.

En resumen, los delitos informáticos requieren de la concurrencia del dolo como elemento subjetivo del tipo penal. Por lo anterior, se descarta que un comportamiento realizado sin el objetivo de afectar los bienes jurídicos tutelados carece de relevancia penal. Bajo esta perspectiva, cuando se accede a una página web en los términos permitidos por su administrador y se recolectan datos sin afectar los derechos que el Régimen General de Protección de Datos Personales confiere a sus titulares, no habría lugar a la aplicación del derecho penal. Así mismo, se descarta la configuración de estos delitos cuando la técnica se emplea respecto de información pública. Finalmente, si bien se ha reconocido que esta técnica puede sobrecargar los sistemas informáticos, el eventual daño que se genere podría dar lugar a la imposición de la sanción legal consagrada en el artículo 269B en la medida en que resulte en la obstrucción del sistema informático.

IV. LICENCIAS EN INTERNET, DOMINIO PÚBLICO Y *WEB SCRAPING*

A. *Open source, open access, creative commons*

En 2022, Microsoft, GitHub y OpenAI fueron demandados por reproducir códigos para el entrenamiento de Copilot. Esta demanda se basa en que la reproducción de los códigos se realizó sin observancia de las condiciones establecidas en las licencias otorgadas por sus autores, pues estas requerían dar crédito a los creadores de los programas. Este caso presenta un interesante debate sobre el *web scraping* y las licencias que se otorgan en Internet.

Las licencias de código abierto (*open source*) otorgan a los usuarios la libertad de ejecutar el programa con cualquier fin, estudiarlo y adaptarlo, distribuirlo o

modificarlo para mejorarlo[53]. Estas pueden incluir restricciones a la forma en que los usuarios pueden ejercer dichas libertades. Por ejemplo, pueden establecer que en caso de modificaciones o re-distribuciones, estas deben seguir la licencia del código inicial[54].

Además del *open source* existen otros tipos de licencias que permiten un acceso libre a la información. Entre estos se encuentra el acceso abierto u *open access*, un concepto que ha adquirido gran relevancia en el contexto de los recursos digitales y se refiere al acceso gratuito y sin restricción a información publicada en línea[55]. La difusión de las obras de *open access* se caracteriza por permitir el acceso libre y universal sin costo alguno y el otorgamiento del titular del derecho de usar, copiar o distribuir las obras a los usuarios[56]. El *open access* busca la compartición del conocimiento para reducir la brecha de información y así contribuir a la generación del conocimiento.

De acuerdo con opendefinition.org citado por la UNESCO, una obra se considera de *open access* si: (1) es accesible a todos, (2) no tiene restricciones para su redistribución ni utilización, (3) no presenta discriminaciones y (4) se confieren los permisos para desarrollar las prerrogativas protegidas por derechos de autor. Igualmente, las licencias pueden contener provisiones sobre el mantenimiento de la integridad de la obra y el reconocimiento de su derecho de paternidad[57]. En consecuencia, ya que por medio de estas iniciativas de *open access* se brinda una licencia para re-usar y re-distribuir las obras y la información en general, el *web scraping* se acomodaría a sus objetivos de compartición y creación de conocimiento.

Por lo anterior, en la medida en que se respeten las condiciones de la licencia *open source* u *open access*, el uso de *web scraping* para obtener datos destinados al entrenamiento de herramientas de inteligencia artificial es legalmente viable. La demanda en contra de Microsoft, GitHub y OpenAI, en todo caso, va a ser un hito en relación con los lineamientos que se deben tener en cuenta en relación con *web scraping* y licencias en Internet.

Con miras a la consecución del fin de compartir conocimiento, información y cultura en interés público surgieron organizaciones como *Creative Commons* (en adelante, CC).[58] Esta organización proporciona un sistema de licencias promueve el conocimiento abierto por medio de licencias sobre los siguientes aspectos: (1) reconocimiento del autor, (2) restricción al uso comercial de la obra, (3) restricciones a distribución de obras derivadas (4) distribución de obras derivadas únicamente

53 W. R. Ríos Ruiz. "Aspectos legales del software libre o de código abierto (open source)". *Revista La Propiedad Inmaterial*, (2003): 41-60.

54 Andrew St. Laurent. "Understanding Open Source and Free Software Licensing" (Sebastopol, USA: O'Reilly Media, Inc., 2004).

55 Thomas Margini y Diane Peters. "Creative Commons Licenses: Empowering Open Access". 2016 [En línea]. [Consulta: 10 de febrero de 2024].

56 UNESCO. "Concepts of openness and open access". 2015 [En línea]. [Consulta: 15 de marzo de 2024].

57 *Ibid*.

58 Creative Commons. "Who we are". s. f. [En línea]. [Consulta: 15 de marzo de 2024].

mediante la misma licencia[59]. Además, existen licencias CC0 mediante las cuales el autor puede renunciar a todos sus derechos patrimoniales para permitir una distribución, reproducción y modificación de la obra libre por parte de terceros[60]. En esencia, esta licencia busca tener un efecto análogo al del dominio público.

Actualmente, no existe claridad total sobre el alcance de estas licencias en relación con el entrenamiento de inteligencia artificial. En efecto, CC recalca que, si bien es posible conceder y delimitar permisos para la utilización de obras con respecto a facultades propias del derecho de autor, las licencias no pueden usarse para sustituir las limitaciones y excepciones consagradas en la normatividad existente[61]. Por lo tanto, es preciso tener en mente que su ámbito de aplicación también depende de la delimitación de los derechos de propiedad intelectual frente a la inteligencia artificial.

Ahora bien, ya que el autor puede usar una licencia de CC para permitir y definir ciertas acciones de terceros respecto de su obra, las licencias que permiten copiar, redistribuir y adaptar las obras para fines comerciales, en principio, podrían dar lugar a la autorización del uso de estas para entrenar herramientas de inteligencia artificial. En este orden de ideas, CC ha mencionado que

> CC apoya, en principio, el acceso amplio y el uso de obras protegidas por derechos de autor, incluido el contenido con licencia abierta, para entrenar a la IA en beneficio público. Tal acceso puede, por ejemplo, ayudar a reducir el sesgo, mejorar la inclusión, promover actividades importantes como la educación y la investigación, y fomentar la innovación beneficiosa en el desarrollo de la IA.

Sobre este punto es crucial tener en cuenta que si bien estas licencias pueden combinar una serie de autorizaciones de uso e incluso materializar la renuncia a todos los derechos patrimoniales de la obra (licencias CC0), no es posible concluir que en todos los casos las licencias conllevan una autorización de uso sin restricción. Por lo anterior, claramente quienes usen las obras deben atender a lo definido por el autor mismo sobre el uso de su obra.

En este sentido, es preciso distinguir los derechos incluidos en cada licencia para determinar la legalidad del despliegue de técnicas de recolección de datos por medio de *web scraping*. Frente a licencias CC0, en tanto el titular permite a terceros disponer libremente de la obra, el *web scraping* no representaría una infracción a derechos de propiedad intelectual. Por el contrario, en otros tipos de licencia CC, el *web scraping* será una infracción cuando exceda los límites impuestos por el autor en la licencia y será permitido cuando se haga en el marco de lo autorizado.

59 Centro Nacional de Desarrollo Curricular en Sistemas no Propietarios. "Las licencias Creative Commons: qué son, por qué utilizarlas y cómo hacerlo". 2021 [En línea]. [Consulta: 15 de marzo de 2024].

60 Creative Commons. "CC0" s. f. [En línea]. [Consulta: 15 de marzo de 2024].

61 Kat Walsh. "Understanding Cc Licenses and Generative Ai". 2023. [En línea]. [Consulta: 15 de marzo de 2024].

Sin embargo, es notable que este asunto depende en gran medida de la definición de la relación entre la propiedad intelectual, la titularidad de la información y la inteligencia artificial.

B. *Open data*

Open data es un concepto que abarca los datos que son de libre acceso y cuyo uso, explotación y distribución está disponible para cualquier propósito y accesible a todos[62]. Esta información se presenta de manera estructurada y permite un fácil aprovechamiento por parte de los usuarios, pues uno de sus principios es el acceso libre de barreras legales y técnicas. Igualmente, para poder ser accesible fácilmente la información se encuentra publicada en línea[63]. Las primeras aproximaciones al sistema de *open data* se inspiraron en el *open source* y se centraron en la publicidad de información de entidades públicas[64]. Este concepto se expandió y ya no aplica únicamente a información de fuentes públicas, sino a datos del sector privado[65].

Respecto de estos datos se permite la reutilización y distribución de estos por terceros. De la forma en la que se concibió el concepto de *datos abiertos*, estos no deberían contar con ninguna restricción de cara a la privacidad, la seguridad o su titularidad. De manera similar al *open access*, se ha determinado que sus características son que los datos deben ser accesibles como un todo, debe permitirse su uso y redistribución y el acceso debe ser universal[66].

Con la expansión del movimiento de *open data*, la amplitud de facultades que confiere al usuario ha suscitado propuestas encaminadas a promover la utilidad de este tipo de datos para el entrenamiento de sistemas de inteligencia artificial[67]. Con esto en mente, las técnicas de *web scraping* estarían comprendidas dentro del catálogo de acceso y uso autorizado.

C. Información en cabeza del estado: ley de transparencia e infomación pública

En Colombia, el acceso a la información pública es un derecho constitucional[68]. En el marco de este mandato constitucional se han expedido varias normas que rigen la forma en la que interactúan los ciudadanos con las entidades públicas,

62 Terzić Rajko M. y N. Majstorović Milosav. "Open Data Concept, Its Application and Experiences" (Belgrado, Serbia: Vojnotehnički Glasnik / Military Technical Courier, 2019).

63 Opendatasoft. "Open Data". s. f. [En línea]. [Consulta: 10 de febrero de 2024].

64 Opendatasoft. "What is open data - Practical Guide". s. f. [En línea]. [Consulta: 10 de febrero de 2024].

65 Terzić Rajko M. y N. Majstorović Milosav. "Open Data Concept, Its Application and Experiences" (Belgrado, Serbia: Vojnotehnički Glasnik / Military Technical Courier, 2019)

66 Open Knowledge foundation. "What is Open Data?" s. f. [En línea]. [Consulta: 23 de marzo de 2024].

67 Telus International. "The essential guide to AI training data". s. f. [En línea]. [Consulta: 1 de marzo de 2024].

68 Constitución Política, artículo 74.

normas que se centran en los derechos y obligaciones de ambas partes, así como en el tipo de información involucrada.

La Ley 1712 de 2014 tiene por objeto regular el derecho de acceso a la información pública y les aplica a entidades estatales. El principio que inspira esta ley es que toda la información en poder, control o custodia de una entidad estatal es pública y no puede ser reservada o limitada, salvo por disposición constitucional o legal.

En virtud del derecho de acceso a los documentos públicos, las entidades públicas solo podrán negar el acceso en el caso de ciertas excepciones expresamente previstas en la Ley 1712 de 2014. Esto significa que los datos generados, obtenidos, adquiridos o controlados por las entidades públicas se presumirán públicos, salvo disposición en contrario.

La Ley 1712 de 2014 establece que existen dos excepciones al principio de transparencia: (1) la información pública clasificada, entendida como aquella que, estando en poder o custodia de una entidad pública, pertenece a la esfera privada o semiprivada de una persona natural o jurídica; y (2) la información pública reservada, entendida como aquella que está en poder o custodia de una entidad pública, pero que está exenta de acceso por daño a los intereses públicos. Las entidades públicas deben mantener actualizado un índice de actos, documentos e información establecida como clasificada o reservada. La Ley 1712 de 2014 establece la obligación de las entidades públicas de elaborar un índice de información clasificada y reservada, en el cual deben identificar qué información debe ser clasificada o restringida para proteger datos personales o por razones de defensa o seguridad nacional. Este índice debe ser publicado en la página web oficial de la entidad, así como en el portal de datos abiertos www.datos.gov.co.

En este orden de ideas, el uso de tecnologías de *web scraping* en datos generados, obtenidos, adquiridos o controlados por entidades públicas, es legal en cuanto no se trate de información pública clasificada o información pública reservada. La normativa actual incentiva la innovación en el mercado a partir de la explotación de los datos a través de instrumentos como el CONPES 3920 de 2018, documento que señala la política nacional de explotación de datos (*big data*).

V. CONCLUSIÓN

En el contexto del desarrollo de modelos de inteligencia artificial generativa, el *web scraping* resulta ser una de las técnicas más usadas y de mayor utilidad para la extracción de datos de Internet. Dada la relevancia y el rápido avance de las tecnologías relacionadas con inteligencia artificial, es necesario considerar sus implicaciones legales.

En Colombia, el *web scraping* no es un asunto legislado ni estudiado con amplitud por la jurisprudencia. Aun cuando no existe un pronunciamiento en

relación con el *web scraping* en el ordenamiento jurídico colombiano, esta práctica no está expresamente prohibida y le son aplicables normas que regulan los contratos, la propiedad intelectual y los datos personales.

Las implicaciones de esta técnica dependen en gran medida de las disposiciones contractuales que regulen el acceso a la información de las páginas web y los términos de servicio. En el caso de páginas a las que se puede acceder sin registrarse (*browse-wrap*), dada la ausencia de vinculación contractual entre las partes, el *web scraping* es válido y no contraría el ordenamiento jurídico. En contraste, en contratos *click-wrap* se debe analizar caso a caso el contenido de los términos de servicio para determinar si se permite el *web scraping* o no.

Este método de recolección de datos puede implicar el riesgo de incurrir en una vulneración a derechos de autor cuando la extracción se presenta respecto de obras protegidas. Por lo tanto, es crucial el análisis del sistema de protección aplicable a la información en cada caso. En efecto, cuando los datos pertenecen al dominio público (p. ej., *open data*, información en cabeza del Estado, están dispuestos en Internet mediante licencias que permiten su libre utilización, etc.) o simplemente no reúnen los requisitos para ser protegidos como obra, pueden ser recolectados por medio de *web scraping* de manera legal.

En la medida en que las páginas web pueden contener datos personales, cuando se emplean técnicas de *web scraping* se deben respetar los derechos que consagra el régimen general de protección de datos personales en Colombia. Como resultado, cuando se trate de datos privados o semiprivados se requiere la autorización previa, expresa e informada del respectivo titular, salvo contadas excepciones. Además, el tratamiento debe llevarse a cabo de la forma y para la finalidad que fue informada al titular, pues de lo contrario la SIC podría imponer sanciones. En cuanto a datos públicos no se presenta esta restricción, pero se debe ser cuidadoso a la hora de determinar que constituye un dato público, pues como lo recalcó la SIC, no todo dato de acceso público tiene naturaleza pública.

El ordenamiento penal colombiano también prevé una serie de conductas tipificadas que buscan proteger bienes jurídicos como la información, los datos, la intimidad, la libertad de la información, la honra, la integridad de los documentos y el patrimonio económico. Las conductas desplegadas con ocasión del *web scraping* podrían llegar a ser penalmente relevantes si dolosamente resultan en el acceso a un sistema informático fuera de la autorización de su administrador, se recolectan datos personales sin observancia del régimen general de protección de datos personales o se genera un daño al sistema informático.

Por otra parte, existen una serie de licencias generadas en el marco de movimiento de acceso abierto, a partir de las cuales se podría concluir la existencia de una autorización de uso de la información en el marco del *web scraping*. Ciertamente, mediante sistemas como el *open access*, *open source* o CC, los autores pueden conferir algunas facultades sobre su información a terceros. Aun así, sería equivocado concluir que en todos los casos estos modelos implican una autorización

sin restricción. Por el contrario, el *web scraping* en estos supuestos debe efectuarse conforme los términos fijados por el titular de las obras para ser conforme a la ley.

BIBLIOGRAFÍA

Andersen v. Stability AI Ltd, U.S. District Court for the Northern District of California, n.° 3:23-cv-00201.

Associated Press. "AP, OpenAI agree to share select news content and technology in new collaboration". Comunicado de prensa. 2023.

Bale, Ajay, Naveen Ghorpade, S. S. Rohith Kamalesh, R. Rohith y S. Rohan. "Web Scraping Approaches and their Performance on Modern Websites". Coimbatore, India: Proceedings of the Third International Conference on Electronics and Sustainable Communication Systems, 2022.

Brittain, Blake. "Lawsuits accuse AI content creators of misusing copyrighted work". Reuters, 2023.

Brittain, Blake. "OpenAI hit with new lawsuits from news outlets over AI training". 2024.

Centro Nacional de Desarrollo Curricular en Sistemas no Propietarios. "Las licencias Creative Commons: qué son, por qué utilizarlas y cómo hacerlo". 2021.

Concord Music Group Inc v. Anthropic PBC, U.S. District Court for the Middle District of Tennessee, n.° 3:23-cv-01092.

Corte Constitucional. Sentencia SU139 de 2021, 14 de mayo de 2021, M.P. Jorge Enrique Ibáñez Najar.

Corte Constitucional. Sentencia T-020 de 2014, 27 de enero de 2014. M.P. Luis Guillermo Guerrero Pérez.

Creative Commons. "CC0".

Creative Commons. "Who we are".

David, Emilia. "Microsoft invokes VCRs in motion to dismiss the New York Times' AI lawsuit". 2024.

De Frutos, Rahn. "Excepciones y limitaciones al derecho de autor en Colombia: propuestas legislativas". 2014.

Dirección Nacional de Derechos de Autor. Resolución 11 de 2017.

Doe et al v. GitHub, Inc. et al, U.S. District Court for the Northern District of California, n.° 4:22-cv-06823 (N.D. Cal. Nov 03, 2022).

Hacker, Philipp. "A legal framework for AI training data—from firs tprinciples to the Artificial Intelligence Act". 2020.

hiQ Labs, Inc. v. LinkedIn Corporation, U.S. Court of Appeals for the Ninth Circuit.

Kroto, Vlad, Leigh Redd y Leiser Silva. "Tutorial: Legality and Ethics of Web Scraping". Communications of the Association for Information Systems, 2020.

Lofti, Chaimaa, Swetha Srinivasan, Myriam Ertz y Imen Latrous. *Web scraping techniques and applications: A literature review*. SCRS Conference Proceedings on Intelligent Systems, pp. 381-394.

Margini, Thomas y Diane Peters. "Creative Commons Licenses: Empowering Open Access". 2016.

Meta Platforms, Inc. v. Bright Data Ltd., U.S. District Court for the Northern District of California, n.° 3:23-cv-00077-EMC.

Metke, Ricardo. *Lecciones de propiedad industrial* (III). Baker & McKenzie, 2006.

Nazemian et al v Nvidia Corp, U.S. District Court, Northern District of California, n.° 24-01454.

New York Times Co v Microsoft Corp et al, U.S. District Court for the Southern District of New York, n.° 23-11195.

OpenAI. "GPT-4 Technical Report". Computation and Language (cs.CL). Nueva York: Cornell University, 2023. DOI: arXiv:2303.08774 [cs.CL].

OpenAI. "GPTBot". 2023.

Opendatasoft. "Open Data".

Opendatasoft. "What is open data - Practical Guide".

Open Knowledge foundation. "What is Open Data?".

Originality.ai. "Websites that have blocked OpenAI's GPTBot CCBot Anthropic Google Extended - 1000 Website Study". 2024.

Patern, Tatjana y Layna Deneen. "AI Threats Emerge in Music Publishers' Battle with Big Tech". 2024.

Perez, Sarah. "Court rules in favor of a web scraper, Bright Data, which Meta had used and then sued". TechCrunch. 2024.

Perez, Sarah. "Meta drops lawsuit against web-scraping firm Bright Data that sold millions of Instagram records". TechCrunch. 2024.

Rajko M. Terzić y N. Majstorović Milosav. "Open Data Concept, Its Application and Experiences" Belgrado, Serbia: Vojnotehnički Glasnik / Military Technical Courier, 2019.

Raw Story Media v. OpenAI Inc, U.S. District Court for the Southern District of New York, No. 1:24-cv-01514

Ríos Ruiz, W. R. "Aspectos legales del software libre o de código abierto (open source)". *Revista la Propiedad Inmaterial*, (2003), 41-60.

Rubin, Aaron y Jackie Li. "Court discovers rare and elusive 'enforceable browsewrap'". JDSupra. 2020.

Sala de Casación Penal. Corte Suprema de Justicia. Sentencia SP592-2022, 2 de marzo de 2022 M.P. Diego Eugenio Corredor Beltrán.

Sala de Casación Penal. Corte Suprema de Justicia. Sentencia SP2699-2023, 3 de agosto de 2023 M.P. Fernando León Bolaños Palacios.

Sanabria, Johan. *Sector privado y libre competencia: implicaciones jurídicas del web*. Bogotá: Universidad Externado de Colombia, 2021.

Schmitz, Christian. "Propiedad intelectual, dominio público y equilibrio de intereses". *Revista Chilena de Derecho* 36, n.° 2 (2009).

Sellars, Andrew. *Twenty Years of Web Scraping and the Computer Fraud and Abuse Act*. Boston: Boston Univeristy School of Law, Scholarly Commons at Boston University School of Law.

Shutterstock. "Shutterstock expands partnership with OpenAI, signis new six-year agreement to provide high-quality training data". Comunicado de prensa. 2023.

Shutterstock. "Shutterstock partners with OpenAI and leads the way to bring AI-Generated content to all". Comunicado de prensa. 2022.

St. Laurent, Andrew. "Understanding Open Source and Free Software Licensing". Sebastopol, USA: O'Reilly Media, Inc. 2004.

Stempel, Jonathan. "Nvidia is sued by authors over AI use of copyrighted works". 2024.

Suárez, Alberto. Delitos informáticos. "Lecciones de derecho penal: parte especial". Bogotá: Universidad Externado de Colombia, 2014.

Superintendencia de Industria y Comercio. "Protección de datos personales: aspectos prácticos sobre el derecho de hábeas data".

Superintendencia de Industria y Comercio. Resolución 58834 de 2023.

Telus International. "The essential guide to AI training data".

The Intercept Media Inc v. OpenAI Inc, U.S. District Court for the Southern District of New York, n.° 1:24-cv-01515.

Tong, Anna, Echo Wang y Martin Coulter. "Exclusive: Reddit in AI content licensing deal with Google". Reuters. 2024.

UNESCO. "Concepts of openness and open access". 2015.

US. Copyright Office. "U.S. Copyright Office Fair Use Index". 2023.

Usma, Fidel. *El consentimiento en los contratos en línea B2C y su protección bajo la ley colombiana*. Cuadernos de la Maestría en Derecho n.° 5. Bogotá: Universidad Sergio Arboleda.

Vincent, James. "The lawsuit that could rewrite the rules of AI copyright". Artificial Intelligence, The Verge, 2022.

Walsh, Kat. "Understanding Cc Licenses And Generative Ai", 2023.

Whittaker, Zack. "Web scraping is legal, US appeals court reaffirms". TechCrunch, 2022.

Zhao, Bo. "Web Scraping". Encyclopedia of Big Data. DOI: 10.1007/978-3-319-32001-4_483-1