

Daniel Kula

Uniwersytet Wrocławski

Wydział Prawa, Administracji i Ekonomii

ORCID: 0009-0007-4911-0488

Web scraping na potrzeby szkolenia sztucznej inteligencji a sytuacja autora

Streszczenie

W dobie sztucznej inteligencji (AI) i powszechnego web scrapingu dzieła dostępne w internecie są codziennie kopiowane na potrzeby zbierania danych treningowych do szkolenia modeli AI. W USA, skąd wiele przedsiębiorstw dostarcza rozwiązania oparte na AI, obserwujemy wzrost liczby postępowań sądowych związanych z tym problemem. Dostawcy tych rozwiązań bronią się, powołując doktrynę *fair use*. W tożsamych przypadkach w Europie argumentacja najprawdopodobniej skupi się na wyjątkach od praw wyłącznych wprowadzonych przez prawodawcę unijnego w dyrektywie Parlamentu Europejskiego i Rady (UE) 2019/790. Analiza wprowadzonych dyrektywą wyjątków prowadzi jednak do wniosku, że regulacje te w istocie nie pozostają optymalne ani z perspektywy podmiotów uprawnionych, ani scrapujących.

Słowa kluczowe

web scraping, eksploracja tekstów i danych, dyrektywa DSM, dozwolony użytek, sztuczna inteligencja, uczenie maszynowe, głębokie uczenie, wielkie modele językowe, zastrzeżenie korzystania, uprawnienia autora

***Web scraping* – odpowiedź na potrzeby społeczeństwa informacyjnego?**

Niewątpliwie społeczeństwo informacyjne cechuje się wysokim zapotrzebowaniem na dostęp do danych. Wraz z rozwojem technologii informacyjnych wytwarzamy i publikujemy w internecie więcej danych niż kiedykolwiek, a rok do roku tempo wytwarzania i publikacji wzrasta. Według [WorldWideWebSize.com](https://www.worldwide-web-size.com) na dzień 30 kwietnia 2024 r. w ramach sieci WWW (World Wide Web) indeksuje się ponad 5 mld stron internetowych¹.

¹ WorldWideWebSize.com. Daily estimated size of the world wide web, [https://www.worldwide-web-size.com/#:~:text=Web \(The Internet\)](https://www.worldwide-web-size.com/#:~:text=Web%20(The%20Internet)) [dostęp 30.04.2024].

Zettabajty danych dostępne w internecie stanowią obiekt zainteresowań wielu podmiotów trudniących się agregacją i analizą tych ogromnych zestawów informacji w każdym możliwym celu. Od statystyki i badań naukowych, przez internetowe porównywarki cen biletów lotniczych, po producentów rozwiązań opartych na AI – wszyscy chcą mieć dostęp do najświeższych i najdokładniejszych danych partykularnych kategorii. Nierzadko są to dane prawnie chronione – dane osobowe, utwory w rozumieniu prawa autorskiego czy bazy danych będących przedmiotem praw *sui generis*. Współcześnie najpopularniejsze metody „wydobycia” danych z internetu określa się terminem web scrapingu.

Web scraping (z ang. skrobanie sieci) to proces, który na potrzeby niniejszego artykułu w uproszczeniu można zdefiniować jako technikę ekstrakcji danych z internetu i zapisywanie ich w systemie plików lub bazie danych dla późniejszego odtwarzania i analizy². Z uwagi na objętość, różnorodność oraz tempo publikacji nowych treści praktycznie niemożliwe jest ręczne pozyskiwanie danych przez człowieka, zwłaszcza jeśli dane podlegają ciągłej aktualizacji³. Wymusza to zastosowanie odpowiednich rozwiązań technicznych w postaci botów scrapingowych – skryptów, których zadaniem jest eksploracja stron internetowych i ekstrakcja określonej kategorii danych lub całej zawartości stron internetowych.

Rola web scrapingu w funkcjonowaniu wielu działów nowoczesnej gospodarki nie może zostać przeceniona. Jeden z najbardziej dynamicznie rozwijających się w ostatnich latach sektorów jest immanentnie związany z tą metodą ekstrakcji danych z internetu – AI, w szczególności rozwiązania oparte na ML (*Machine Learning* – z ang. uczenie maszynowe) oraz jego dalszych iteracjach – DL (*Deep Learning* – z ang. uczenie głębokie) czy LLM (*Large Language Models* – z ang. wielkie modele językowe). Te zaawansowane algorytmy szkoli się w celu osiągnięcia zamierzonej funkcjonalności przy użyciu danych treningowych. Rozwiązania oparte na ML wymagają odpowiedniej ilości danych do pomyślnego szkolenia.

Pokłosie beztroskiego web scrapingu jest widoczne w licznych postępowaniach sądowych toczących się przeciwko przedsiębiorstwom wprowadzającym na rynek produkty oparte na AI, których głównym elementem pozostaje kwestia naruszenia

² B. Zhao, *Web Scraping*, [w:] L.A. Schintler, C.L. McNeely (red.), *Encyclopedia of Big Data*, Springer International Publishing, 2018, s. 1, https://www.researchgate.net/publication/317177787_Web_Scraping [dostęp 30.04.2024].

³ V. Krotov, L. Silva, *Legality and ethics of web scraping*, [w:] *Twenty-fourth Americas Conference on Information Systems*, New Orleans 2018, s. 2, https://www.researchgate.net/publication/324907302_Legality_and_Ethics_of_Web_Scraping [dostęp 30.04.2024].

praw autorskich do utworów wykorzystywanych do szkolenia modeli⁴. Głównym argumentem pozwanych w tych postępowaniach jest odwołanie do doktryny *fair use*, której najbliższym, ale jednak nieitożsamym odpowiednikiem pozostaje funkcjonujący w prawie kontynentalnym dozwolony użytek⁵.

Niniejszy artykuł ma na celu zbadanie, w jakim zakresie praktyka web scrapingu, obejmująca utwory oraz inne przedmioty objęte ochroną na gruncie prawa własności intelektualnej, może naruszać zastrzeżone prawa podmiotów uprawnionych. Analiza koncentruje się na ocenie, w jaki sposób wyjątki i ograniczenia wprowadzone przez dyrektywę Parlamentu Europejskiego i Rady (UE) 2019/790 z 17.04.2019 r. w sprawie prawa autorskiego i praw pokrewnych na jednolitym rynku cyfrowym oraz zmiany dyrektyw 96/9/WE oraz 2001/29/WE⁶ wpływają na zdolność podmiotów udostępniających swoje utwory lub inne przedmioty objęte ochroną w internecie do skutecznego zastrzeżenia swoich praw wobec praktyk web scrapingu. Artykuł dąży do ukazania, w jakim stopniu istniejące regulacje prawne uniemożliwiają podmiotom uprawnionym ochronę swoich praw przed nieautoryzowanym pozyskiwaniem i wykorzystywaniem utworów i innych przedmiotów objętych ochroną, w szczególności wprowadzenie zastrzeżenia za pomocą środków nadających się do odczytu maszynowego.

Dozwolony użytek w prawie polskim a *web scraping*

Istota dozwolonego użytku sprowadza się do możliwości korzystania z chronionego utworu bez konieczności uzyskania odrębnej zgody uprawnionego. Eksploatacja utworu w zakresie dozwolonego użytku wyłącza bowiem możliwość dochodzenia przez uprawnionego roszczeń z tytułu naruszenia praw autorskich⁷. W przypadku polskich uregulowań ustawy z dnia 4 lutego 1994 r. o prawie autorskim i prawach pokrewnych⁸ dozwolony użytek został uregulowany w Rozdziale 3 Oddziale 3. Jest on co do zasady nieodpłatną formą korzystania z utworów przez użytkowników. W niektórych przypadkach twórcy przysługuje prawo do wynagrodzenia

⁴ Zob. pozew Getty Images Inc. v. Stability AI inc., https://fingfx.thomsonreuters.com/gfx/legal-docs/byvrlkmwnve/GETTY_IMAGES_AI_LAWSUIT_complaint.pdf [dostęp 30.04.2024].

⁵ K. Grzybczyk, *Fair use jako przykład ograniczenia monopolu autorskiego klauzulą generalną*, „Zeszyty Naukowe Uniwersytetu Jagiellońskiego” 2020, nr 1, s. 24.

⁶ Dz. Urz. UE L 160/62 z 17.05.2024 r.; dalej: „dyrektywa DSM”

⁷ W. Machała, *Dozwolony użytek chronionych utworów w polskim prawie autorskim w świetle cywilistycznej koncepcji prawa*, „Zeszyty Naukowe Uniwersytetu Jagiellońskiego” 2001, nr 78, s. 59.

⁸ Dz. U. z 1994 r. Nr 24, poz. 83 ze zm.; dalej: „pr. aut.”

(np. rozpowszechnianie aktualnych wypowiedzi i fotografii reporterskich w prasie, radiu i telewizji – art. 25 ust. 2 w zw. z ust. 1 pkt 1 lit. b pr. aut.). W doktrynie różni się dozwolony użytek osobisty (prywatny) oraz publiczny.

Postaci dozwolonego użytku zostały uregulowane w modelu kazuistycznym. Zgodnie z dominującym poglądem doktryny enumeratywnie wskazanie postaci dozwolonego użytku wyłącza stosowanie wykładni rozszerzającej, zwłaszcza przyjmowanie nowych postaci dozwolonego użytku *per analogiam*⁹. Co więcej, korzystanie z utworów w granicach dozwolonego użytku powinno odbywać się pod warunkiem wymienienia imienia i nazwiska twórcy oraz źródła przy uwzględnieniu istniejących możliwości¹⁰ (art. 34 pr. aut.), a jednocześnie taki dozwolony użytek nie może naruszać normalnego korzystania z utworu lub godzić w słusne interesy twórcy (art. 35 pr. aut.). Dla korzystania z utworu w ramach dozwolonego użytku wymagane jest więc spełnienie łącznie tych trzech warunków.

Przenosząc poczynione rozważania na grunt web scrapingu, należy zauważyć w ślad za doktryną, że „(e)tap polegający na skopiowaniu danych, o ile mamy do czynienia z utworem chronionym prawem autorskim i nie możemy oprzeć się na ustawowym wyjątku, zawsze wkracza w wyłączne prawo do zwielokrotniania, gwarantowane zarówno traktatami międzynarodowymi, jak i ustawami krajowymi”¹¹. Przed uchwaleniem dyrektywy DSM słusznie wskazywało się również, że nie istniały podstawy do zakwalifikowaniem zbierania danych dla potrzeb wygenerowania danych treningowych poza zakresem monopolu autorskiego, nawet jeśli dane te były pobierane i przechowywane przez krótki czas, a co więcej – sam proces przetworzenia (wektoryzacji) do postaci czytelnej dla sieci neuronowej stanowi zwielokrotnianie¹². Ten stan prawny wykluczał dopuszczalność takiego zwielokrotniania bez odpowiedniej zgody uprawnionego podmiotu.

Przyjmując, że nowa postać dozwolonego użytku zostałaby wprowadzona do porządku prawnego przez ustawodawcę (przykładowo w postaci implementacji wyjątków zawartych w dyrektywie DSM jako nowych postaci dozwolonego użyt-

⁹ J. Barta, R. Markiewicz, *Prawo autorskie i prawa pokrewne*, Warszawa 2021, s. 81.

¹⁰ Uwzględnienie istniejących możliwości powinno jednak być interpretowane restrykcyjnie, z uwagi na treść art. 5 ust. 3 lit. a) dyrektywy 2001/29/WE Parlamentu Europejskiego i Rady z dnia 22 maja 2001 r. w sprawie harmonizacji niektórych aspektów praw autorskich i pokrewnych w społeczeństwie informacyjnym (Dz. Urz. UE L 167/44 z 22.06.2001 r.), gdzie kształt tej przesłanki został określony zwrotem „poza przypadkami, w których okaże się to niemożliwe” – zob. *ibidem*, s. 84.

¹¹ E. Traple, [w:] B. Fischer, A. Pązik, M. Świerczyński (red.), *Prawo sztucznej inteligencji i nowych technologii*, Warszawa 2021, s. 22.

¹² Z. Okoń, [w:] K. Flaga-Gieruszyńska, J. Gołaczyński, D. Szostek (red.), *Sztuczna inteligencja, blockchain, cyberbezpieczeństwo oraz dane osobowe. Zagadnienia wybrane*, Warszawa 2019, s. 109–110.

ku¹³), nadal pozostawałaby konieczność kumulatywnego wypełnienia pozostałych warunków ustawowych w postaci wymienienia imienia i nazwiska twórcy oraz źródła, a ponadto niegodzenia w słusne interesy twórcy i normalne korzystanie przez niego z utworu. Możliwość spełnienia obu z tych warunków należy podać w wątpliwość. Mówiąc o warunku niegodzenia w słusne interesy twórców, jest to wątpliwe tym bardziej w przypadku web scrapingu dokonywanego na potrzeby szkolenia narzędzi generatywnej AI, które wraz ze wzrostem popularności stanowią realne zagrożenie dla interesów uprawnionych. W literaturze problem nie pozostaje obojętny. Postuluje się nawet jego potencjalne rozwiązania, np. w postaci ustanowienia systemu wynagradzania podmiotów uprawnionych, których utwory były wykorzystywane do szkolenia generatywnej AI, a jej wytwory zastępują w obrocie te, które zostały stworzone przez ludzkich twórców¹⁴.

De lege lata nie ma możliwości wykonywania web scrapingu w ramach dozwolonego użytku w rozumieniu pr. aut., ponieważ nie wchodzi on w zakres żadnej z ustawowo uregulowanych jego form. W obliczu obiektywnej potrzeby zapewnienia możliwości rozwoju technologii wymagających web scrapingu (w tym rozwoju AI) prawodawca unijny zawarł w ramach dyrektywy DSM dwa wyjątki od funkcjonujących w unijnym systemie prawnym praw wyłącznych (art. 3 i 4 dyrektywy DSM), w odniesieniu do utworów, które są dostępne zgodnie z prawem na potrzeby eksploracji tekstów i danych (dalej: „TDM”).

Web scraping jako zwielokrotnianie utworów na potrzeby TDM

Idąc dalej, należy zadać sobie pytanie, czy *web scraping* dokonywany na potrzeby uzyskania danych treningowych do szkolenia AI może być traktowany jako przejaw „zwielokrotniania i pobrań do celów eksploracji tekstów i danych” zgodnie z zawartą w art. 2 pkt 2) dyrektywy DSM definicją TDM. W rozumieniu dyrektywy mianem eksploracji tekstów i danych określa się „zautomatyzowaną technikę analityczną służącą do analizowania tekstów i danych w postaci cyfrowej w celu wygenerowania informacji, obejmujących między innymi wzorce, tendencje i korelacje”. Szkolenie modeli AI zawiera w sobie element TDM, a faktem jest, że prawodawca unijny jednoznacznie

¹³ Najnowsza dostępna wersja projektu nowelizacji pr. aut. zakłada implementację tych wyjątków właśnie w ramach sekcji poświęconej dozwolonemu użytkowi, w projektowanych art. 26² oraz 26³, <https://legislacja.rcl.gov.pl/docs//2/12382002/13037430/dokument664634.zip> [dostęp 30.04.2024].

¹⁴ Zob. M. Senfleben, *Generative AI and Author Remuneration*, „International Review of Intellectual Property and Law” 2023, nr 54, s. 1535–1560, *passim*, <https://doi.org/10.1007/s40319-023-01399-4>.

uznaje, że *web scraping* dokonywany na potrzeby szkolenia AI mieści się w zakresie tej definicji. Ewidentny tego przykład stanowi wskazanie na potrzebę zachowania zgodności tych działań z dyrektywą DSM (w szczególności w zakresie zastrzeżenia, o którym mowa w art. 4 ust. 3 dyrektywy DSM) w motywach¹⁵ rozporządzenia Parlamentu Europejskiego i Rady (UE) 2024/1689 z dnia 13 czerwca 2024 r. w sprawie ustanowienia zharmonizowanych przepisów dotyczących sztucznej inteligencji oraz zmiany rozporządzeń (WE) nr 300/2008, (UE) nr 167/2013, (UE) nr 168/2013, (UE) 2018/858, (UE) 2018/1139 i (UE) 2019/2144 oraz dyrektyw 2014/90/UE, (UE) 2016/797 i (UE) 2020/1828 (akt w sprawie sztucznej inteligencji)¹⁶.

Wspólną cechą obu wyjątków (tj. wyjątków z art. 3 i 4 dyrektywy DSM) jest istnienie warunku „zgodnego z prawem dostępu” do utworów. Prawodawca unijny wyjaśnia w motywie 14 dyrektywy DSM, odnoszącym się do wyjątku zawartego w art. 3, rozumienie dostępu zgodnego z prawem jako „dostępu do treści opartego na polityce otwartego dostępu lub na ustaleniach umownych pomiędzy podmiotami uprawnionymi i organizacjami badawczymi lub instytucjami dziedzictwa kulturowego, takich jak abonamenty lub na innych zgodnych z prawem środkach. [...] Zgodny z prawem dostęp powinien również obejmować dostęp do treści, które są swobodnie dostępne w internecie”. Natomiast w przypadku art. 4 dyrektywy DSM zgodnie z motywem 18 dyrektywy DSM w szczególności chodzi o sytuacje, gdy utwór lub inny przedmiot objęty ochroną „został podany do publicznej wiadomości w internecie”, o ile podmioty uprawnione nie zastrzegły w odpowiedni sposób praw do zwielokrotniania i pobierania do celów TDM. Dokonanie tej dystynkcji w motywach dyrektywy prowadzi do wniosków, że w przypadku wyjątku z art. 4 dyrektywy DSM samo nabycie abonamentu na korzystanie z utworów nie będzie warunkowało samo w sobie wystąpienia przesłanki „zgodnego z prawem dostępu” na potrzeby TDM. Przy tym warto zaznaczyć, że w przypadku wyjątku zawartego w art. 4 dyrektywy DSM prawodawca unijny nie wyłączył możliwości zawierania umów sprzecznych z jego postanowieniami tak, jak uczynił to w przypadku art. 3, 5 i 6. Nic nie stoi więc na przeszkodzie, aby uprawnieni na mocy umowy ustanowili wynagrodzenie za korzystanie z utworów i innych przedmiotów objętych ochroną, do których przysługują im prawa wyłączne¹⁷.

¹⁵ W szczególności należy wskazać tu na treść motywu 105, który w ostatnim zdaniu jednoznacznie wskazuje: „W przypadku gdy prawo do wyłączenia z eksploracji zostało w odpowiedni sposób wyraźnie zastrzeżone, dostawcy modeli AI ogólnego przeznaczenia muszą uzyskać zezwolenie od podmiotów uprawnionych, jeżeli chcą dokonywać eksploracji tekstów i danych odnośnie do takich utworów”.

¹⁶ Dz. Urz. UE L z 12.07.2024 r.

¹⁷ R. Markiewicz, *Prawo autorskie na jednolitym rynku cyfrowym. Dyrektywa Parlamentu Europejskiego i Rady (UE) 2019/790*, Warszawa 2021, s. 64.

Wyjątek z art. 3 dyrektywy DSM

Pierwszy wyjątek umożliwiający skorzystanie z dozwolonego użytku na potrzeby web scrapingu uregulowany jest w art. 3 ust. 1 dyrektywy DSM i odnosi się do zwielokrotnień i pobrań dokonywanych przez organizacje badawcze i instytucje dziedzictwa kulturowego „w celu przeprowadzenia, do celów badań naukowych, eksploracji tekstów i danych na utworach lub innych przedmiotach objętych ochroną, do których mają zgodny z prawem dostęp”. Na potrzeby TDM podmioty te mogą skorzystać z wyjątku od praw wyłącznych skonkretyzowanych w: a) art. 5 lit. a oraz art. 7 ust. 1 dyrektywy 96/9/WE Parlamentu Europejskiego i Rady z dnia 11 marca 1996 r. w sprawie ochrony prawnej baz danych¹⁸ i b) art. 2 dyrektywy 2001/29/WE Parlamentu Europejskiego i Rady z dnia 22 maja 2001 r. w sprawie harmonizacji niektórych aspektów praw autorskich i pokrewnych w społeczeństwie informacyjnym¹⁹.

W ocenie autora wprowadzenie wyjątku na rzecz organizacji badawczych i instytucji dziedzictwa kulturowego nie budzi większych kontrowersji oraz pozostaje w pełni uzasadnione z uwagi na potrzebę rozwoju technologii i nauki właśnie przez te podmioty. Z tego powodu rozważania na temat wyjątku wprowadzonego przez art. 3 dyrektywy DSM należy ograniczyć do krótkiej charakterystyki przedstawionej powyżej, a dalsza część niniejszego artykułu zostanie poświęcona wyjątkowi przewidzianemu w art. 4.

Wyjątek z art. 4 dyrektywy DSM

Wprowadzony w art. 4 dyrektywy DSM wyjątek, który nie posiada tak skonkretyzowanego zakresu podmiotowego jak w art. 3, przy jednoczesnym rozszerzeniu zakresu przedmiotowego (o prawa wyłączne zawarte w art. 4 ust. 1 lit. a) i b) dyrektywy 2009/24/WE z dnia 23 kwietnia 2009 r. w sprawie ochrony prawnej programów komputerowych²⁰), budzi zdecydowanie większe kontrowersje. W art. 4 prawodawca unijny nie zdefiniował bowiem adresata normy, którą państwa członkowskie mają zaimplementować, wskazując jedynie na „znaczenie technik eksploracji tekstów i danych dla podmiotów prywatnych i publicznych do analizowania dużych ilości danych w różnych sferach życia codziennego i do różnych celów, w tym przez służby państwowe, do podejmowania złożonych decyzji biznesowych oraz do rozwijania nowych aplikacji i technologii” (motyw 18 dyrektywy DSM).

¹⁸ Dz. Urz. UE L 77/39 z 27.03.1996 r.

¹⁹ Dz. Urz. UE L 167/44 z 22.06.2001 r.

²⁰ Dz. Urz. UE L 111/52 z 5.05.2009 r.

Analizując dalej motyw 18, potrzeba takiego uregulowania wynika również z konieczności zniwelowania niepewności co do możliwości zastosowania art. 5 ust. 1 dyrektywy 2001/29/WE w sprawie harmonizacji niektórych aspektów praw autorskich i pokrewnych w społeczeństwie informacyjnym do procesu TDM (tak jak w przypadku web scrapingu, zakładającego potrzebę trwałego zwielokrotnienia danych, w celu ich następczej obróbki dla szkolenia modeli AI). Niewątpliwie prawodawca unijny miał na celu stworzenie przyjaznego środowiska dla podmiotów prywatnych do rozwijania technologii AI na terenie UE, rozpoznając wyzwania natury prawnej związane z pozyskiwaniem danych treningowych.

Temporalna granica przechowywania zwielokrotnionych utworów i innych pobranych przedmiotów objętych ochroną została zakreślona w ust. 2 omawianego artykułu i obejmuje przechowywanie „[...] tak długo, jak jest to konieczne do celów eksploracji tekstów i danych”. Dane powinny więc zostać usunięte w momencie, w którym zostały już wygenerowane informacje obejmujące wzorce, tendencje i korelacje, o których mowa w art. 2 ust. 2 dyrektywy DSM. Takie rozwiązanie z jednej strony pozwala na ograniczenie skutków sanowanego naruszenia praw wyłącznych poprzez minimalizację czasu, w którym będą one przechowywane przez podmiot dokonujący TDM, z drugiej jednak może stanowić istotne ograniczenie możliwości następczego zbadania plików na potrzeby przeprowadzenia dowodu w przypadku powstania sporu pomiędzy uprawnionym z tytułu praw autorskich a dokonującym TDM.

Prawodawca unijny w art. 4 ust. 3 dyrektywy DSM wprowadza warunek, który musi zostać spełniony dla skorzystania z wyjątku zawartego w art. 4 – brak wyraźnego zastrzeżenia korzystania z utworów i innych przedmiotów objętych ochroną przez podmiot uprawniony. Zastrzeżenie to musi być wyrażone „[...] w odpowiedni sposób, na przykład za pomocą środków nadających się do odczytu maszynowego w przypadku treści, które zostały podane do publicznej wiadomości w internecie”. Prawodawca unijny wprowadza więc domniemanie zgody uprawnionego na korzystanie z utworów i innych przedmiotów objętych ochroną do celów TDM udostępnionych w internecie. Możliwość zastrzeżenia przez podmioty uprawnione korzystania z utworów i innych przedmiotów objętych ochroną jest – zgodnie z dominującymi poglądami – niedostateczna dla realizowania uzasadnionych potrzeb korzystających z TDM oraz stawia te podmioty z krajów UE w gorszej pozycji w stosunku do tych, które operują chociażby w jurysdykcji USA, Japonii czy Chin²¹.

²¹ Tak m.in. E. Traple, *op. cit.*, s. 26 oraz R. Markiewicz, *op. cit.*, s. 64.

W opinii autora domyślna zgoda uprawnionego wprowadzona w art. 4 ust. 3 dyrektywy DSM pozostaje problematyczna. Jest tak z uwagi na fakt, że podmioty uprawnione nie zawsze udostępniają swoje utwory i inne przedmioty objęte ochroną w sieci samodzielnie, a wydaje się, że częstszym zjawiskiem jest udostępnianie utworów za pośrednictwem zewnętrznych podmiotów takich jak dostawcy platform internetowych, którym uprawnieni udzielają licencji na rozpowszechnienie utworu w sieci lub przenoszą na nich autorskie prawa majątkowe do utworu. To właśnie te zewnętrzne podmioty mają faktyczną (techniczną) możliwość zastrzeżenia korzystania z utworów w odpowiedni sposób, w tym z użyciem środków nadających się do odczytu maszynowego, o których mowa w ust. 3 omawianego artykułu. Z tego względu przed analizą skutków, jakie art. 4 dyrektywy DSM może wyrzucić w sferze interesów podmiotów uprawnionych, należy przeanalizować, w jaki sposób utwory są udostępniane w sieci i jak te przypadki rzutują na możliwość wprowadzenia zastrzeżenia.

Podmiot uprawniony udostępnia samodzielnie

Przypadki, gdy utwór lub inny przedmiot objęty ochroną jest udostępniony bezpośrednio przez podmiot uprawniony (w witrynie operowanej osobiście przez niego) lub inny podmiot²² (w taki sposób, że podmiot uprawniony ustala warunki korzystania przez użytkowników), pozostają raczej rzadkością w przypadku niezależnych twórców, jednak zdarzają się²³. Często można natomiast zauważyć udostępnianie w sieci utworów przez wydawców, którym przysługują autorskie prawa majątkowe – prowadzących działalność gospodarczą opartą na publikacji treści w internecie. W takim układzie korzystanie z utworów przez użytkowników serwisu może odbywać się zarówno nieodpłatnie, jak i za wynagrodzeniem,

²² Z uwagi na niniejszy śródtytuł celowo pominięte zostały rozważania dotyczące dostawców usług pośrednich innych niż platformy internetowe, w rozumieniu rozporządzenia Parlamentu Europejskiego i Rady (UE) 2022/2065 z dnia 19 października 2022 r. w sprawie jednolitego rynku usług cyfrowych oraz zmiany dyrektywy 2000/31/WE (Dz. Urz. UE L 277/65 z 27.10.2022 r.), (dalej: „DSA”), ponieważ rola tych podmiotów pozostaje marginalna z perspektywy ustalania warunków korzystania z utworów udostępnianych w witrynie operowanej bezpośrednio przez podmiot uprawniony.

²³ Np. polski artysta Taco Hemingway udostępnił swoją twórczość (albumy muzyczne) do pobrania w formie skompresowanych folderów w formacie .zip na stronie <https://pobierz.1-800-oswiecenie.com/> [dostęp 30.04.2024], a adres strony internetowej zmienia się wraz z wydaniem każdego kolejnego albumu. Na stronie nie sposób szukać jednak jakichkolwiek warunków umowy licencyjnej, której autor udziela swoim słuchaczom, należy więc przyjąć, że mogą z nich korzystać w ramach dozwolonego użytku osobistego.

często w postaci subskrypcji lub jednorazowej opłaty za dostęp do utworu lub ich grupy, a warunki i pola eksploatacji, na jakich użytkownicy mogą korzystać z utworu w ramach licencji²⁴ będą ustalane bezpośrednio przez odpowiedni podmiot uprawniony.

Przenosząc te rozważania na grunt web scrapingu dokonywanego w ramach art. 4 dyrektywy DSM, należy przyjąć, że konsekwencje dla legalności TDM w zależności od warunków korzystania z utworu narzuconych przez podmiot uprawniony mogą przyjmować zróżnicowane postaci. Przede wszystkim warunki korzystania z utworu narzucone przez podmiot uprawniony mogą całkowicie wyłączać możliwość zwielokrotnienia utworu na potrzeby TDM. Podmiot uprawniony posiada więc faktyczną możliwość zastrzeżenia korzystania, w tym za pomocą środków możliwych do odczytu maszynowego. W świetle treści art. 7 ust. 1 dyrektywy DSM nie stoi również na przeszkodzie, aby podmiot uprawniony, udostępniając utwór lub inny przedmiot objęty ochroną, zastrzegł możliwość korzystania z utworu na potrzeby TDM, jednocześnie uzależniając zgodę na takie korzystanie od zapłaty wynagrodzenia – np. w postaci zwiększonej opłaty za dostęp czy indywidualnie negocjowanych warunków²⁵.

Podmiot trzeci

Powszechnym zjawiskiem jest udostępnianie utworów i innych przedmiotów objętych ochroną za pośrednictwem podmiotów trzecich, w szczególności platform internetowych w rozumieniu DSA czy dostawców usług udostępniania treści online w rozumieniu dyrektywy DSM. W tym zakresie na potrzeby udostępnienia utworu przez podmiot trzeci, autor lub inny podmiot uprawniony udziela administratorowi witryny licencji, często również w zakresie udzielania sublicencji (w dyspozycji art. 68 ust. 3 pr. aut.) lub przenosi na niego autorskie prawa majątkowe do utworu.

Przypadek przeniesienia autorskich praw majątkowych wydaje się zdecydowanie łatwiejszy do zakwalifikowania w zakresie ewentualnego zastrzeżenia korzy-

²⁴ Mając na względzie sposób zawarcia umowy licencyjnej w tym modelu, tj. w sposób dorozumiany i z użyciem wzorca umownego, należy wykluczyć możliwość przeniesienia autorskich praw majątkowych do utworu zgodnie z regulacją art. 53 pr. aut. Oczywiście zakres licencji może zostać określony przez podmiot uprawniony w bardzo zróżnicowany sposób, w tym np. z użyciem powszechnych licencji typu Creative Commons.

²⁵ Zob. np. GETTY IMAGES CONTENT LICENSE AGREEMENT, Restricted Uses, lit. k, <https://www.gettyimages.com/eula> [dostęp 21.05.2024]

stania z utworów na potrzeby TDM. W tym przypadku bowiem to administrator serwisu staje się podmiotem uprawnionym w znaczeniu art. 4 i jego pozycja prawna zostaje praktycznie zrównana z tą podmiotu uprawnionego opisywaną w poprzedniej jednostce redakcyjnej²⁶.

W pierwszym przypadku główną rolę będzie grała treść umowy zawieranej przez podmiot uprawniony z administratorem witryny. To właśnie jej warunki będą w istocie określać dalsze losy utworu lub innego przedmiotu objętego ochroną na potrzeby TDM. W opinii autora w jej treści powinny znaleźć się postanowienia dotyczące ewentualnego wprowadzenia zastrzeżenia do struktury strony internetowej przez podmiot trzeci. Niestety, w większości przypadków jej treść będzie kształtowana przez nienegocjowalne wzorce umów narzucane przez platformy internetowe i innych administratorów witryn w postaci np. regulaminu usługi. Wydaje się, że sens wprowadzenia prawa do zastrzeżenia, o którym mowa w art. 4 ust. 3 dyrektywy DSM, pozostaje wypaczony właśnie z uwagi na tę okoliczność.

Wniosek taki wynika z wykładni literalnej ust. 3 – wskazania *explicite* podmiotu uprawnionego jako zdolnego do poczynienia zastrzeżenia przez unijnego prawodawcę. Natomiast patrząc na uwarunkowania rynkowe, chociażby analizując warunki korzystania jednej z popularniejszych platform internetowych – YouTube – podmiot uprawniony, udostępniając na niej swój utwór lub inny przedmiot objęty ochroną, będzie udzielał „[...] każdemu innemu użytkownikowi Usługi ogólnoświatowej, niewyłącznej i nieodpłatnej licencji na korzystanie z jego Treści w ramach Usługi (w tym na zwielokrotnianie, rozpowszechnianie, modyfikowanie, wyświetlanie i wykonywanie) wyłącznie w zakresie dopuszczonym w ramach Usługi”²⁷. Dopuszczalny zakres, o którym mowa powyżej, jest kształtowany przez wolę samego dostawcy platformy internetowej. Ten musi co prawda uwzględnić powszechnie obowiązujące przepisy prawa, które jednak na ten moment nie wskazują *explicite* obowiązku uwzględnienia woli podmiotu uprawnionego do dokonania zastrzeżenia. Pozostaje to niezrozumiałe w obliczu tej obiektywnej przewagi kontraktowej i faktu, że to podmiot trzeci ma faktyczną (techniczną) możliwość wprowadzenia zastrzeżenia. Takiego obowiązku nie nałożono bowiem na platformy internetowe w DSA ani na dostawców usług udostępniania treści online w dyrektywie DSM.

²⁶ „Praktycznie”, ponieważ twórcy w dalszym ciągu będą przysługiwać niezbywalne autorskie prawa osobiste, o których mowa w art. 16 pr. aut.

²⁷ *Licencja dla innych użytkowników*, <https://www.youtube.com/t/terms>, [dostęp 30.04.2024].

Udostępnienie przez osobę trzecią nieuprawnioną

Równie powszechnym zjawiskiem jest udostępnianie utworów i innych przedmiotów objętych ochroną przez osoby nieuprawnione do takiego udostępnienia. Już w momencie udostępnienia przez ww. podmiot trzeci dochodzi do naruszenia monopolu prawnoautorskiego i podmiot dokonujący takiego udostępnienia jest odpowiedzialny za naruszenie. Słusznie wskazuje się bowiem w literaturze, że „[...] jeżeli osoba na własnej stronie internetowej chce zamieścić utwory osób trzecich, które są objęte ochroną prawnoautorską, musi uzyskać zezwolenie twórcy na korzystanie z utworu w ramach tego pola eksploatacji, aby nie narazić się na zarzut naruszenia praw autorskich, niezależnie od tego, czy robi to w ramach działalności komercyjnej czy niekomercyjnej”²⁸. Każdy przejaw udostępniania utworów i innych przedmiotów podlegających ochronie może być rozpatrywany w kategoriach naruszenia autorskich praw majątkowych, niezależnie od tego, czy takie udostępnienie odbywa się na własnej stronie internetowej udostępniającego, za pośrednictwem dostawców usług pośrednich, czy np. za pośrednictwem programów *peer-to-peer*.

Nasuwa się więc pytanie, czy podmiot dokonujący TDM ze źródeł, które pierwotnie już zostały obciążone naruszeniem monopolu autorskiego, może skutecznie powołać się na wyjątek zawarty w art. 4 dyrektywy DSM. Przywołany już motyw 18 wskazywałby, że utwór lub inny przedmiot objęty ochroną, który „został podany do publicznej wiadomości w internecie”, *prima facie* samoistnie gwarantuje spełnienie przesłanki dostępu zgodnego z prawem. Takie stwierdzenie istotnie mogłoby zostać podniesione przez dokonującego *web scraping*. Patrząc na regulacje dyrektywy DSM, istnieje uzasadnione przypuszczenie, że utwór lub inny przedmiot objęty ochroną podany do publicznej wiadomości w internecie może być przedmiotem zwielokrotnienia w zakresie art. 4, o ile nie zostało dokonane zastrzeżenie. Kolejnym argumentem za wskazaną tezą jest ocena możliwości, jakie posiada taki podmiot w zakresie analizy uwarunkowań prawnoautorskich, które są przedmiotem umowy pomiędzy podmiotami od niego niezależnymi. Konstrukcja ww. argumentów oparta jest jednak na niewiedzy podmiotu dokonującego *web scrapingu* o nielegalności źródła, natomiast wydaje się, że należałoby jednoznacznie oddzielić przypadki, gdzie podmioty świadomie scrapują z nielegalnych źródeł. Na ten moment, z uwagi na stosunkowo krótki okres po uchwaleniu dyrektywy DSM, ciężko udzielić jednoznacznej odpowiedzi – nie wykształciła się jeszcze praktyka orzecznicza. W opinii autora więcej argumentów

²⁸ M. Nowikowska, [w:] J. Sieńczyło-Chlabicz (red.), *Prawo własności intelektualnej. Teoria i praktyka*, Warszawa 2021, s. 374.

przemawia jednak za uznaniem zwielokrotnienia obciążonego pierwotnym naruszeniem monopolu autorskiego za wykraczające poza zakres art. 4 dyrektywy DSM. Biorąc pod uwagę obowiązek zawężającej interpretacji wyjątków oraz dotychczasowe orzecznictwo TSUE w sprawach dotyczących wyjątków od praw wyłącznych, zdecydowanie bardziej uprawnione jest przyjęcie kryterium legalności źródła – tak jak TSUE przyjął, odpowiadając na pytanie prejudycjalne w odniesieniu do dozwolonego użytku osobistego w sprawie *ACI Adam BV i in. v. Stichting de Thuiscope i Stichting Onderhandeligen Thuiscope vergoeding*²⁹. Kwestia wykorzystywania utworów i innych przedmiotów objętych ochroną pochodzących z nielegalnych źródeł dla TDM niewątpliwie pozostaje problematyczna. Praktyka pokazuje bowiem, że podmioty dokonujące web scrapingu na potrzeby szkolenia AI korzystają również z nich³⁰.

Ułomność „możliwości odczytania maszynowego”

Powracając jednak do sytuacji, w których podmiot uprawniony ma możliwość dokonania zastrzeżenia, o którym mowa w art. 4 ust. 3 dyrektywy DSM, warto zwrócić uwagę na nieprecyzyjność prawodawcy unijnego w formułowaniu wymogu co do formy, w jakiej zastrzeżenie może być poczynione. Użycie sformułowania „na przykład za pomocą środków nadających się do odczytu maszynowego” wymaga uzupełnienia o treść motywu 18 dyrektywy DSM. Prawodawca unijny wyjaśnia w nim, że „w przypadku treści, które zostały podane do publicznej wiadomości w internecie za stosowne powinno się wyłącznie uznać zastrzeżenie tych praw za pomocą środków nadających się do odczytu maszynowego, łącznie z metadanymi i warunkami korzystania ze strony internetowej lub usługi”. Prawodawca dopuszcza zatem możliwość dokonania zastrzeżenia w warunkach korzystania ze strony internetowej lub usługi, wyraźnie nie zdając sobie sprawy, że boty scrapingowe wykorzystywane do dokonywania zwielokrotniania i pobierania treści na potrzeby TDM to w istocie proste programy, które zazwyczaj nie dokonują analizy warunków czy regulaminów zamieszczonych w witrynach, których zawartość pobierają.

²⁹ Wyrok Trybunału z dnia 10 kwietnia 2014 r. w sprawie C-435/12, *ACI Adam BV i in. v. Stichting de Thuiscope i Stichting Onderhandeligen Thuiscope vergoeding*, EU:C:2014:254.

³⁰ Np. podczas szkolenia swojego LLM Meta użyła i przyznała się do użycia zestawu danych zawierającego książki z nielegalnego źródła „the Books3 section of ThePile”. Według agencji Reuters osoba, która stworzyła ten zestaw danych, podaje, że sporządziła go z ponad 196 640 książek, K. Paul, *Meta used copyrighted books for AI training despite its own lawyers' warnings, authors allege*, 13.12.2023, <https://www.reuters.com/technology/meta-used-copyrighted-books-ai-training-despite-its-own-lawyers-warnings-authors-2023-12-12/> [dostęp 30.04.2024].

Brak regulacji odnoszącej się bezpośrednio do tego, co to znaczy, że zastrzeżenie „jest możliwe do odczytu maszynowego”, przy jednoczesnym wymienieniu warunków strony internetowej lub usługi jako jeden ze sposobów zastrzeżenia, prowadzi do powstania dysonansu, który w ocenie autora prawodawca unijny powinien uzupełnić o wprowadzenie jednolitego sposobu wprowadzenia zastrzeżenia, a to z kilku powodów.

Po pierwsze, na rynku wewnętrznym warunki korzystania ze stron internetowych występują w znacznej liczbie języków i często w ramach jednej witryny nie są nawet dostępne dodatkowo w języku angielskim, z uwagi na fakt, że taki warunek nie w każdym przypadku³¹ przewidują akty prawa unijnego, natomiast akty prawa krajowego często wymuszają na administratorach jedynie udostępnianie warunków w języku narodowym (np. art. 7 ustawy z dnia 7 października 1999 r. o języku polskim [Dz. U. z 1999 r. Nr 90, poz. 999 ze zm.] w odniesieniu do obrotu konsumenckiego). Prowadzi to do sytuacji, w której podmiot planujący wykorzystać bota scrapingowego na terenie UE musi go zaprogramować w sposób, który umożliwi mu co najmniej samodzielną analizę warunków korzystania we wszystkich językach urzędowych UE, co znacząco zwiększa poziom skomplikowania takiego bota, który z założenia ma być prostym środkiem do pozyskania treści z internetu. Nietrudno wywnioskować również, że z uwagi na takie utrudnienia w zakresie stworzenia bota scrapingowego podmioty będą: a) rezygnować z web scrapingu, pozostając w jurysdykcji UE, lub b) scrapować, nie zwracając uwagi na warunki korzystania z witryn.

Po drugie, wykształcona praktyka pozwala na przyjęcie innego środka w postaci pliku *robots.txt* zamieszczanego w metadanych witryny (co również stara się zauważyć prawodawca unijny w motywie 18, odnosząc się do „metadanych”), w tym zakresie jednak poszczególni scrapujący nie posiadają jednolitych warunków wyłączenia. Przykładowo, OpenAI w odniesieniu do swojego GPTBot’a scrapującego dane na potrzeby działania modelu GPT-4 może zostać wyłączony w odniesieniu do strony, w której w pliku *robots.txt* znajdują się dwie linijki tekstu³²:

User-agent: GPTBot

Disallow: /

³¹ Np. w art. 14 ust. 6 DSA prawodawca unijny wprowadził obowiązek dla bardzo dużych platform internetowych i bardzo dużych wyszukiwarek do publikowania swoich warunków korzystania z usług w językach urzędowych wszystkich państw członkowskich, w których oferują usługi, jednak zakres podmiotowy tego przepisu pozostaje istotnie ograniczony.

³² GPTBot, OpenAI Platform, <https://platform.openai.com/docs/gptbot> [dostęp 30.04.2024].

Nie wszyscy producenci popularnych modeli AI są tak transparentni w kwestii blokowania swoich botów scrapingowych. Poszukując informacji, można natrafić chociażby na artykuły popularnonaukowe, które zawierają zbiory takich komend³³, które muszą zostać zaimplementowane do pliku *robots.txt*, aby ten potencjalnie ograniczył dostęp dla poszczególnych botów. Są one jednak najczęściej wskazywane przez podmioty zewnętrzne, niezwiązane z działalnością poszczególnych przedsiębiorstw zajmujących się web scrapingiem. Analiza tych zbiorów prowadzi do wniosku, że każdy bot wymaga osobnej komendy w pliku *robots.txt*. Wydaje się to kuriozalne w obliczu konieczności zapewnienia respektowania prerogatywy podmiotów uprawnionych do zastrzeżenia korzystania, o której mowa w art. 4 ust. 3 dyrektywy DSM.

Z uwagi na to należy postulować ujednolicenie przez prawodawcę unijnego metod zastrzeżenia, tak aby umożliwić korzystanie z uprawnień przez podmioty uprawnione oraz pewność co do prawa dla podmiotów dokonujących TDM na podstawie art. 4 ust. 3.

Podsumowanie i postulaty

Wyjątek zawarty w art. 4 dyrektywy DSM, pozostaje niewystarczający z uwagi na dalszy brak pewności co do prawa z perspektywy podmiotów dokonujących web scrapingu danych szkoleniowych dla AI oraz z perspektywy podmiotów uprawnionych. Dla uprawnionych w praktyce wprowadzenie zastrzeżenia, o którym mowa w art. 4 ust. 3, do witryny, w której udostępniają oni swoje utwory, może okazać się nadmiernie utrudnione lub niemożliwe. Pozostaje niepewność co do znaczenia terminu „możliwości odczytania maszynowego”, na którego interpretacji orzecznictwo zapewne „połamie sobie zęby” z uwagi na fakt, że prawodawca unijny nie ujednolicił ram dla zastrzeżeń w metadanych. Dostawcy rozwiązań opartych na AI aktualnie ewidentnie stawiają tę metodę przed zastrzeżeniem dokonany w warunkach strony internetowej, a prawodawca unijny powinien rozważyć wprowadzenie jednolitej metody zastrzeżenia, do której to dostawcy rozwiązań będą musieli się dostosować.

W istocie projektując rozwiązania, które miały na celu wywarzyć zapewnienie rozwoju technologii z przeciwnymi interesami podmiotów uprawnionych, w opinii autora prawodawca unijny powinien mieć na względzie najbardziej interesujący dla obu stron aspekt – finansowy.

³³ N. Clarke, *Block the Bots that Feed “AI” Models by Scraping Your Website*, <https://neil-clarke.com/block-the-bots-that-feed-ai-models-by-scraping-your-website/> [dostęp 30.04.2024].

Wydaje się, że zdecydowanie bardziej praktycznym rozwiązaniem, które mogłoby ograniczyć przyszłe spory wynikające z poszczególnych nieścisłości dotyczących legalnego przeprowadzania procedury TDM, byłoby upoważnienie organizacji zbiorowego zarządzania do negocjowania zbiorowych umów licencyjnych z podmiotami prywatnymi, które dokonują web scrapingu, tak jak prawodawca unijny uczynił to w przypadku korzystania przez instytucje dziedzictwa kulturowego z niedostępnych w obrocie handlowych utworów i innych przedmiotów objętych ochroną w art. 8 dyrektywy DSM.

Co więcej, niejednolita praktyka podmiotów rynkowych co do plików *robots.txt* oraz brak realnego wpływu na te metadane przez podmioty uprawnione udostępniające swoje utwory za pośrednictwem podmiotów zewnętrznych zdecydowanie warunkują potrzebę wprowadzenia jednolitych ram dla zastrzeżenia, o którym mowa w art. 4 ust 3 dyrektywy DSM, oraz obowiązek uwzględnienia woli podmiotów uprawnionych w zakresie dokonania zastrzeżenia. W tym zakresie wydaje się, że wystarczające byłoby wprowadzenie wymogu uzyskania stanowiska podmiotu uprawnionego (np. w postaci prostego *checkboxu*).

Bibliografia

Źródła

- Dyrektywa Parlamentu Europejskiego i Rady (UE) 2019/790 z 17.04.2019 r. w sprawie prawa autorskiego i praw pokrewnych na jednolitym rynku cyfrowym oraz zmiany dyrektyw 96/9/WE oraz 2001/29/WE (Dz. Urz. UE L 160/62 z 17.05.2024 r.).
- Ustawa z dnia 4 lutego 1994 r. o prawie autorskim i prawach pokrewnych (Dz. U. z 1994 r. Nr 24, poz. 83 ze zm.).
- Rozporządzenie Parlamentu Europejskiego i Rady (UE) 2024/1689 z dnia 13 czerwca 2024 r. w sprawie ustanowienia zharmonizowanych przepisów dotyczących sztucznej inteligencji oraz zmiany rozporządzeń (WE) nr 300/2008, (UE) nr 167/2013, (UE) nr 168/2013, (UE) 2018/858, (UE) 2018/1139 i (UE) 2019/2144 oraz dyrektyw 2014/90/UE, (UE) 2016/797 i (UE) 2020/1828 (akt w sprawie sztucznej inteligencji) (Dz. Urz. UE L z 12.07.2024 r.).
- Rozporządzenie Parlamentu Europejskiego i Rady (UE) 2022/2065 z dnia 19 października 2022 r. w sprawie jednolitego rynku usług cyfrowych oraz zmiany dyrektywy 2000/31/WE (akt o usługach cyfrowych) (Dz. Urz. UE L 277/65 z 27.10.2022 r.).
- Dyrektywa 96/9/WE Parlamentu Europejskiego i Rady z dnia 11 marca 1996 r. w sprawie ochrony prawnej baz danych (Dz. Urz. UE L 77/39 z 27.03.1996 r.).
- Dyrektywa 2001/29/WE Parlamentu Europejskiego i Rady z dnia 22 maja 2001 r. w sprawie harmonizacji niektórych aspektów praw autorskich i pokrewnych w społeczeństwie informacyjnym (Dz. Urz. UE L 167/44 z 22.06.2001 r.).
- Dyrektywa Parlamentu Europejskiego i Rady 2009/24/WE z dnia 23 kwietnia 2009 r. w sprawie ochrony prawnej programów komputerowych (Dz. Urz. UE L 111/52 z 5.05.2009 r.).
- Ustawa z dnia 7 października 1999 r. o języku polskim (Dz. U. z 1999 r. Nr 90, poz. 999 ze zm.).

Orzecznictwo

Wyrok Trybunału Sprawiedliwości Unii Europejskiej z dnia 10 kwietnia 2014 r. w sprawie C-435/12, *ACI Adam BV i in. v. Stichting de Thuiscope i Stichting Onderhandeligen Thuiscope vergoeding*, EU:C:2014:254.

Literatura

- Barta J., Markiewicz R., *Prawo autorskie i prawa pokrewne*, Warszawa 2021.
- Fischer B., Pązik A., Świerczyński M., *Prawo sztucznej inteligencji i nowych technologii*, Warszawa 2021.
- Flaga-Gieruszyńska K., Gołaczyński J., Szostek D., *Sztuczna inteligencja, blockchain, cyberbezpieczeństwo oraz dane osobowe. Zagadnienia wybrane*, Warszawa 2019.
- Grzybczyk K., *Fair use jako przykład ograniczenia monopolu autorskiego klauzulą generalną*, „Zeszyty Naukowe Uniwersytetu Jagiellońskiego” 2020, nr 1.
- Krotov V., Silva L. et. al., *Twenty-fourth Americas Conference on Information Systems*, New Orleans 2018.
- Machała W., *Dozwolony użytek chronionych utworów w polskim prawie autorskim w świetle cywilistycznej koncepcji prawa*, „Zeszyty Naukowe Uniwersytetu Jagiellońskiego” 2001, nr 78.
- Markiewicz R., *Prawo autorskie na jednolitym rynku cyfrowym. Dyrektywa Parlamentu Europejskiego i Rady (UE) 2019/790*, Warszawa 2021.
- Senftleben M., *Generative AI and Author Remuneration*, „International Review of Intellectual Property and Law” 2023, nr 54.
- Schintler A., McNeely C.L., *Encyclopedia of Big Data*, „Springers International Publishing” 2017,
- Sięńczyło-Chlabicz J., *Prawo własności intelektualnej. Teoria i praktyka*, Warszawa 2021.

Internet

- Clarke N., *Block the Bots that Feed “AI” Models by Scraping Your Website*, neil-clarke.com/category/technical, 23.8.2023.
- GETTY IMAGES CONTENT LICENSE AGREEMENT, Restricted Uses, lit. k, <https://www.gettyimages.com/eula>
- GPTBot, OpenAI Platform, <https://platform.openai.com/docs/gptbot>
- <https://neil-clarke.com/block-the-bots-that-feed-ai-models-by-scraping-your-website/>
- <https://pobierz.1-800-oswiecenie.com/>
- Licencja dla innych użytkowników, <https://www.youtube.com/t/terms>
- Paul K., *Meta used copyrighted books for AI training despite its own lawyers’ warnings, authors allege*, <https://www.reuters.com/technology/meta-used-copyrighted-books-ai-training-despite-its-own-lawyers-warnings-authors-2023-12-12/>
- Pozew Getty Images Inc. v. Stability AI inc., <https://fingfx.thomsonreuters.com/gfx/legaldocs/byvrlkmwnve/GETTY IMAGES AI LAWSUIT complaint.pdf>
- Projekt z dnia 18 kwietnia 2024 r. ustawa o zmianie ustawy o prawie autorskim i prawach pokrewnych oraz niektórych innych ustaw, <https://legislacja.rcl.gov.pl/docs//2/12382002/13037430/dokument664634.zip>

Reuters, Technology, [reuters.com/technology](https://www.reuters.com/technology/), 13.12.2023.

WorldWideWebSize.com. Daily estimated size of the world wide web, <https://www.worldwide-websize.com/#:~:text=Web> (The Internet)

Web scraping for artificial intelligence training v. author's position

Abstract

In the age of artificial intelligence (AI) and widespread web scraping, artworks available on the internet are copied daily for the purpose of obtaining training data for AI models. In the U.S., where many companies are providing AI-based solutions, we are seeing an increase in the number of lawsuits related to this issue. The providers of these solutions defend themselves by invoking the doctrine of fair use. In the same cases in Europe, the arguments will most likely focus on the exceptions to the exclusive rights introduced by the EU legislator in directive (EU) 2019/790 of the European Parliament and of the Council. However, an analysis of the exceptions introduced by the directive leads to the conclusion that these regulations do not in fact remain optimal from the perspective of rightholders or scrapers.

Keywords

webscraping, text and data mining, DSM Directive, fair use, artificial intelligence, machine learning, deep learning, large language models, use restriction, author's rights