Možnosti ochrany vydavatelského obsahu před zneužíváním AI a jejich LLM

Autor: Ondřej Ježek
Datum: 6. března 2025

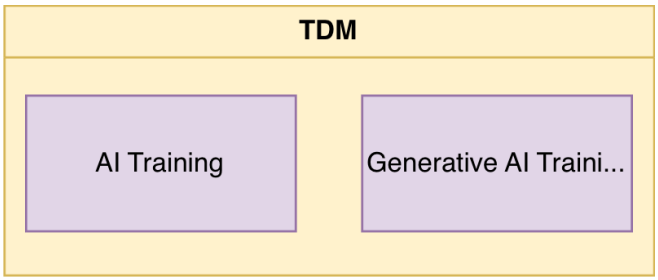
Úvod

Blokování robotů a ochrana obsahu vůči AI je pro české vydavatele **strategická nutnost v digitální éře**, ale vyžaduje vyvážený přístup. Český trh má svá specifika (např. podíl Seznamu na vyhledávání nebo aktuální stav mediální legislativy), přesto platí obecná zásada: **chraňte svůj obsah chytře, ne na úkor vlastní viditelnosti**.

V tomto dokumentu naleznete seznam doporučení, které mohou zvýšit ochranu českých vydavatelů proti vytěžování dat AI platformami a jejich Language Learning Modely (LLM).

Níže uvedená doporučení vycházejí z veřejně dostupných dat, debat v rámci pracovní skupiny Platformy při Asociaci online vydavatelů (AOV). Hlavním zdrojem informací jsou závěry ze setkání s kolegy z News Media Europe, kteří v této oblasti poskytují AOV dostatečnou znalostní bázi. Oblast AI a LLM se pravidelně řeší v pracovní skupině „Copyright“, ve které je členem předseda AOV Libor Matoušek.

Obecně je potřeba rozlišovat pojmy TDM (Text and data mining), AI a generativní AI. Z hlediska struktury stojí TDM nad AI a generativní AI. Negenerativní AI vztažena k vydavatelskému obsahu většinou zobrazuje výsledky, popisuje trendy a netvoří nový obsah. Oproti tomu generativní AI z vydavatelských dat (novinových článků) tvoří obsah nový.



Kombinace opatření je klíč

Kombinací opatření – od technických (robots.txt, meta tagy, omezení přístupu) přes právní (TDMRep, úprava podmínek) až po spolupráci napříč trhem – lze **minimalizovat riziko, že AI modely nekontrolovaně vytěží obsah**, a současně **neublížit SEO**. Vydavatelé by měli průběžně vyhodnocovat efekt svých kroků: ideálem je, aby běžný uživatel nic nepocítil (nadále našel obsah ve vyhledávači), zatímco neautorizovaný bot narazí na překážky.

Do budoucna se dá očekávat, že se prostředí ustálí – buď vzniknou **standardy (formální „stopky“ pro AI)** uznávané napříč, nebo se prosadí **licenční modely** (kde vydavatelé budou za využití obsahu AI systémy kompenzováni). Pro české vydavatele je klíčové být u toho od počátku: **nastavit si interně jasnou politiku** k AI, implementovat snadná ochranná opatření hned teď a aktivně se účastnit diskuse o možných řešeních. Tím si zajistí, že je nepřekvapí ani rychlý nástup AI vyhledávání, ani případné změny algoritmů, a ochrání hodnotu svého obsahu pro budoucnost.

Technické metody ochrany obsahu

Robots.txt a blokování AI botů

Nástrojem je soubor **robots.txt**, v němž lze zakázat přístup konkrétním robotům. Vydavatel může explicitně blokovat známé AI crawlery – například GPTBot (crawler OpenAI pro ChatGPT), CCBot (Common Crawl), Google-Extended (Google pro AI účely), Anthropic-ai a další. V souboru robots.txt se uvedou pravidla typu:

User-agent: GPTBot
Disallow: /

Při nasazování omezení je nutné přesně zacílit na nežádoucí boty. Chybné nebo příliš široké pravidlo v robots.txt by mohlo **vyloučit i důležité indexovací boty**. Obecně platí zásada: **povol explicitně hlavní vyhledávače, blokuj konkrétně ty ostatní**.

Efektivita: Robots.txt je tradiční a standardní způsob, jak deklarovat, které crawlery smí na web. Je však nutné, aby jej roboti respektovali – soubor sám technicky nebrání přístupu, jen dává pokyn k dobrovolnému dodržení. Renomované firmy (Google, OpenAI aj.) avizovaly, že své AI crawlery budou pokyny robots.txt dodržovat a nabízejí tak opt-out. Problém nastává u méně známých či nepoctivých botů – jsou jich potenciálně tisíce a ne všichni se identifikují zvláštním user-agentem. Udržovat kompletní seznam je prakticky nemožné. Někteří AI roboti dokonce používají falešný user-agent (imitují běžné prohlížeče), aby se vyhnuli detekci.

Závěr: Robots.txt je základ (mnoho velkých vydavatelů typu NY Times či Reuters už plošné blokování AI botů zavedlo), ale sám o sobě nezaručí stoprocentní ochranu – slouží spíše k vymezení právně doložitelného nesouhlasu (opt-out) a k blokaci slušných crawlerů.

W3C TDM Reservation Protocol

Jde o nově navržený standard vyvinutý ve W3C komunitě jako reakce na čl. 4 evropské směrnice DSM (2019/790), který umožňuje strojově čitelně vyhradit práva k textovému a datovému dolování (Text and Data Mining). Cílem protokolu je poskytnout jednotný způsob, jak dát najevo TDM Reserved – tedy že si nositel práv vyhrazuje souhlas s využitím svého online obsahu pro účely dolování dat.

Implementace: TDM Reservation Protocol definuje dvě hlavní informace: boolean příznak, zda jsou těžební práva vyhrazena, a URL tzv. TDM policy, kde jsou podmínky a kontakt na vydavatele. Tyto informace lze publikovat několika způsoby:

- Jako JSON soubor v adresáři */.well-known/tdmrep.json* na webu.
- Pomocí HTTP hlaviček vracených u každé stránky.
- Jako meta tag v HTML stránce. Crawler tedy může snadno zjistit, jestli daný obsah smí použít pro TDM, případně najde odkaz na licenci/kontakt pro získání povolení.

Hodnocení: Jedná se komplexnější a robustnější přístup než jednoduchý robots.txt. Umožňuje sdělit i "chcete-li těžit, kontaktujte nás zde a licenci lze získat za XY podmínek". **Velcí vydavatelé v EU jej prosazují jako jednotný standard.** Nevýhodou je, že jde o novinku. Žádný velký vyhledávač či AI crawler zatím otevřeně neoznámil podporu TDMRep a implementace vyžaduje úpravy na straně vydavatele (vygenerovat soubor či meta tag). **Do budoucna ale může jít o klíčový nástroj v rámci EU právního rámce** – pokud vydavatel takto deklaruje opt-out, AI těžař v EU, který by to ignoroval, by jednal protiprávně (porušení autorského práva dle implementace čl. 4 DSM směrnice).

Meta tagy a HTML pokyny (noai apod.)

Další metodou jsou speciální metatagy v HTML stránce nebo záhlaví HTTP, které robotům sdělují, jak mohou obsah používat. Nově se objevují neformální meta-štítky jako „noai" a „noimageai", které vyjadřují nesouhlas s použitím obsahu (textu či obrázků) k trénování AI. Tyto tagy byly původně navrženy komunitou kolem DeviantArt pro ochranu děl výtvarníků a postupně je adoptují i blogeři a vydavatelé. Přidávají se do sekce *<head>* stránky (např. *<meta name="robots" content="noai, noimageai">*).

Výhoda: snadná implementace.

Nevýhoda: nejde o oficiální standard a nelze zaručit, že je současné boty budou respektovat. Jde spíše o signalizaci postoje a snahu vytvořit tlak na vznik standardu, který by byl respektovaný AI společnostmi. Tuto hodnotu ignoruje i Google.

`ai.txt` soubor:

Jedná se o **neoficiální iniciativu** organizace Spawning.ai a dalších, inspirovanou robots.txt, avšak zaměřenou přímo na AI trénování. Soubor *ai.txt* umístěný v kořenové složce webu by obsahoval pravidla, která upřesňují povolení či zákaz použití obsahu webu pro AI. Na rozdíl od robots.txt, který řeší spíše kdo smí/leze kam, by ai.txt rozlišoval typy obsahu (text, obrázky, video, kód) a k nim přiřadil, zda smějí být použity ke komerčnímu trénování AI. Tvůrci jej propagují jako jednoduchý způsob, jak dát AI firmám vědět, co z vašeho webu smějí využít – například povolit trénování na obrázcích, ale zakázat na textech, atd.

Stav: Podobně jako meta „noai" jde o komunitní návrh. Nejsou známé AI crawlery, které by ai.txt skutečně četly. Někteří vydavatelé jej však začali přidávat preventivně. Slouží tedy spíše jako vývěska pro AI firmy o podmínkách použití. Do podobné kategorie patří i různé manuální blokace na serverové úrovni.

SEO a indexace

Implementace výše zmíněných opatření může výrazně ovlivnit viditelnost obsahu ve vyhledávačích. Vydavatelé proto musí postupovat opatrně, aby **nepoškodili vlastní SEO**. Zde je několik aspektů:

- **Dopad robots.txt na Google a spol.:** Tradiční vyhledávače (Google, Seznam, Bing) respektují robots.txt striktně. Pokud by vydavatel omylem zablokoval jejich uživatelské agenty (Googlebot apod.), daný obsah se přestane indexovat.
- **Indexace vs. zneužití obsahu:** Někteří vydavatelé zvažují omezit zveřejňovaný rozsah obsahu. Například publikovat jen perexy a část článku volně a zbytek mít za přihlášením/paywallem. Vyhledávač indexuje základ, ale plný text není volně ke stažení pro scrapery. To sice sníží riziko kompletního zneužití obsahu, ale zároveň to může snížit SEO hodnotu (Google nemá celý text pro hodnocení relevance) a uživatelskou vstřícnost. Je to tedy vhodné spíše pro prémiový obsah.
- **AI vyhledávače a nové indexace:** Objevují se nové AI-driven vyhledávače (např. Perplexity či Bing Chat apod.), které fungují jinak než tradiční search. Perplexity má vlastní crawl („PerplexityBot“) pro indexování webů, ale při zodpovídání dotazů používá i on-demand načítání stránek přes agenta „Perplexity-User“. Pro vydavatele to znamená, že **i neindexovaný obsah může být načtený, pokud se dostane do kontextu dotazu uživatele**. Navíc Perplexity-User **ignoruje robots.txt, pokud se jedná o přímý požadavek vyvolaný uživatelem** (bere to, jako by si stránku otevřel uživatel v prohlížeči). Perplexity bylo dokonce obviněné, že jeho bot se maskoval jako běžný návštěvník, aby obešel omezení. Z hlediska SEO je otázka, zda být v indexu takové služby výhodou, či nikoli. Pokud povolíte PerplexityBot indexaci, vaše stránky se mohou zobrazovat jako zdroje v odpovědích – Perplexity u odpovědí uvádí citace s odkazem, takže uživatel může na váš web přejít. Teoreticky to může přinést nějakou návštěvnost od uživatelů AI asistentů. Pokud však váš obsah AI plně zodpoví dotaz uživatele, mnoho lidí už na odkaz neklikne, protože odpověď získali okamžitě. Rizikem tedy je **kanibalizace organické návštěvnosti** – AI vyhledávač spotřebuje váš obsah k uspokojení dotazu, aniž by uživatel navštívil web. První data z nasazení Google SGE (Search Generative Experience) naznačují, že dopad na prokliky zatím není tak dramatický, jak by se možná čekalo. Odhady hovoří i o 25% poklesu návštěv, což není ani málo. Google tvrdí, že AI overviews mohou naopak zvýšit počet kliků na zdroje, ale konkrétní data zatím neposkytl. Perplexity a jemu podobní jsou zatím minoritní hráči, ale mohou naznačovat trend. Pro SEO strategii to znamená: **Měli byste sledovat, odkud doopravdy chodí návštěvy**. Pokud zjistíte, že např. Perplexity sice cituje váš web, ale návštěv z něj mnoho nepřichází, můžete zvážit jeho blokování (např. v robots.txt povolit jen vyhledávače, ale zablokovat PerplexityBot). Tím se váš obsah přestane v této službě objevovat – uživatel Perplexity dostane méně kvalitní odpověď nebo odkaz na jiný zdroj. Otázkou je, zda si tím nepodřezáváte větev v dlouhodobějším horizontu, kdy by AI vyhledávání získalo větší podíl uživatelů.

Alternativní právní a licenční strategie

Technické prostředky samy o sobě nemusí stoprocentně zabránit, aby někdo váš obsah použil v tréninkových datech. Proto je nutné opřít se také o **právní ochranu a licencování**. Několik možností:

- **Uplatnění autorských práv a licenčních nároků:** V České republice je obsah článků (pokud je původní a kreativní) chráněn autorským zákonem. **Autorská ochrana vzniká automaticky** v okamžiku vytvoření díla, není třeba nikde nic registrovat. Nicméně pro posílení své pozice může autor či vydavatel **zaregistrovat dílo u kolektivního správce** 🙂.

- **Práva vydavatelů (licenční poplatky od platforem):** Pokud AI chatbot cituje celé pasáže z článků, mělo by se to považovat za užití tiskové publikace vyžadující licenci. Vydavatelé v EU tak mají **páku pro vyjednávání** – mohou říct AI firmám: „Trénovali jste či používáte naše články? Pak potřebujete licenci, jinak porušujete autorský zákon.“ **To by s ohledem na velikost českého trhu mělo být vymáháno spíše kolektivně**. Již nyní některé zahraniční redakce uzavírají dohody – např. agentura **AP (Associated Press) licencovala svůj archiv OpenAI** pro trénování AI modelů. Podobné dohody zvažují další mediální domy výměnou za přístup k technologiím nebo finanční plnění. **Pro české vydavatele** to znamená: Měli by se hlásit o svá práva – dát najevo, že obsah je chráněn a není volně k tréninku.
- **DMCA a odstraňování nelegitimního obsahu:** Pokud zjistíte, že vaše články byly zkopírovány či zneužity na konkrétní platformě či webu (např. někdo vytvoří stránku generující obsah z vašich textů, nebo váš obsah je součástí tréninkového datasetu zveřejněného online), můžete využít **DMCA (Digital Millennium Copyright Act)** takedown procesu. DMCA je americký mechanismus, ale prakticky ho respektují globální platformy (Google, Github atd.). Umožňuje nahlásit porušení copyrightu a dosáhnout stažení obsahu. Pro AI je to trochu obtížnější – pokud je obsah přímo na webu, DMCA ho může odstranit (např. datový soubor s vašimi články někde na GitHubu by byl stažený). U již natrénovaného modelu je odstranění konkrétního díla prakticky nemožné bez přeškolení celého modelu. Přesto lze DMCA využít nepřímo: například podat žádost Googlu, aby odstranil ze svých indexů stránky, kde AI model (nebo kdokoli) zveřejňuje vaše texty bez licence. Někteří autoři už DMCA využívají proti výstupům ChatGPT, pokud generuje doslovné části jejich děl. Google i další vyhledávače také označují stránky, proti nimž se vzneslo DMCA, což znesnadní jejich šíření. **Shrnutí:** Mějte připravené postupy pro případ, že objevíte neoprávněné užití obsahu – zdokumentujte ho a využijte právní nástroje (DMCA, výzva k ukončení protiprávního jednání).
- **Další mechanismy:** Doporučuje se upravit podmínky užití na webu – vložit ustanovení, že veškerý obsah je chráněn a „je zakázáno ho strojově sbírat či využívat k trénování AI bez výslovného povolení“. Tím se posílí vaše právní postavení (porušení podmínek = neoprávněný přístup).
- Celkově platí, že **právní rámec se teprve vyvíjí**. Vydavatelé by měli společně tlačit na vytvoření jasných pravidel a precedentů. První větší spory (např. žaloby Getty Images proti tvůrcům AI za zneužití dat, žaloby autorů kódu proti GitHub Copilot apod.) už probíhají a jejich výsledek naznačí, co si AI firmy mohou dovolit. **Aktivní přístup je na místě.**

Dopady na ekosystém: Pokud by většina vydavatelů striktně zablokovala AI přístup, AI vyhledávače by měly hlad po datech a možná by musely uzavírat dohody (což by bylo pro vydavatele finančně pozitivní). Na druhou stranu, kdyby uživatelé masově přešli k AI asistentům, kteří by kvůli blokacím měli nekvalitní odpovědi, mohl by to snížit důvěru v AI vyhledávání – a uživatelé by se vrátili ke klasickým vyhledávačům nebo přímo k ověřeným zdrojům. Je také možné, že se **změní model návštěvnosti**: méně kliků z vyhledávačů bude muset kompenzovat více **přímé návštěvnosti** (např. lidé budou chodit přes záložky, sociální sítě, newslettery) – protože pokud budou vědět, že AI jim dá jen obecný přehled, pro hlubší info půjdou přímo na oblíbený web. Vydavatelé by se tedy měli zaměřit i na **posilování přímého vztahu s publikem** (brand, předplatitelské modely, komunitní obsah), aby nebyli zcela závislí na zprostředkovatelích (ať už vyhledávačích nebo AI).

Doporučení

Optimální ochrana obsahu spočívá v kombinaci technických a právních opatření a ve strategické úvaze o budoucím směřování online konzumace obsahu. Vydavatelé mohou využít technické možnosti k vyjádření nesouhlasu se zneužíváním obsahu. Existuje však riziko, ač minimální, že tím omezí svou indexaci v důležitých vyhledávačích. Zároveň by měli aktivně uplatňovat svá práva. Klíčové je jednotné vystupování trhu: pokud většina významných médií nastaví podobné bariéry a začne jednat s AI firmami jednotně, zvýší šanci na uzavření výhodných dohod či vytvoření standardů, které budou klíčové společnosti respektovat.

AI vyhledávače a chatboti jsou novou realitou – mohou být hrozbou pro tradiční model návštěvnosti, ale mohou se stát i novým zdrojem příjmů i trafficu, pokud se podaří nastavit pravidla férové spolupráce. Znovu platí, že pokud dokáže trh postupovat jednotně je větší šance na vytvoření prostředí, kde AI modely nemohou volně krást obsah bez následků. Balancování mezi otevřeností webu a ochranou obsahu je nepochybně výzvou.

Cílené blokování AI robotů:

- Pomocí robots.txt explicitně blokování známé AI crawlery (např. GPTBot, Google-Extended).
- Nepoužívat příliš obecná pravidla, protože nemusí zachytit všechny nežádoucí roboty a riskujete tím omezení přístupu i pro neškodné systémy.

Implementace strojově čitelné rezervace práv:

- Integrace TDMRep do webu.
- Jasně deklarovat podmínky použití obsahu – například uvedením informací, že obsah je chráněn a není určen k automatizovanému dolování dat či tréninku AI bez licenční dohody. Tím si posílíte právní postavení.

Diferencovaný přístup k ochraně obsahu:

- Ochranu přizpůsobit hodnotě a trvanlivosti jednotlivých typů obsahu. Vysoce hodnotné materiály (investigativní reportáže, datové analýzy) mohou vyžadovat přísnější ochranu (např. registraci), zatímco pro rychle zastarávající zpravodajské články postačí základní ochrana přes robots.txt.

Monitorování vývoje AI vyhledávání a návštěvnosti:

- Sledovat trendy u AI vyhledávačů (Perplexity, YouChat, Bing Chat, Google SGE apod.) a analyzujte, jaký vliv mají na organickou návštěvnost.

Pravidelná analýza a kolektivní jednání:

- Využívat nástroje jako Google Search Console a monitorovat přístupové logy k identifikaci neznámých nebo problematických botů.
- Analyzovat dopady zavedených opatření a být připravený na úpravu
- Spolupráce s ostatními vydavateli (prostřednictvím například AOV) na tvorbě jednotných standardů a právních postupů proti neoprávněnému využívání obsahu.