GOUVERNEMENT

Liberté
Égalité
Fraternité

Pôle d'Expertise de la Régulation Numérique - PEReN

Note d'état des lieux sur l'opt-out

Réalisée par le PEReN au titre de son programme de travail 2024, cette note a été produite pour le ministère de la Culture qui a souhaité la rendre publique.

La multiplication des systèmes d'intelligence artificielle générative soulève des enjeux sur les droits des éditeurs de presse, leurs contenus étant largement utilisés par ces systèmes, dans un rapport qui ne fait pas toujours consensus entre ces acteurs.

Il existe un cadre technique minimal qui permet aux éditeurs de sites de déclarer quelles parties de leurs publications peuvent être protégées des robots de collecte de données grâce à des protocoles d'opt-out comme robots.txt. Ce dernier, de loin le plus largement utilisé présente quelques limitations : d'une part il n'est pas toujours adapté (manque de granularité ou tout simplement mauvaise configuration par certains sites), d'autre part il repose sur un système de confiance, dans lequel le robot s'auto déclare et doit ensuite respecter de lui-même les consignes indiquées par le site qu'il visite.

Ces limites peuvent éroder la confiance entre certains acteurs, alors que la bonne appropriation de ces technologies de réservation et de collecte pourrait permettre d'apporter aux acteurs une plus grande transparence et une objectivation des enjeux, comme préalable à la recherche d'un équilibre entre valorisation des données et contenus et innovation dans un internet qui reste ouvert.

Cette note technique à visée pédagogique présente un état des lieux à mi 2024 des principaux systèmes d'opt-out.

Table des matières

Contexte et état des lieux

Les robots d'exploration (*moissonneurs de données* ou *crawlers* en anglais) sont des automates qui récupèrent du contenu sur internet. Ces robots existent depuis de nombreuses années et collectent historiquement différents contenus dans le cadre du fonctionnement des moteurs de recherche comme *Google Search*. Les robots des moteurs de recherche collectent ainsi le contenu des sites (notamment les sites de presse) et les métadonnées associées (URL, titre, mots-clés, etc.) afin de pouvoir répondre à des recherches d'utilisateurs en leur proposant des ressources pertinentes. Dans ce cadre, l'*indexation* du contenu collecté (c'est-à-dire son intégration dans le moteur de recherche) est essentielle. C'est elle seule qui permet au moteur de recherche d'identifier le contenu et de le référencer, le rendant ainsi accessible et visible aux internautes.

D'autres acteurs historiques comme les agrégateurs et curateurs de contenus ont également recours à des robots d'exploration pour automatiser la collecte des données leur servant à développer leurs produits, par exemple pour créer des newsletters citant les articles de presse sur une thématique spécifique.

Avec le développement de l'intelligence artificielle (IA) générative ces dernières années, en particulier des grands modèles de langue (appelés par la suite « LLMs » pour *Large Language Models*), de technologies associées comme les agents conversationnels (tels que ChatGPT, lancé en novembre 2022) et de systèmes intégrant plus explicitement des contenus connus dans leurs réponses, le paysage formé par les robots d'exploration s'est largement transformé. Les développeurs et fournisseurs de systèmes d'IA ont besoin d'importantes quantités de données pour entraîner leurs systèmes et plus généralement pour assurer leur fonctionnement. À cette fin, ils opèrent leurs propres robots de moissonnage, ou bien utilisent des données préalablement collectées par des tiers. Les acteurs qui opèrent les robots peuvent être déjà bien implantés (acteur hybride qui opérait déjà un service de recherche par exemple, à l'instar de Google ou Microsoft) ou bien totalement nouveaux (tels que OpenAI, Perplexity ou Anthropic par exemple). Ainsi, les robots représentent aujourd'hui une part importante du trafic sur internet. En France, l'entreprise *Cloudflare* estime à environ 26 % la part des requêtes liées aux robots (dédiés ou non à la collecte) dans le trafic total sur son réseau de diffusion[1].

Dans un contexte où les réglementations ne sont pas harmonisées au niveau mondial, un volume immense de données – parmi lesquelles les œuvres de l'esprit (œuvres audiovisuelles, musicales, littéraires, journalistiques, etc) occupent une place très importante – a déjà été collecté et utilisé pour l'entraînement d'un grand nombre de modèles d'IA, dans la plupart des cas, sans accord préalable ni même information, des propriétaires de ces données. Et cette collecte se poursuit de façon continue afin d'améliorer et d'actualiser les modèles d'IA.

Aux États-Unis, le débat autour des droits de propriété intellectuelle afférents aux œuvres utilisées pour ces entraînements est suspendu à des décisions de cours de justice devant préciser les contours de l'exception de *fair use*.

[1] https://radar.cloudflare.com/traffic/fr?dateRange=52w

Dans l'Union Européenne, la directive sur le droit d'auteur de 2019 fixe un cadre juridique, relatif à la « fouille de textes et de données » (*text and data mining*), dont l'application à l'entraînement des modèles d'IA est aujourd'hui assez largement admise. Une première exception, réservée au secteur académique, bénéficie aux organismes de recherche et institutions du patrimoine culturel qui effectuent des fouilles à des fins de recherche scientifique. Une seconde exception est ouverte à tous les usages, quelle que soit la finalité, y compris commerciale, sous réserve toutefois que le titulaire n'ait pas exprimé son opposition (*opt-out*) de manière appropriée, notamment par des procédés lisibles par machine pour les contenus mis à la disposition du public en ligne. Lorsqu'une telle opposition est exprimée, les fournisseurs de modèles d'IA doivent obtenir une autorisation des éditeurs s'ils souhaitent procéder à un entrainement sur leurs publications. Ce retour à l'exclusivité permet aux éditeurs de recouvrer une capacité de négociation en vue d'obtenir une rémunération.

Face à l'hétérogénéité des pratiques des éditeurs pour déclarer la manière dont ils réservent leurs droits, et à la course à la collecte massive de données que se livrent certains acteurs, les pratiques tant des acteurs qui mettent à disposition des contenus que ceux qui les collectent pourraient être améliorées.

Les *crawlers*

Afin de pouvoir exploiter les informations présentes sur internet[2], divers acteurs opèrent des robots en charge d'explorer internet et de collecter ces informations. Ces robots sont notamment identifiables par leur adresse IP ainsi que par un nom technique (une *immatriculation*) qu'ils indiquent aux sites web consultés dans un en-tête HTTP dédié : le *User-Agent* (« agent utilisateur »). Il est à noter que ce *User-Agent* est purement déclaratif et totalement à l'initiative de l'opérateur de ce robot.

Exemple d'un champ User-Agent. L'en-tête de requête HTTP indique qu'elle provient du robot GPTBot, qui collecte les données utilisées par OpenAI pour l'entraînement de modèles de fondation.

```
User-Agent: Mozilla/5.0 AppleWebKit/537.36 (KHTML, like Gecko); compatible; GPTBot/1.1; +https://openai.com/gptbot
```

Les robots n'ont pas tous les mêmes finalités et peuvent être catégorisés en fonction de ces dernières : indexation de contenu à référencer dans un moteur de recherche, archivage du web, utilisation pour de l'IA, etc. Nous reprenons ici la typologie utilisée par Dark Visitors[3], consolidée avec celle de Cloudflare Radar[4], afin d'expliciter les différentes catégories de robots utilisées en nous focalisant sur celles liées aux intelligences artificielles.

Il n'existe ainsi pas d'obligation pour les robots de s'identifier et déclarer leur finalité. Toutefois, ils sont incités à le faire afin d'établir un cadre de travail de confiance avec les éditeurs dont ils

[2] L'entraînement des modèles d'IA peut être réalisé à partir des œuvres mises à disposition du public sur Internet, sous réserve que l'accès aux œuvres concernées se fasse de manière licite.

[3] https://darkvisitors.com/agents

[4] https://radar.cloudflare.com/fr-fr/traffic/verified-bots

moissonnent les sites.

Catégorisation des *crawlers* liés à l'IA

Les différents types d'acteurs liés aux intelligences artificielles relevés par Dark Visitors sont les suivants :

- « *AI data scrapers* » (moissonneurs de données pour IA) pour désigner les robots utilisés afin de collecter des données en vue de l'entraînement d'intelligences artificielles. Leur objectif sera alors de collecter des contenus variés et de bonne qualité. Notons qu'un temps conséquent peut s'écouler entre la date de la collecte par ces robots, la date de l'entraînement des intelligences artificielles, et la date de mise en production des modèles entraînés sur les données collectées, ce qui peut rendre ardu le suivi de l'utilisation des contenus moissonnés ;
- « *AI search crawlers »* (indexeurs de données pour les recherches basées sur IA) : afin de proposer des réponses cohérentes avec l'actualité, certaines entreprises mettent en place des systèmes permettant aux modèles d'intelligence artificielle de s'appuyer sur des contenus récents, trouvés sur internet. Les *AI search crawlers* sont des robots ayant pour fonction de lister les contenus pouvant être utilisés d'une telle manière, afin d'alimenter en contenus récents un modèle déjà entraîné et mis en production.
- « *AI assistants* » (assistants d'IA) : les assistants d'IA sont quant à eux des robots spécifiques mis en place par un nombre plus réduit d'acteurs, afin de répondre aux requêtes spécifiques des usagers (*user-triggered fetchers*). Par exemple, lorsqu'un utilisateur demande explicitement à un modèle d'IA de résumer le contenu d'une page spécifique désigné par une URL, certains acteurs considèrent que les autorisations liées à cette demande devraient être distinctes des autorisations liées aux collectes faites directement pour le compte de l'entreprise[5]. Les robots correspondants à cette catégorie découlent de cette distinction et interviennent en cas de demandes explicites des utilisateurs et font office d'intermédiaires entre un utilisateur et un site.
- « *Undocumented AI Agents* » (robots d'IA non documentés) : certains robots ont pu être détectés par Dark Visitors, qui les soupçonne d'être liés à des entreprises d'IA, sans que celles-ci ne documentent lesdits robots, empêchant ainsi de connaître leur finalité précise.

La multiplicité des robots, aux finalités diverses, a des conséquences observables par les éditeurs sur la hausse significative de leur trafic, et peut les amener à de nouveaux investissements dans la capacité de leurs serveurs afin de maintenir une bonne qualité de service pour leur audience.

Il est extrêmement difficile d'établir un panorama objectif des pratiques de moissonnage (fréquences et dynamiques de collecte, part de trafic, etc) des différents acteurs en présence du

[5] Cette distinction a notamment été faite par l'IA Perplexity, ce qui a pu mener à la controverse explicitée ici : https://rknight.me/blog/perplexity-ai-is-lying-about-its-user-agent/ et par la suite à un changement de politique de la part de Perplexity, comme indiqué ici : https://www.perplexity.ai/hub/technical-faq/how-does-perplexity-follow-robots-txt.

fait de l'envergure et de la diversité d'Internet. Une des seules sources crédibles pour de telles statistiques est à trouver du côté des CDN (*Content Delivery Networks*) qui sont des points d'entrée incontournables de l'infrastructure moderne du Web. Ainsi, à notre connaissance, Cloudflare est le seul acteur à éditer un tableau de bord public permettant de visualiser de nombreuses statistiques pertinentes sur le sujet[6].

Enjeux posés par les *crawlers* aux finalités multiples

Cette catégorisation masque le fait que des acteurs de l'IA peuvent collecter des données dans un objectif principal, mais réutiliser la donnée collectée pour une autre finalité. Inversement, des acteurs dont les robots collectent de la donnée pour une activité différente peuvent vouloir réutiliser cette donnée pour de l'IA, sans nouvelle demande d'accord ou notification au propriétaire des données initialement collectées. On pense en particulier aux entreprises éditant des moteurs de recherche ou des agrégateurs d'actualités qui ont par ailleurs développé une activité dans le domaine de l'IA (Google et Microsoft notamment). Étant donné qu'il serait inefficace pour ces entreprises et pour les sites visités par leurs robots de collecter de multiples fois leurs données en fonction de la finalité (une fois pour l'indexation du moteur de recherches, une autre pour des finalités d'IA qui peuvent être elles-mêmes multiples), une unique collecte, mutualisée, est effectuée et il n'est donc pas évident pour un site visité par ces robots de savoir quelle finalité est réellement utilisée et encore moins de s'opposer à la collecte.

Dans les faits, Google et Bing ont mis en place des dispositions pour permettre aux sites de s'opposer en fonction des finalités avec un *opt-out* distinct entre leurs différents produits[7]. De plus, Google, acteur en position dominante sur le marché de la recherche, a indiqué que l'*opt-out* sur ses robots d'IA (Google-Extended ou GoogleOther) n'avait pas d'impact sur ses autres produits et notamment sur le référencement d'un site dans Google Search[8].

Ces mécanismes ne prévoient pas à l'heure actuelle de moyens permettant aux sites visités de voir en transparence que leurs directives sont respectées. En effet, ces acteurs hybrides ne visitent bien souvent les pages d'un site web qu'avec un seul robot, qui permet d'identifier l'acteur mais pas la finalité. Par exemple, Google visitera `robots.txt` avec son Googlebot (utilisé pour Google Search) puis toutes les autres pages que les directives lui permettent de visiter et d'indexer (cf. section infra sur le protocole des `robots.txt`). Néanmoins l'interprétation des directives qui déterminent l'usage des données collectées conformément au product-token Google-Extended[9] se fait sur ses serveurs et sans vérification possible et donc en toute opacité pour le site visité.

---

[6] https://radar.cloudflare.com/fr-fr/ai-insights?dateRange=52w

[7] Cette opposition peut se faire via un *product token* dédié dans les `robots.txt` (voir infra) pour Google, ou des balises meta spécifiques pour Microsoft Bing.

[8] https://developers.google.com/search/docs/crawling-indexing/google-common-crawlers?hl=fr#google-extended

[9] En particulier, ce comportement ne respecte pas strictement les recommandations du protocole `robots.txt` qui stipule, comme nous le verrons plus tard, que le nom utilisé par le bot devrait décrire son objectif (« The identification string SHOULD describe the purpose of the crawler ») et que le *product-token* devrait être une sous-partie du champ *User-Agent* du robot visitant le site : https://www.rfc-editor.org/rfc/rfc9309#name-the-user-agent-line.

La question sous-jacente du partage de valeur

Si les possibilités d'*opt-out* existent pour que les éditeurs puissent réserver les droits liés à leur contenu, d'une part les conditions techniques de cet *opt-out* sont essentielles à l'exercice des droits de manière effective, et d'autre part la question du partage de valeur est un sous-jacent particulièrement complexe et sensible.

L'Autorité de la Concurrence a ainsi prononcé en mars 2024 une sanction de 250 millions d'euros à l'encontre de Google pour le non-respect de certains de ses engagements pris en juin 2022 en lien avec les droits voisins. En particulier, « L'Autorité a [...] constaté que Google avait utilisé aux fins d'entraînement de son modèle fondateur des contenus des éditeurs et agences de presse, sans avertir ces derniers », ajoutant que Google ne proposait à cette date « pas de solution technique permettant aux éditeurs et agences de presse de s'opposer à l'utilisation de leur contenu par Bard (*opt-out*) sans affecter l'affichage des contenus protégés au titre des droits voisins sur les autres services de Google et en obérant ainsi la capacité des éditeurs et agences de presse à négocier une rémunération »[10].

Le règlement européen sur l'IA (AI Act)[11], ou RIA, qui entre progressivement en vigueur depuis août 2024, précise les obligations auxquelles sont soumis les fournisseurs d'IA. Ce règlement définit plusieurs catégories de modèles d'IA, parmi lesquelles se trouve celle des modèles d'IA à usage général, dont font partie les LLMs en raison de leur capacité à servir à un grand nombre de tâches. Dans l'article 53, qui doit être appliqué à partir d'août 2025, l**e RIA prévoit plusieurs niveaux d'obligation pour cette catégorie de modèles, dont des mesures de transparence et de documentation minimales : les fournisseurs d'IA seront dans l'obligation de produire un « résumé suffisamment détaillé du contenu utilisé pour entraîner le modèle », résumé, public, dont l'implémentation opérationnelle est en cours de discussion au niveau de l'Union européenne.** Cette transparence sur les sources ayant permis l'entraînement des systèmes d'IA en amont et la granularité de ce qui doit apparaître dans ce résumé sont au cœur de ce débat. Le 11 mars dernier, la Commission Européenne a publié la troisième ébauche de ce que devrait être le résumé des données d'entrainement des modèles génératifs[12].

Dans ce contexte, des acteurs déjà connus dans le marché de la diffusion de contenu commencent à proposer des solutions pour permettre aux éditeurs de site web de mesurer la proportion de robots d'exploration dans le trafic HTTP, afin d'éventuellement objectiver l'importance de l'accès au contenu par ces robots, en préalable souvent à des réflexions sur la valorisation de cet accès. C'est le cas notamment de Cloudflare[13] qui réfléchit à proposer ce service à ses clients.

D'autres acteurs, comme BotsCorner[14], se proposent d'analyser le trafic des sites pour comprendre

[10] https://www.autoritedelaconcurrence.fr/fr/communiques-de-presse/droits-voisins-lautorite-prononce-une-sanction-de-250-millions-deuros

[11] https://eur-lex.europa.eu/legal-content/FR/TXT/?uri=CELEX:32024R1689

[12] https://digital-strategy.ec.europa.eu/en/library/third-draft-general-purpose-ai-code-practice-published-written-independent-experts

[13] https://blog.cloudflare.com/cloudflare-ai-audit-control-ai-content-crawlers/#step-5-prepare-your-site-to-capture-value-from-ai-scanning

[14] https://www.botscorner.fr/

le comportement des robots (volume et fréquence de collecte, accès à des pages nécessitant des comptes utilisateurs, etc.) et leur typologie. BotsCorner propose également un service d'intermédiation entre les sites internet et les entreprises qui moissonnent les données des sites pour faciliter la mise en place d'accords sur l'utilisation des données.

De manière symétrique, des créateurs de contenus de plus en plus nombreux utilisent des IA génératives, dont certaines sont mises à disposition libres de droits, pour accélérer ou améliorer leur propre processus de création. Les éditeurs de presse peuvent ainsi dans certains cas bénéficier en retour des innovations produites par les développeurs d'IA.

Ces équilibres sont nouveaux et difficiles à établir. Cependant, on atteste une augmentation d'accords conclus entre fournisseurs de systèmes d'IA et groupes de médias (OpenAI avec Le Monde depuis mars 2024[15], et depuis janvier 2025, Google avec Associated Press[16], Perplexity avec Humanoid (Ebra)[17] et Mistral avec l'AFP[18]). A noter que ceux-ci portent à la fois sur l'entraînement des modèles d'IA et sur les services associés. Si les détails de ces contrats ne sont pas communiqués, ils attestent de la relation forte qui lie les concepteurs de systèmes d'IA et les créateurs de contenus dont font partie les éditeurs de presse.

Les protocoles d'*opt-out*

Plusieurs protocoles et standards permettent aujourd'hui d'édicter des directives relatives à l'exploration ou l'indexation de contenu de tout ou partie de son site par des robots, tout en restant dans un internet ouvert. Ces protocoles ne sont pas exclusifs les uns des autres.

`robots.txt`

Le protocole le plus largement respecté aujourd'hui est le protocole d'exclusion des robots ou *Robots Exclusion Protocol (RFC 9309[19])*. Ce protocole est souvent appelé `robots.txt`[20] car sa mise en œuvre repose sur l'utilisation d'un fichier nommé `robots.txt` placé à la racine de son site. Si le *crawler* respecte le protocole[21], avant d'accéder à une page d'un site web, il tentera d'accéder au fichier `robots.txt` et respectera les directives qui y figurent et s'appliquent à lui. Ainsi, si la page est indiquée comme interdite pour le *user-agent* concerné, le robot ne récupérera pas le contenu de la page. La syntaxe du protocole consiste en une succession d'interdictions (ou d'autorisations, bien que celle-ci soit accordée par défaut) pour chaque *product-token* mentionné. Par exemple, le texte suivant permet d'interdire le *crawling* des pages avec les chemins d'accès `/articles` et `/private-content` au robot `ClaudeBot`.

[15] Intelligence artificielle : un accord de partenariat entre « Le Monde » et OpenAI
[16] Google works with Associated Press for fresher results on Gemini app
[17] Perplexity signe avec Humanoid (Ebra) en France - mind Media
[18] L'AFP et Mistral AI annoncent un partenariat mondial | AFP.com
[19] https://www.rfc-editor.org/rfc/rfc9309.html
[20] https://robots-txt.com/
[21] C'est le cas de la majorité des grands acteurs, dont Anthropic : https://support.anthropic.com/en/articles/10023637-does-anthropic-crawl-data-from-the-web-and-how-can-site-owners-block-the-crawler.

```
User-agent: ClaudeBot
Disallow: /articles/
Disallow: /private-content/
```

L'utilisation du caractère * permet de donner des instructions pour l'ensemble des robots, ce qui inclut les robots d'exploration des moteurs de recherche et peut ainsi nuire au référencement du site. Par exemple, le texte suivant interdit l'ensemble du site à tous les robots d'exploration.

```
User-agent: *
Disallow: /
```

Il est crucial de noter qu'un acteur opérant un robot choisit unilatéralement, au moment d'interpréter les directives du `robots.txt`, le ou les *product-token* qui s'appliquent à son robot. Les seules contraintes énoncées dans la spécification du protocole concernant ce *product-token* sont d'ordre typographique (caractères autorisés, insensibilité à la casse). À cela s'ajoutent deux recommandations notables :

- que le *product-token* soit une sous-partie de son en-tête *User-Agent* (que l'acteur opérant le robot choisit également de manière unilatérale),
- et surtout qu'il décrive la finalité (*purpose*) de la collecte ; il est cependant à noter qu'aucun détail n'est fourni concernant ce qu'est une finalité de collecte, ce qui laisse en l'état une grande liberté d'interprétation sur cette spécification.

Le comportement à suivre dans le cas ou plusieurs *product-token* seraient applicables à un robot n'est en revanche pas spécifié dans le protocole[22].

Par ailleurs, un protocole connexe, non standardisé, existe : il s'agit des balises (*tags*) HTML meta robots[23] et de leur extension basée sur des en-têtes de requêtes HTTP (X-Robots-Tag ou Robots-Tag)[24], qui visent à préciser de manière plus granulaire les pratiques de collecte de données autorisées. Par exemple, à l'instar des instructions `noindex`, `nofollow` ou `nosnippet`, plus ou moins respectées par les crawlers de recherche[25], de nouvelles instructions `noai` et `noimageai` ont été proposées par DevianART fin 2022[26], mais aucune trace publique d'adoption de ces mots-clés par des moissonneurs de données n'a été repérée à ce jour.

Text and Data Mining Reservation Protocol (TDMRep)

Le *Text and Data Mining Reservation Protocol* (TDMRep)[27] est un protocole communautaire du

[22] On pense par exemple au *product-token* GoogleOther, et à ses versions spécialisées GoogleOther-Image|Video.

[23] Cf. https://robots-txt.com/meta-robots/, https://developers.google.com/search/docs/crawling-indexing/robots-meta-tag par exemple ou https://www.bing.com/webmasters/help/robots-meta-tags-and-attributes-that-bing-supports-5198d240.

[24] Cf. par exemple : https://robots-txt.com/x-robots-tag/.

[25] Cf. spécifications pour Google et Bing dans les liens de la note

[26] Cf. communiqué : https://www.deviantart.com/team/journal/UPDATE-All-Deviations-Are-Opted-Out-of-AI-Datasets-934500371

[27] Version courante : https://www.w3.org/2022/tdmrep/ (cf. https://www.w3.org/community/tdmrep/ pour les actualités du groupe et https://w3c.github.io/tdm-reservation-protocol/spec/index.html pour un brouillon de la prochaine version)

World Wide Web Consortium (W3C)[28], porté depuis 2021 par l'ONG française EDRLab[29], spécifiquement conçu pour traiter techniquement la question de l'*opt-out* dans le cadre de l'article 4 de la directive européenne de 2019 sur le droit d'auteur et les droits voisins dans le marché unique numérique[30]. A ce jour ce protocole semble très faiblement considéré par les moissonneurs de données, bien que des éditeurs, notamment français, ont commencé à l'adopter. Par ailleurs le protocole pourrait encore évoluer malgré une stabilisation depuis mai 2024. Le protocole repose sur trois mécanismes complémentaires que les ayants droit peuvent mobiliser pour exprimer leurs choix :

1. L'inclusion dans les en-têtes des réponses HTTP du serveur d'un champ booléen `tdm-reservation` exprimant la réservation ou non des droits TDM, et d'un champ optionnel `tdm-policy` permettant de préciser en langage machine les contours de la réservation des droits ainsi que l'éventuelle marche à suivre pour pouvoir exploiter le contenu.

2. L'utilisation d'un fichier `tdmrep.json`, qui définit la politique sur l'ensemble du site. Il doit être accessible à une URL spécifique et suivre une syntaxe particulière détaillée dans le standard. On y retrouve en particulier les champs `tdm-reservation` et `tdm-policy` (optionnel), cette fois accompagnés d'un champ `location`, indiquant les chemins correspondants aux descriptions de droits correspondants. Le fichier doit ainsi contenir une liste d'objets JSON, où chaque objet représente une règle et contient les clés `location`, `tdm-reservation` et `tdm-policy` (optionnellement)[31].

3. Une troisième option pour les ayants-droit est d'inscrire leurs préférences dans le contenu HTML des pages avec une balise `meta`, en utilisant les mêmes champs que dans la première modalité décrite[32].

Dans le cadre du protocole proposé, un robot d'IA doit d'abord vérifier la présence du fichier `tdmrep.json` sur le serveur avant de commencer à récupérer le contenu du site. De plus, il doit vérifier la présence des champs d'en-tête HTTP spécifiques au protocole TDMRep dans chaque page qu'il explore : lorsque présents les champs `tdm-reservation` et `tdm-policy` prévalent sur toute autre valeur déduite de l'éventuel fichier `tdmrep.json`. Enfin, il doit vérifier la présence de métadonnées TDM dans le contenu HTML récupéré, qui remplacent les valeurs précédentes.

Outre les sites Web, le protocole a été conçu pour pouvoir s'appliquer aux livres numériques (EPUB) ainsi qu'à des fichiers PDF, grâce à des métadonnées spécifiques incluses dans ces fichiers.

Autres initiatives

- Le fichier `ai.txt` est une autre initiative portée par la société privée américaine Spawning[33]

[28] Ce n'est donc pas un standard officiel du W3C, cf. la différence : https://www.w3.org/standards/types/#x2-1-w3c-community-group-report-or-w3c-business-group-report

[29] https://www.edrlab.org/about/

[30] https://eur-lex.europa.eu/legal-content/FR/TXT/?uri=CELEX:32019L0790

[31] https://www.w3.org/community/reports/tdmrep/CG-FINAL-tdmrep-20240510/#example-2

[32] https://www.w3.org/community/reports/tdmrep/CG-FINAL-tdmrep-20240510/#example-5

[33] https://spawning.ai/

qui reprend essentiellement la syntaxe des `robots.txt` et permet aux ayants-droit de préciser quelles données ne sont pas utilisables par des systèmes d'IA (registre « *Do Not Train* »). Les données spécifiquement couvertes par le protocole sont des fichiers hébergés sur un site Web (images/vidéos/audio en particulier) plutôt que le texte brut qu'on peut trouver sur ses différentes pages. Spawning propose une API qui vérifie si l'ayant-droit du contenu figurant à une URL a exercé son droit d'*opt-out* sur ce type de contenu via le fichier `ai.txt`. À ce jour, l'adoption de ce protocole semble encore moindre que celle du TDMRep. A noter que la société édite également la plateforme *Have I Been Trained?*[34] qui vise à identifier si des contenus font partie de base de données d'entraînement IA connues.

- Le protocole TDM·AI[35] est une autre initiative, encore en cours d'élaboration et relativement confidentielle portée par la société privée néerlandaise Liccium. Contrairement aux protocoles `robots.txt`, le protocole entend attacher des métadonnées exprimant l'*opt-out* au niveau des contenus eux-mêmes[36], et vise donc spécifiquement des fichiers originaux (images/vidéos/audio/EPUB/PDF) davantage que le texte brut présent sur des pages Web. Il repose sur des mécanismes cryptographiques non triviaux pour garantir l'authenticité des déclarations d'*opt-out* et être partiellement résistant aux manipulations des fichiers lors de leur dissémination sur le Web.
- Une dernière initiative portée par la Commission européenne est également en préparation : il s'agirait de mettre en place un registre unifié d'expression des réservations de droits, le *Open Rights Data Exchange*[37].

Il est à noter qu'un travail de synthèse intéressant publié en mars 2025 a été mené par l'ONG néerlandaise OpenFuture à propos du paysage des protocoles d'*opt-out* IA[38].

Limites des protocoles

Une caractéristique que partagent tous les protocoles décrits ci-dessus est que les informations fournies par les acteurs opérant des robots sont purement déclaratives. En effet, le choix technique de son *User-Agent* (et éventuellement d'un *product-token* associé pour le protocole des `robots.txt`) est totalement souverain, tout comme la décision de respecter la politique TDM d'un site décrit dans un fichier `tdmrep.json`. Certains opérateurs adoptent quant à eux des interprétations qui dévient parfois des spécifications officielles des protocoles de réservation de droits, exploitent les ambiguïtés ou non-dits de ces protocoles[39], ou encore ajoutent explicitement des spécifications pour répondre à leurs besoins propres[40].

---

[34] https://haveibeentrained.com/

[35] https://docs.tdmai.org/

[36] Cf. schéma https://docs.tdmai.org/options-for-metadata-binding

[37] https://ec.europa.eu/digital-building-blocks/sites/display/EBSI/Open+Rights+Data+Exchange

[38] https://openfuture.eu/publication/a-vocabulary-for-opting-out-of-ai-training-and-other-forms-of-tdm/ (en anglais)

[39] Par exemple, dans le cadre du `robots.txt`, certains robots, peuvent ne pas considérer les directives s'appliquant au robot universel * (cf. par exemple la mention explicite pour les crawlers spéciaux de Google), ou suivre les directives qui s'appliquent à un *product-token* plus connu que le leur (par exemple celles applicables au *Googlebot,* comme c'est le cas pour *Applebot* ou *Imagesift Bot*).

[40] Cf. https://developers.google.com/search/docs/crawling-indexing/special-tags ou https://www.bing.com/webmasters/help/which-robots-metatags-does-bing-support-5198d240 par exemple.

Le protocole `robots.txt` étant le plus largement adopté, nous proposons de développer plus spécifiquement certaines de ses limites. La première est due à l'asymétrie d'information entre les acteurs qui opèrent les robots et les propriétaires des sites *crawlés*. En effet, de nouveaux robots apparaissent régulièrement, et les comportements ou périmètres des robots identifiés peuvent également changer. Ces évolutions rapides induisent le besoin d'une veille permanente et proactive de la part des éditeurs de site s'ils veulent maintenir une réservation de droits la plus complète possible sans restreindre drastiquement l'accès à leur site (cf. infra la section Le risque d'un Internet qui se referme). En pratique ce travail est chronophage, difficilement exhaustif et arrive souvent trop tard (décalage entre l'édiction des directives à l'encontre d'un nouveau robot ou d'un robot dont le périmètre a été modifié, et un moissonnage massif par ce robot au moment où il était encore admis[41]). Un résumé intéressant des limites du protocole `robots.txt` par les créateurs du protocole TDMRep peut se trouver en ligne (en anglais)[42].

Une autre limite importante concerne en particulier les acteurs hybrides qui opèrent des robots avec des finalités multiples (notamment Google), et dont l'acceptation des directives ne peut être contrôlée facilement. Elle est détaillée dans la section sur les enjeux posés par ces acteurs plus haut.

[41] Cf. par exemple cet article de presse : https://www.404media.co/websites-are-blocking-the-wrong-ai-scrapers-because-ai-companies-keep-making-new-ones/.

[42] https://github.com/w3c/tdm-reservation-protocol/blob/main/docs/robots.md#conclusion

Le risque d'un Internet qui se referme

Pour les raisons évoquées précédemment, les protocoles permettant de réserver des contenus sont parfois perçus comme insuffisamment efficaces ou transparents par les éditeurs de contenus. Ainsi, certains adoptent des mesures supplémentaires pour protéger leurs sites de ce qu'ils perçoivent comme une aspiration indue de contenus par des robots qui ne respecteraient pas les protocoles précédents :

- Les fichiers de configuration des serveurs Web (e.g. : `.htaccess` pour Apache) fournissent des méthodes pour bloquer de manière stricte les requêtes associées à certains *User-Agents ou certaines plages d'IPs ;*
- Les pare-feux d'applications web (ou WAF pour *Web Application Firewalls*) sont des filtres protégeant des applications Web. Ces pare-feux servent en particulier à prévenir d'attaques classiques comme les attaques DDoS ou d'injection SQL grâce à la mise en place de politiques (c'est-à-dire de règles) d'identification des requêtes à filtrer. Des politiques spécifiques peuvent naturellement être écrites pour filtrer les requêtes provenant de robots d'exploration.

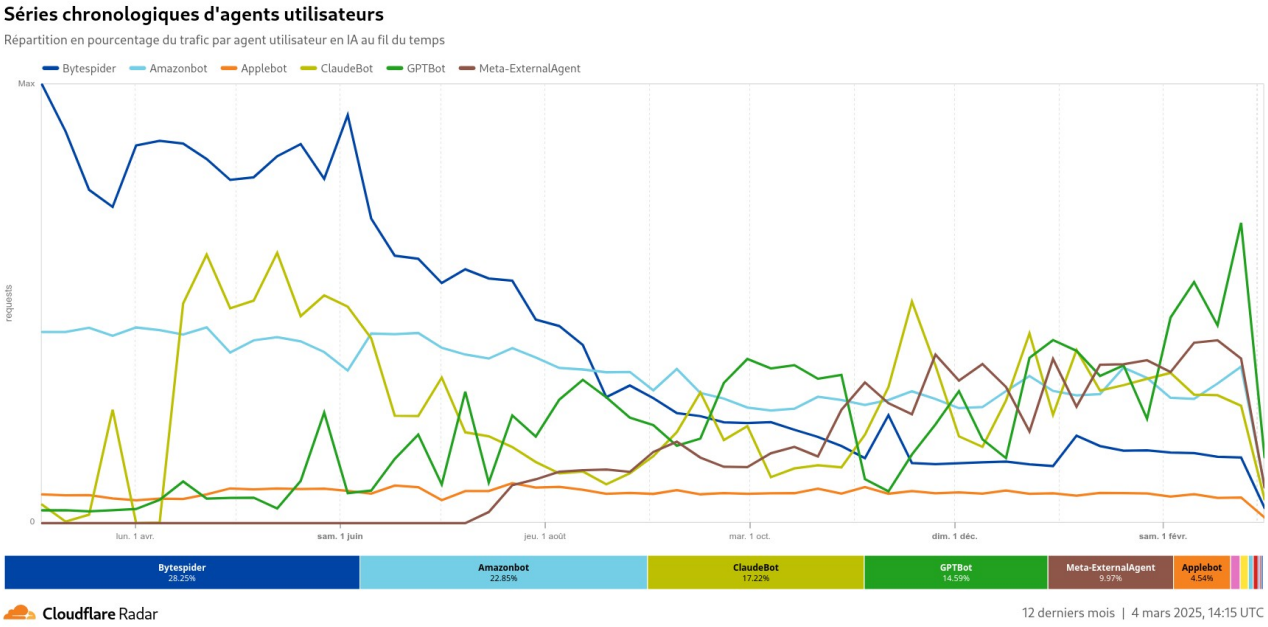

Figure 1: Evolution comparée du nombre de requêtes (unités arbitraires) effectuées par les principaux robots IA (France en haut, monde en bas) sur une année glissante jusqu'en mars 2025 ; le cartouche sous les figures présente la proportion moyenne des différents robots au sein de la catégorie des robots liés à l'IA. Sur la figure du haut, on aperçoit nettement la périodicité des moissonnages pour le robot CCBot *du* CommonCrawl *au niveau français. (source Cloudflare Radar)*

Les méthodes décrites ci-dessus demandent l'établissement de règles explicites portant sur les *User-Agents* ou des adresses *IP*. Cependant l'identification de bots est une tâche difficile[43], et l'écriture manuelle de règles ne suffit en général pas à bloquer toutes les requêtes jugées

[43] Cf. https://datadome.co/fr/bot-management-protection-fr/waf-vs-protection-robot/ par exemple.

indésirables. Plusieurs solutions techniques sont utilisées pour imposer des blocages supplémentaires :

- Les réseaux de diffusion de contenu (ou CDN pour *Content Delivery Networks,* comme ceux des entreprises *Cloudflare ou Akamai*) sont des points d'entrée externes qui permettent à leurs clients de rendre leurs contenus accessibles rapidement depuis un grand nombre de localisations. Les entreprises opérant ces réseaux proposent en général des solutions de filtrage de trafic, à partir d'adresses IP par exemple, ou de règles éventuellement plus complexes[44]. L'analyse de l'intégralité du trafic du réseau peut en effet permettre d'identifier des adresses IP présentant un comportement inhabituel et donc une probabilité importante de se trouver derrière des robots ;
- Les CAPTCHA sont des dispositifs d'authentification par question-réponse qui permettent de vérifier si l'utilisateur qui tente d'accéder au contenu est un humain et non un robot[45] ;
- Les verrous d'accès, payants (*paywalls*) ou non (navigation avec un compte utilisateur) demandent à un utilisateur de s'identifier, voire de souscrire un abonnement s'il souhaite accéder au contenu d'une page web. Si ces dispositifs sont utilisés par les éditeurs de presse pour s'assurer des revenus complémentaires aux revenus d'abonnement des éditions papier, ils peuvent constituer une protection contre les robots. Cette protection restera néanmoins limitée contre des acteurs malveillants, car il leur suffirait d'être titulaire d'un compte ou d'un abonnement pour obtenir l'accès au contenu pour son robot.

Ces dernières mesures offrent une protection des contenus au prix de contrôles pour l'accès de ressources en ligne. Si elles se généralisent et deviennent trop étendues ou mal calibrées, elles pourraient compromettre les principes d'ouverture d'internet.

Des pistes de bonnes pratiques à explorer…

La multiplication des systèmes d'intelligence artificielle générative peut parfois poser des questions sur l'effectivité des droits dont disposent les éditeurs de presse sur leur contenu publié en ligne, lorsque ce dernier est utilisé pour le développement ou le fonctionnement de tels systèmes.

Plusieurs protocoles existent pour permettre aux éditeurs d'exercer ces droits, dont le protocole `robots.txt` qui est le plus largement diffusé aujourd'hui même s'il peut parfois être absent ou mal configuré. La plupart des acteurs opérant ces robots considèrent respecter ce protocole somme toute peu contraignant. En particulier, il repose sur l'*immatriculation* volontaire des robots, laissant de fait une importante liberté aux moissonneurs de données (utilisation de plusieurs robots ou un unique robot pour plusieurs finalités, renouvellement régulier des *immatriculations* des robots). Cette asymétrie par nature vis à vis des éditeurs de site peut opacifier les pratiques de *crawling*. Des protocoles émergents cherchent à pallier ces défauts, mais ils sont encore très faiblement

[44] Pour Cloudflare, cf. https://blog.cloudflare.com/cloudflare-ai-audit-control-ai-content-crawlers/#step-3-control-the-bots-you-do-want-to-allow.

[45] Les bots CAPTCHA sont des robots qui résolvent des CAPTCHA (éventuellement avec une aide humaine) et sont utilisés par des *crawlers* qui souhaitent contourner cette mesure de protection. Certaines sociétés proposent des mesures de protection spécifiques contre ces CAPTCHA bots.

adoptés par les éditeurs de sites, et encore moins par les moissonneurs de données.

Tous les protocoles cités supposent que le robot se déclare correctement auprès du site qu'il visite. Dès lors, pour faire face à un risque de potentiels robots non identifiés préalablement ou déclarés incorrectement, certains ayants-droit choisissent de bloquer en amont l'accès à leur contenu à ces robots, via des dispositifs de pare-feu, de CAPTCHA, ou grâce aux services d'acteurs tiers comme les opérateurs de CDN. Les mois et années à venir pourraient s'avérer déterminants dans l'élaboration de modèles de valorisation des données utilisées par les systèmes d'IA. En attendant, il est important que certaines bonnes pratiques avec les technologies existantes se généralisent aussi bien du côté des éditeurs que des moissonneurs de données pour prévenir les restrictions d'accès généralisées aux contenus sur Internet.

Du côté des éditeurs :

- assurer un travail de veille potentiellement mutualisée pour avoir une vision en temps réel du paysage des robots d'exploration ;
- systématiser le recours au standard du `robots.txt` ;
- faire évoluer le standard `robots.txt` pour y ajouter les finalités autorisées et/ou non autorisées d'utilisation des données crawlées. Cette évolution permettrait par exemple de réserver des droits pour des robots inconnus des éditeurs de contenu, ce qui est impossible à l'heure actuelle[46] ;
- exploiter leurs registres de visites pour vérifier le respect des règles édictées vis-à-vis des robots d'exploration, et dans un format qui permette le rapprochement avec les registres tenus par les *crawlers*.

Du côté des moissonneurs de données :

- tenir à jour une documentation publique sur les différents robots opérés et les finalités de collecte associées ;
- mettre en place en interne des registres de visite (incluant les finalités de collecte), conservés sur plusieurs années et requêtables dans un format standardisé ;
- rendre ces registres accessibles aux éditeurs des sites visités, de manière à permettre une comparaison avec leurs propres registres de visites.

[46] Une première étape préalable à cette mesure serait que l'ensemble de l'écosystème s'accorde sur une taxonomie, exhaustive, de finalités de collecte.