RFC 9969 : IAB AI-CONTROL Workshop Report

Stéphane Bortzmeyer
<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 20 mai 2026

Date de publication du RFC : Mai 2026

https://www.bortzmeyer.org/9969.html

Ah, l'IA...Vaste sujet, et d'actualité. Une des questions qui reviennent souvent est celle de l'utilisation du contenu qu'on trouve sur le Web pour entrainer les grands modèles, sa légitimité, la charge qu'elle induit pour les serveurs, les moyens de la contrôler, etc. Un colloque avait été organisé <https://www.ietf.org/blog/impressions-ai-control-workshop/> par l'IAB en septembre 2024 sur ces questions et ce RFC en est le compte-rendu. Ce colloque avait lancé le projet IETF aipref <https://datatracker.ietf.org/wg/aipref/>.

Une petite précision politique d'abord : le RFC précise bien qu'il s'agit d'un compte-rendu et que l'IAB n'approuve pas forcément tout ce qui a été dit à ce colloque (section 1.2 du RFC). Je rajoute que j'ai aussi des opinions sur le sujet, donc je mettrais [entre crochets] ce qui est mon opinion, et ne vient pas du RFC. Le reste n'est donc pas de moi, j'en rends compte, c'est tout, ne me tapez pas.

Ce colloque <https://www.ietf.org/blog/impressions-ai-control-workshop/> fait partie de la série de colloques qu'organise régulièrement l'IAB pour explorer des tendances à plus ou moins long terme, sans les obligations de l'IETF de produire normes et documents.

Donc, les LLM (qui ne sont qu'une partie des techniques qu'on regroupe sous le terme marketing d'« IA » <https://www.bortzmeyer.org/pses-ia.html>) fonctionnent en deux phases : on entraine le modèle en lui faisant ingérer une grande quantité de contenu (textes, images ou autres), puis il va pouvoir inférer du contenu à partir de ce qu'il a digéré pendant la phase d'entrainement, et d'une demande (dite "*prompt*"). Le contenu inféré n'est jamais tout à fait identique à celui utilisé pour l'entrainement (autrement, cela serait du plagiat, potentiellement illégal). Les LLM ne marchant bien, à l'heure actuelle, que si le corpus d'entrainement était énorme, ils sont très gourmands en données et une source évidente de contenu en grande quantité est le Web. Des "*bots*" ou "*crawlers*" parcourent donc le Web, ramassant du contenu. Voici par exemple un extrait du journal du serveur qui héberge ce blog, montrant le "*bot*" de Perplexity récoltant du contenu :

```
18.210.92.235:64884 - - [09/Jan/2026:07:26:15 +0000] "GET /images/TCP_state_diagram.jpg HTTP/1.1" 200 96551
18.97.9.103:64398 - - [09/Jan/2026:08:31:39 +0000] "GET /bitcoin-metamorphoses.html HTTP/1.1" 200 8982 "-" "
18.97.9.101:57493 - - [09/Jan/2026:08:31:39 +0000] "GET /robots.txt HTTP/1.1" 404 3250 "-" "Mozilla/5.0 Appl
18.97.9.103:48122 - - [09/Jan/2026:08:47:36 +0000] "GET /nist-pq.pdf HTTP/1.1" 200 84543 "-" "Mozilla/5.0 Ap
18.97.9.96:25157 - - [09/Jan/2026:08:51:37 +0000] "GET /files/capitole-libre-2019-quic-pour-impression.pdf H
```

Une des questions soulevées par cette récolte de données est qu'elle n'était pas prévue à l'origine. Certains webmestres qui mettent du contenu en ligne estiment donc qu'un nouvel usage (l'entrainement des LLM) justifie de nouvelles règles et de nouvelles possibilités de contrôle par le serveur Web. C'est par exemple ce que prévoit l'"*AI Act*" européen.

Le colloque (ou atelier) de l'IAB s'est tenu les 19 et 20 septembre 2024 (oui, le RFC met trop longtemps à sortir) et prévoyait de travailler sur tous les aspects liés à cette récolte de données (cf. l'appel à participation <https://datatracker.ietf.org/group/aicontrolws/about/>). Le colloque regroupait des personnes de divers horizons, experts techniques, entreprises d'IA, fournisseurs de contenu, décideurs politiques, etc. La liste figure dans l'annexe A.2. La règle suivie était celle de Chatham House (tout ce qui se dit est public mais sans le lier à un·e participant·e particulier·ère). D'ailleurs, le RFC note qu'au moins un participant n'a pas voulu que son identité soit dévoilée, juste qu'il était un représentant officiel d'un gouvernement. Et, comme déjà dit, le RFC rend compte des discussions, cela ne signifie pas que l'IAB approuve tout ce qui a été dit.

La section 2 du RFC résume les discussions (la totalité des soumissions sont en ligne <https://datatracker.ietf.org/group/aicontrolws/materials/>). Aujourd'hui, les fournisseurs de contenu, les webmestres, peuvent exprimer leurs choix quant à la récolte de données par divers moyens. Il y a par exemple une solution technique existante, c'est le `robots.txt`, qui est décrit dans le RFC 9309 [1]. Notez que, en l'absence du fichier `robots.txt`, les "*bots*" peuvent tout récolter. C'est donc une solution "*opt-out*". Il y a aussi des solutions non-techniques par exemple les conditions d'utilisation (ainsi, ce blog </> est sous licence GFDL et les contenus peuvent donc être réutilisés, à la condition que les destinataires jouissent des mêmes droits de réutilisation). Tiens, par contre, il n'y a pas de licence CC-BY-NoAI ? Pour revenir à la technique, le webmestres peuvent aussi bloquer les "*crawlers*" par leur adresse IP, ou leur `User-Agent` (RFC 9110, section 10.1.5), ou carrément tout mettre derrière un "*paywall*".

Comme indiqué plus haut, ces solutions sont en général "*opt-out*", donc par défaut, la récolte est autorisée. (Sur ce blog, et c'est apparemment le cas de la plupart des serveurs HTTP, plus de la moitié des requêtes sont faites par des "*bots*", mais pas forcément liés à l'IA, c'était déjà le cas avant les LLM.) On constate (cf. l'exposé « "*Consent in Crisis : The Rapid Decline of the AI Data Commons*" <https://www.ietf.org/slides/slides-aicontrolws-consent-in-crisis-the-rapid-decline-of-the-ai-data pdf> ») une tendance à la fermeture : la récolte devient de plus en plus difficile car de nombreux serveurs bloquent les accès qu'ils pensent dûs à un "*bot*". Le Web tend donc à se fermer.

[Le RFC n'est pas clair sur ce point mais, pour moi, il est important de faire la différence entre les problèmes techniques et opérationnels posés par certains "*bots*" qui, qu'ils travaillent pour l'IA ou pas, « matraquent » avec excès les serveurs, et les problèmes politiques et financiers liés à l'**utilisation** qui est faite des données récoltées. Les problèmes techniques et opérationnels causés par des « "*bots*" fous » existaient bien avant les LLM. Par contre, les problèmes politiques (légitimité à réutiliser le contenu <https://www.bortzmeyer.org/collecte-pour-l-ia.html>) et financiers (perte de revenus pour les ayant-droits, comme mentionné dans le RFC) sont plus spécifiques de l'IA. Le RFC ne

1. Pour voir le RFC de numéro NNN, `https://www.ietf.org/rfc/rfcNNN.txt`, par exemple `https://www.ietf.org/rfc/rfc9309.txt`

parle pas vraiment des problèmes opérationnels posés par l'agressivité de certains "*bots*" - pas forcément liés à l'IA, d'ailleurs - mais la réunion IETF 123 à Madrid <https://datatracker.ietf.org/meeting/123/proceedings> avait vu de très intéressants exposés à ce sujet <https://mastodon.gougere.fr/@bortzmeyer/114901375864309123>.]

À l'heure actuelle, il est difficile de savoir ce qui est permis, au delà des simples consignes du `robots.txt`. Les gérants des serveurs n'ont pas de moyen standard et automatiquement analysable de faire connaitre leurs conditions d'utilisation, et les "*bots*" n'ont donc pas non plus de moyen de savoir ce qui est permis. [Il va de soi qu'il y a des "*bots*" qui, de toute façon, s'en foutent. Le travail de normalisation à l'IETF ne pourra concerner que les "*bots*" honnêtes et ne dispensera pas de mesurs de sécurité contre les malhonnêtes.]

Le RFC creuse certain aspects de la question. Par exemple, en section 2.1, le problème de la différence entre le moment du ramassage des données et celui de leur utilisation. Les consignes du serveur (comme le `robots.txt`) sont lues au moment du ramassage mais certains responsables de contenu voudraient exprimer des choix concernant l'utilisation, or celle-ci se fait à un autre moment, décorrélé du premier. Certaines récoltes, comme celle faite par Common Crawl, peuvent servir à de multiples usages et des consignes concernant le ramassage ne sont donc pas appropriées. Autre exemple que Common Crawl, on peut avoir une organisation qui gère un moteur de recherche du Web **et** développe un LLM, et qui utilise le même "*crawler*" pour les deux usages. Certains webmestres estiment que la première utilisation ne pose pas de problème (au bout du compte, cela ramènera du trafic sur leur site Web) mais s'opposent à la seconde car elle n'apportera pas de trafic, le LLM donnant des réponses qui suffiront à l'utilisateur.

Du point de vue technique, il faut aussi noter que le principe d'entrainement d'un LLM fait qu'on utilise toutes les données et, qu'une fois le modèle créé, il n'y a pas d'étiquetage spécifique de la source de telle ou telle réponse du LLM. (C'est pour cela que les LLM ont du mal à indiquer leurs sources.) Un webmestre qui souhaiterait dire « d'accord pour servir à l'entrainement des IA mais pas pour que ces IA aient un usage militaire, ou bien pas un usage commercial » ne le peut pas, en raison de cette limite technique. Et même si ce moyen existait, le gérant du LLM serat obligé de faire N modèles, pour toutes les permutations des différents critères (ou tout simplement d'exclure tous les contenus ayant une licence restrictive, ce qui limiterait la représentativité du corpus d'entrainement du modèle).

Enfin, les préférences changent dans le temps et celles exprimées au moment de la récolte des données peuvent ne pas être à jour lorsque les données seront utilisées pour l'entrainement d'un LLM.

Le problème est déjà compliqué si on suppose que tous les acteurs sont de bonne foi et respectent les règles. Mais, évidemment, la confiance ne règne pas, et pour de bonnes raisons. [Les entreprises capitalistes trichent, que ce soit celles qui entrainent les LLM ou bien celles des ayant-droits.] Il n'y a pas de moyen facile de vérifier le respect des préférences exprimées par les gérants du contenu. Et c'est d'autant plus inquiétant que les entreprises de l'IA n'ont pas vraiment de motivation pour respecter les règles : aucun risque de sanction [surtout compte-tenu des déclarations de Trump contre tout projet de régulation de l'IA]. Cette absence de confiance entraine l'utilisation importante de moyens techniques de blocage, comme de bloquer les adresses IP des "*bots*" connus. Il y a même un "*bot*" qui suggère cette solution :

```
119.28.89.249:58834 - - [21/Jan/2026:15:32:02 +0000] "GET /5153.xml HTTP/1.1" 200 6891 "-" "Mozilla/5.0 (compatil
```

L'atelier de l'IAB a passé du temps sur la question de l'attachement des préférences au contenu (section 2.3). Le `robots.txt` (RFC 9309) est très bien, très déployé et largement reconnu. Mais il manque de souplesse pour les gros sites qui souhaitent un système plus granulaire. Par exemple, si un site de vidéos

souhaitait restreindre l'accès à certaines vidéos, la seule solution est de les placer dans un espace particulier (par exemple un répertoire distinct), et donc de devoir changer l'URL si la classification change. Et, comme le `robots.txt` est à la racine du site Web, il n'est pas sous le contrôle des créateurs de contenu qui ont accès à un espace dédié mais pas à la totalité du site. Si le CMS que vous utilisez permet des créations de contenu et des mises à jour décentralisées, où certaines personnes peuvent modifier une partie du site, regardez s'il permet à ces personnes d'influencer le `robots.txt`. Je suis preneur d'exemples.

Une autre solution (qui ne serait pas forcément exclusive du `robots.txt` mais complémentaire) serait d'inclure les préférences d'utilisation dans le contenu lui-même. C'est ce que permet l'élément HTML `<meta>` ou le format XMP pour les images. Des formats comme XML ou JSON permettraient certainement d'ajouter ces préférences d'utilisation, qui ont l'avantage de forcément voyager avec le contenu, contrairement au `robots.txt`. Évidemment, cela ne marchera pas si ces méta-données sont retirées par le programme de collecte (pas forcément pour des raisons malveillantes, cela peut être pour diminuer la taille des données). Et certains formats ne se prêtent pas à cette inclusion des préférences d'utilisation, comme le texte brut, ou comme les contenus qui ont plusieurs auteurs (pensez à un fil de discussion sur les réseaux sociaux).

Une autre solution serait de placer les préférences d'utilisation dans un registre, extérieur aux œuvres, comme cela se fait souvent pour, par exemple, la musique ou les photographies. C'est plus robuste que l'inclusion de métadonnées mais ça passe mal à l'échelle de l'Internet (les registres existants avaient été conçus pour des écosystèmes plus petits et relativement fermés).

Enfin, parmi les difficultés, il faut noter qu'exprimer préférences et conditions d'utilisation un peu fines nécessite de disposer d'un vocabulaire (par exemple pour décrire les différentes techniques qui sont regroupées sous le terme marketing et flou d'« IA ») et qu'il n'existe pas de vocabulaire standard. Ce serait un tâche difficile que d'en établir un (un travail est en cours, dans `draft-ietf-aipref-vocab`). Je me souviens d'une réunion IETF où il y avait eu un long débat sur la question de savoir si la traduction rentrait dans la catégorie « IA générative » (après tout, elle génère des textes…).

La section 3 du RFC, en conclusion, essaie de synthétiser et d'identifier les points sur lesquels l'IETF pourrait travailler. L'atelier avait un relatif consensus sur le fait que la situation actuelle est mauvaise et que le principal outil technique disponible, `robots.txt`, ne convient pas. Les pistes de travail discutées ont été :

- Améliorer `robots.txt` ou bien développer un meilleur système d'attachement aux sites Web,
- Définir des attachements pour les protocoles IETF (par exemple dans l'en-tête HTTP, HTML ou XML dépendant d'un autre organisme),
- Définition d'un vocabulaire commun (le groupe de travail IETF aipref `<https://datatracker.ietf.org/wg/aipref/>` y travaille),
- Description de comment les différentes techniques (attachées au site Web ou bien attachées au contenu) se combinent.

Par contre, le consensus était que les points suivants n'étaient pas du ressort de l'IETF ou ne pouvaient pas, pour l'instant, faire l'objet d'un travail concret :

- Faire respecter les directives données aux récoltants (les licences du logiciel libre ont un problème analogue),
- Développer des mécanismes de transparence de la récolte,
- Créer des registres des contenus,
- Identifier et authentifier les "*bots*" (à noter que Cloudflare a lancé cette idée `<https://blog.cloudflare.com/web-bot-auth/>`, ce qui a donné naissance aux brouillons `draft-meunier-web-bot-` et `draft-meunier-web-bot-auth-glossary`).

Notez qu'un résumé de l'atelier avait été publié juste après `<https://www.ietf.org/blog/impressions-ai-control-workshop/>`. Et, sinon, vous pouvez regarder l'intéressant site Web « "*Dealing With Bots*" `<https://dealing-with-bots.coar-repositories.org/>` » et, sur les projets de contrôle de l'accès aux ressources et leurs risques, l'excellent article « "*No One Should Control the Internet After AI : Freedom to Build Cleopatra GPT*" `<https://digitalmedusa.org/no-one-should-control>` ».