Engineer

robots.txt

Fichier signalant aux robots les chemins à ne pas explorer

Ce que c’est

Chaque hôte du web peut publier, au chemin fixe `/robots.txt`, un fichier texte de demandes adressées aux programmes qui l’explorent. Martijn Koster a défini cette convention en 1994. En septembre 2022, l’IETF l’a publiée sous la forme de la RFC 9309, le protocole d’exclusion des robots, en précisant le langage des règles et les instructions relatives aux erreurs et au cache.

Le fichier est une liste de groupes. Un groupe commence par une ou plusieurs lignes `user-agent`, chacune nommant un robot par son identifiant de produit, comme `ExampleBot`, ou par `*` pour tous les robots. Suivent les lignes `allow` et `disallow`, contenant chacune un chemin. Un robot cherche le groupe portant son identifiant, sans tenir compte de la casse, et lui obéit. Si plusieurs groupes le nomment, leurs règles fusionnent. Seul un robot sans groupe propre se rabat sur le groupe `*`, et celui qui ne trouve ni l’un ni l’autre n’a aucune règle à suivre.

Dans un groupe, une règle s’applique lorsque son chemin correspond au début de l’adresse. La correspondance la plus précise l’emporte, c’est-à-dire celle qui comporte le plus d’octets. Une égalité entre `allow` et `disallow` se tranche en faveur de `allow`, et une adresse sans règle correspondante est autorisée. Deux caractères élargissent un chemin : `*` représente une suite quelconque de caractères et `$` fixe la fin de l’adresse. Un `#` commence un commentaire, et le chemin du fichier lui-même est toujours autorisé.

La RFC précise aussi la récupération du fichier. Il se trouve à ce chemin en minuscules à la racine du service. Chaque hôte répond donc pour lui-même, et un fichier sur un sous-domaine ne dit rien d’un autre. C’est du texte brut UTF-8. Un robot devrait suivre au moins cinq redirections pour l’atteindre, ne devrait pas utiliser une copie en cache pendant plus de 24 heures et peut plafonner la quantité analysée, mais jamais sous 500 KiB.

Les lignes extérieures au protocole sont tolérées. La RFC les appelle autres enregistrements, donne `Sitemap` comme exemple et laisse le robot les interpréter ou non, à une condition : une telle ligne ne doit jamais terminer un groupe.

Utilisez le concept pour prendre une décision

Utilisez « Qu'est-ce que robots.txt ? » pour éclairer une décision réelle. Nous pouvons vous aider à relier ce concept au système que vous construisez ou achetez. Un site ou une application que les assistants d'IA de vos clients peuvent trouver, lire et utiliser : une capacité vérifiée à chaque livraison, plutôt qu'une promesse ponctuelle.

Découvrir nos prestations Contactez-nous →

Vérifiez les hypothèses derrière le terme

Appuyez-vous sur l'explication et les réalisations liées pour formuler les questions de votre projet. Nous pouvons examiner les hypothèses avec vous avant le choix d'un outil. Site Hugo multilingue en 3 langues — 108 fichiers de modèles pour trois langues et quatre formats

Site Hugo multilingue en 3 langues Contactez-nous →

Pourquoi c'est important

La RFC pose la limite dans son introduction : « Ces règles ne constituent pas une forme d'autorisation d'accès. » Le fichier est une demande. Un robot qui la respecte reste à l'écart ; celui qui ne la respecte pas lit tout ce que lui fournit le serveur. La section sur la sécurité ajoute qu'inscrire un chemin dans le fichier le rend public et découvrable. Tout ce qui doit rester privé nécessite donc un véritable contrôle, comme l'authentification HTTP.

La documentation Google restreint encore l'objectif. Elle décrit le fichier comme « principalement utilisé pour éviter de surcharger votre site de requêtes » et indique que « ce n'est pas un mécanisme permettant d'exclure une page web de Google ». Une adresse interdite peut tout de même être indexée si d'autres pages pointent vers elle. Elle apparaît alors dans les résultats sans description, parce que le robot n'a jamais lu la page. Pour l'exclure des résultats, Google renvoie à une règle `noindex` ou à un mot de passe. Une règle `noindex` ne fonctionne que sur une page que le robot peut récupérer.

Le fichier remplit bien sa fonction d'origine : détourner les robots des parties d'un site qui coûteraient des requêtes au serveur sans apporter grand-chose au moteur de recherche, comme les pages peu importantes ou presque identiques. Sa portée est aussi son risque. Une ligne sous `*` s'applique à toutes les adresses et à tous les robots qui respectent le fichier. Un changement peut prendre jusqu'à un jour pour être vu : un `Disallow: /` dans le mauvais groupe ferme donc un site entier.

À quoi cela ressemble en pratique

Le fichier de ce site est long et presque entièrement constitué de commentaires. La partie sur laquelle agit un robot compte quatre lignes :

```
User-agent: *
Content-Signal: search=yes, ai-input=yes, ai-train=yes
Allow: /

Sitemap: https://engineer.company/sitemap.xml
```

Un groupe couvre tous les robots et `Allow: /` correspond à toutes les adresses. Un groupe sans règles permettrait exactement la même chose, puisqu'une adresse sans règle correspondante est autorisée. La ligne est néanmoins écrite, car cette ouverture est un choix. Le commentaire qui la précède dit « Decided rather than inherited », autrement dit une décision plutôt qu'un héritage, et la page consacrée aux besoins des agents IA sur un site présente la politique d'exploration comme l'une des décisions délibérées du site.

`Content-Signal` n'est pas un enregistrement défini par la RFC 9309. Il vient du vocabulaire Content Signals et autorise ici la collecte pour la recherche, pour répondre à une question au moment de l'inférence et pour l'entraînement. Selon la RFC, c'est un de ces autres enregistrements : un robot connaissant la convention peut le lire, un robot antérieur l'ignore, comme le dit le commentaire du fichier, et aucun ne peut le traiter comme la fin du groupe. La ligne `Sitemap:` est un autre enregistrement de ce type et pointe vers l'index des plans du site. La page des crédits résume le fichier de la même manière : il « autorise tout, aussi bien aux robots de recherche qu'aux systèmes d'IA », et elle nomme le plan du site.

Les absences du fichier ont aussi été décidées. Parmi les cinq recommandations refusées, la page sur la préparation aux agents IA cite les règles par agent. Elle conserve une règle commune à tous les agents « parce qu'une règle écrite pour un robot remplace silencieusement la règle générale pour lui ». Si le fichier avait donné son propre groupe à un robot IA, celui-ci ne lirait que ce groupe et ne recevrait jamais un `Disallow` ajouté plus tard sous `*`.

Le reste est constitué de commentaires destinés à la personne qui ouvre le fichier : les raisons qui motivent cette politique, puis les adresses de `llms.txt`, `agents.txt`, `humans.txt`, `security.txt` et d'autres fichiers lisibles par machine,

gardées en commentaires « because robots.txt has no field for them », c'est-à-dire parce que robots.txt ne leur réserve aucun champ. Les commentaires font donc partie du contenu livré. Une ligne `#` du modèle est servie octet pour octet à chaque robot. À deux reprises, l'une contenait une note destinée à un mainteneur, pointant vers un document de planification dans un dépôt non public. Un contrôle du site construit échoue désormais lorsqu'un fichier servi contient un chemin interne au dépôt. Les notes de maintenance du modèle restent dans des commentaires de modèle, que le générateur supprime.

Où les problèmes surviennent

Une règle générale cesse de s'appliquer à un robot. Quelqu'un ajoute un groupe pour un seul robot, peut-être pour l'écarter d'un dossier. Dès lors, ce robot ignore tout ce qui se trouve sous `*`, y compris chaque `Disallow`. La RFC 9309 n'envoie un robot au groupe `*` que si aucun groupe ne le nomme. La documentation Google précise que les groupes spécifiques et le groupe `*` « ne sont pas combinés ». Répétez chaque règle générale dans chaque groupe nommé ou gardez un seul groupe, comme ce site.

Une page bloquée apparaît encore dans les résultats. Son adresse apparaît sans description, éventuellement accompagnée des mots employés par d'autres sites dans leurs liens. Google n'a jamais récupéré la page : il sait seulement que l'adresse existe et comment d'autres la nomment. Un `Disallow` contrôle l'exploration, pas l'indexation. Retirez la règle et servez `noindex` sur la page, ou protégez-la par mot de passe. Un `noindex` caché derrière un `Disallow` n'est jamais lu.

Le fichier annonce ce qu'il devait cacher. Une ligne telle que `Disallow: /admin/` indique à chaque lecteur, robot poli ou non, exactement où chercher, et le fichier est public par conception. La section sécurité de la RFC précise que l'énumération des chemins les rend découvrables et demande plutôt un contrôle d'accès. Placez les contenus privés derrière une authentification et ne les inscrivez pas du tout dans le fichier.

Une erreur serveur arrête l'exploration du site entier. Les pages se chargent, mais `/robots.txt` renvoie un statut 5xx ou dépasse le délai. Un robot suivant la RFC doit traiter cela comme une interdiction totale. Google arrête l'exploration pendant les 12 premières heures puis utilise sa dernière copie valide jusqu'à 30 jours. L'erreur inverse est plus silencieuse : Google traite une réponse 4xx autre

que 429, y compris 401 et 403, comme une absence de fichier. Toutes les adresses sont donc ouvertes. Servez le fichier en texte statique avec le statut 200 et supervisez `/robots.txt` lui-même, pas seulement la page d'accueil.

Pour aller plus loin

- RFC 9309 : protocole d'exclusion des robots — la norme elle-même : groupes, correspondance, gestion des erreurs, cache et note de sécurité.
- Introduction à robots.txt — objectif donné par Google au fichier et raison pour laquelle il ne cache pas une page des résultats.
- Interprétation de la spécification robots.txt par Google — codes de statut, cache, champs ignorés par Google et choix du groupe.
- Bloquer l'indexation avec noindex — la règle qui exclut effectivement une page des résultats et la raison pour laquelle elle doit pouvoir être explorée.

Glossaire

Choisissez selon vos contraintes réelles

Présentez le processus de travail, les interfaces existantes et les utilisateurs qui ont besoin du changement. Nous pouvons définir une livraison testable et une transmission permettant à votre équipe d'assurer la maintenance des éléments livrés.

Consulter les tarifs Contactez-nous →

Rendez la terminologie concrète

Demandez comment « Qu'est-ce que robots.txt ? » s'applique à votre organisation. Un exemple concret tiré de votre système actuel suffit pour engager la discussion.

Découvrir nos prestations

welcome@engineer.company +45 50 10 61 84 Copenhague, Danemark

Emportez cette page avec vous

Fiche de contact

Créé avec soin et amour par D.

CVR 45969703 · Tagensvej 105C, 3. th, 2200 København N

Droit d'auteur © 2025 – aujourd'hui · Engineer ApS · Tous droits réservés.

Imprimé depuis https://engineer.company/fr/glossary/robots-txt/