Optimisez vos contenus

Guide pour des éditeurs de sites web

Google™

Sommaire

Introduction

Si vous cherchez à vous faire connaître, Internet est l'outil qu'il vous faut. Si vous en doutez, il vous suffit d'interroger un annonceur qui a réussi à augmenter ses ventes grâce à la publicité en ligne, un blogueur qui a décroché un contrat avec un éditeur grâce à sa popularité sur le Web ou un directeur dont le journal touche désormais un public international grâce à Internet.

Nous recevons fréquemment de nombreuses questions portant sur la manière dont fonctionnent les moteurs de recherche Web, ainsi que sur la façon dont les éditeurs Web peuvent optimiser leur présence sur Internet.

Ce petit guide vous aidera à mieux comprendre comment les moteurs de recherche "perçoivent" votre contenu. Vous apprendrez à adapter ce dernier pour que les internautes trouvent facilement les informations que vous désirez communiquer, sans qu'ils puissent toutefois accéder à celles que vous ne souhaitez pas diffuser.

Ce petit guide, qui contient des conseils destinés aux administrateurs de sites Web, des informations sur des outils en ligne, ainsi qu'une foire aux questions étape par étape, est conçu aussi bien pour les petits éditeurs de sites Web que pour les propriétaires de grands sites.

A l'instar d'Internet qui a connu une évolution fulgurante au cours de la dernière décennie, l'approche de Google en matière de recherche Web et ses relations avec les propriétaires de sites Web ont connu d'importantes transformations. Nous avons mis au point de nombreux outils qui permettent aux administrateurs de sites Web d'optimiser la visibilité de leur contenu et de mieux contrôler les modalités d'indexation de leurs pages Web. Mais le mieux n'est pas toujours l'ennemi du bien. N'hésitez pas à nous faire part de vos commentaires sur ce guide et de vos idées et suggestions d'amélioration. Nous mettons tout en oeuvre pour que le Web devienne un outil encore plus convivial, aussi bien pour les internautes que pour les éditeurs de sites Web.

- L'équipe Google Webmaster

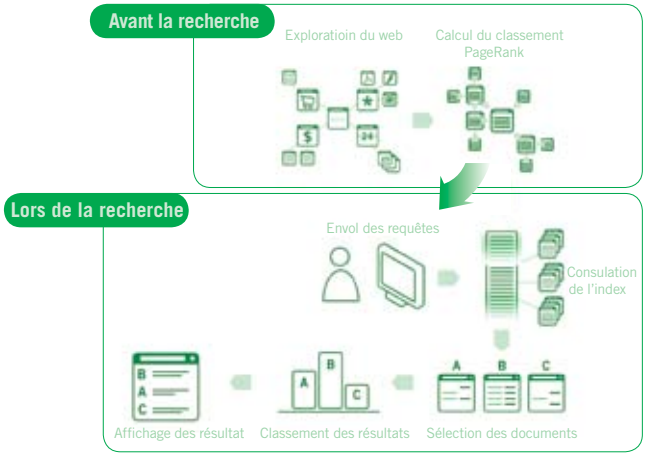
Rapide présentation de la recherche Web : *fonctionnement*

En termes simples, on peut dire que le Web s'apparente à un gigantesque livre dont l'index tout aussi imposant vous indique avec précision où trouver ce que vous cherchez.

Google dispose d'un groupe d'ordinateurs nommé Googlebot qui explore en permanence des milliards de pages Web. Ce processus d'exploration (le crawling, en anglais) est algorithmique. Cela signifie que des programmes informatiques déterminent quels sites explorer, à quelle fréquence et combien de pages récupérer pour chacun de ces sites. Google n'accepte pas de paiement pour qu'un site soit exploré plus fréquemment, et les activités liées au moteur de recherche sont totalement séparées de notre service AdWords, qui lui, est source de revenus.Il existe cependant des moyens gratuits et simples pour optimiser la fréquence de référencement de son site, notamment le protocole Sitemaps (voir page 14).

Notre groupe d'ordinateurs commence son processus d'exploration en parcourant une liste d'adresses URL de pages Web. Cette exploration permet aux robots Googlebot de détecter sur chacune de ces pages d'éventuels liens qu'ils ajoutent ensuite à la liste des pages à parcourir. Les robots Googlebot font également une copie de chaque page explorée pour compiler ensuite tous les mots rencontrés et créer ainsi un large index. Cette liste indique également l'emplacement exact de chaque mot sur chaque page.

Lorsqu'un utilisateur saisit une requête, nos moteurs de recherche parcourent cet index afin de trouver les pages correspondantes, puis affichent les résultats les plus pertinents. Le taux de pertinence est déterminé selon plus de 200 critères, parmi lesquels le score PageRank de chacune des pages. Ce dernier évalue l'importance d'une page en fonction des liens sur les autres pages Web renvoyant à ladite page. En d'autres termes, le score PageRank d'une page augmente à chaque fois qu'une autre page renvoie à celle-ci.



Nouveautés Google en matière de recherche Web

Bien qu'en matière de recherche Web les grands principes demeurent, Google cherche en permanence à améliorer ses résultats de recherche.

Quelle différence existe-t-il par rapport à la recherche Web, telle qu'elle était pratiquée il y a cinq ans ? Tout d'abord, elle est bien plus rapide.

En outre, nos systèmes d'exploration et d'indexation sont bien plus intelligents qu'auparavant. Nos robots parcourent désormais en continu les pages Web et planifient leur visite de manière plus efficace afin d'optimiser l'actualisation des résultats. Cette nouvelle approche prend en compte le fait qu'un journal en ligne nécessite des visites plus fréquentes qu'un site Web statique dont le contenu n'est mis à jour qu'une fois par mois par exemple. En fait, grâce aux outils disponibles sdans le Centre pour les webmasters, les administrateurs de sites Web peuvent à présent décider de la fréquence d'exploration de leurs sites par nos robots. Dans l'ensemble, ces améliorations permettent la création d'un index mieux actualisé et exhaustif.

Si aujourd'hui la recherche Web est plus rapide et plus efficace que jamais, les facteurs jouant un rôle déterminant dans la visibilité des sites Web ont toujours été notre priorité et ce dès le lancement de notre moteur de recherch:

Google peut-il trouver votre site?

Le référencement de votre site Web dans les résultats d'une recherche Google est gratuit et ne nécessite pas d'action préalable de votre part. En fait, la grande majorité des sites répertoriés dans nos résultats ne sont pas soumis manuellement, mais trouvés et ajoutés automatiquement par les robots Googlebot qui explorent le Web.

Bien que Google explore des milliards de pages, l'omission de certains sites reste inévitable. Ce genre d'omission se produit le plus souvent pour l'une des raisons suivantes :

- Les sites omis sont peu liés aux autres sites du Web en raison d'une insuffisance de liens ;
- les sites omis ont été mis en ligne après la toute dernière exploration effectuée par les robots Googlebot ;
- les sites omis étaient momentanément indisponibles lors de l'exploration ou nous avons reçu un message d'erreur tandis que nous tentions de les explorer.

Les outils Google destinés aux administrateurs de sites Web, tels que le protocole Sitemaps par exemple, peuvent vous aider à savoir si votre site est actuellement référencé dans l'index de Google ou si nous recevons des messages d'erreur lorsque nous tentons de l'explorer (voir page 14). Vous pouvez également utiliser ces outils pour ajouter manuellement l'adresse URL de votre site à l'index de Google ou nous fournir un plan Sitemap de votre site afin que nous disposions d'un meilleur aperçu de son contenu. Ce plan nous aidera à extraire le nouveau contenu et les nouvelles sections de votre site.

Google peut-il indexer votre site?

De temps à autre, les administrateurs de sites Web s'aperçoivent que leurs sites n'apparaissent pas dans nos résultats de recherche. Il peut s'agir d'un problème d'"indexabilité". Ce terme désigne la possibilité pour les robots Googlebot de faire ou non une copie des pages Web en question afin de l'inclure dans nos résultats de recherche.

Structure et Contenu

L'impossibilité d'inclure des pages Web dans nos résultats de recherche est souvent due à leur structure et leur contenu. Par exemple, une page Web sur laquelle les utilisateurs doivent renseigner les champs d'un formulaire peut ne pas être indexable par Google. De même, les moteurs de recherche peuvent avoir des difficultés à indexer une page contenant des données dynamiques (Flash, JavaScript, cadres et adresses URL générées dynamiquement). Pour vous assurer que votre site ne connaît pas ce problème, essayez de l'afficher à l'aide d'un navigateur texte tel que Lynx, ou de tout autre navigateur après avoir désactivé l'option permettant d'afficher les images, les contenus Javascript et Flash. Vous verrez alors si tout le contenu de votre site est effectivement accessible.

Si votre site contient un grand nombre d'images, assurez-vous que le texte ou les légendes y faisant référence décrivent de manière précise leur contenu. Cela permet non seulement aux moteurs de recherche d'indexer correctement vos images, mais aussi de les rendre accessibles aux internautes malvoyants. Vous pouvez également utiliser la fonction "alt text" pour vos images et attribuer à leur fichier des noms descriptifs comme dans l'exemple ci-dessous (il s'agit d'une image correspondant au logo d'une société dénommée La cuisine de Véronique) :

```
<img src=""cuisineveronique.jpg" alt="Bienvenue
sur le site consacrée à la cuisine de Véronique !">
```

Les adresses URL

Un autre obstacle à l'indexation de votre site peut être son adresse URL. Si l'adresse URL de votre site contient plusieurs paramètres ou comprend des identifiants de session ou si cette adresse renvoie automatiquement à plusieurs autres adresses successives, Google peut ne pas être mesure de l'indexer.

Serveur et réseau

Des problèmes de serveur ou de réseau peuvent également nous empêcher d'accéder à certaines pages de votre site. Grâce aux outils du Centre pour les webmasters développés par Google, les éditeurs de sites Web peuvent désormais voir une liste des pages Web auxquelles les robots Googlebot ne peuvent pas accéder. Pour en savoir plus sur les outils du Centre pour les webmasters, voir page 13.

Protocole d'exclusion des robots

Il peut arriver que certaines pages soient bloquées par le protocole d'exclusion des robots. Il s'agit d'une norme technique qui permet aux éditeurs Web d'indiquer aux moteurs de recherche de ne pas indexer le contenu de leur site (voir ci-dessous). Si votre site Web

n'apparaît pas dans les résultats de recherche Google, assurez-vous que les données du fichier robots.txt ou qu'une balise Méta ne bloquent pas l'accès de votre contenu à nos robots d'exploration.

Contrôle des éléments indexés par Google

Chaque éditeur Web cherche à atteindre un objectif différent sur Internet. Certains éditeurs de journaux choisissent par exemple de permettre à leurs lecteurs d'accéder gratuitement à leurs articles les plus récents, mais de rendre payant l'accès à leurs archives. Certains souhaitent que leur site apparaisse dans toutes les catégories d'un moteur de recherche (par exemple sur Google Mobile, Google Images, etc.), tandis que d'autres préfèrent qu'il figure uniquement dans les résultats de recherche Web.

Il est important que les moteurs de recherche respectent les souhaits des éditeurs, puisqu'il s'agit de leur contenu. Toutefois, nous ne sommes pas devins ! Il est donc crucial que les administrateurs de sites Web nous communiquent la manière dont ils souhaitent que leurs contenus soient indexés. Pour ce faire, il est possible de faire appel au protocole d'exclusion des robots. Il s'agit d'une norme technique éprouvée qui indique aux moteurs de recherche quels sites ou parties de site doivent ou non apparaître dans les résultats de recherche.

Robots.txt: contrôle à l'échelle du site

Au cœur du protocole d'exclusion des robots se trouve un simple fichier texte dénommé robots.txt, devenu la norme du secteur depuis de nombreuses années. Le fichier robots.txt vous permet de contrôler l'accès au contenu de votre site à plusieurs niveaux : intégralité de votre site, répertoires individuels, pages d'un type spécifique ou même pages individuelles.

> Sur mon site, il y a certaines pages que je ne souhaite pas voir indexées dans Google. Que dois-je faire pour empêcher que ces pages ne s'affichent dans les résultats de recherche de Google ?

En général, la plupart des propriétaires de sites souhaitent que Googlebot puisse accéder à leur contenu afin que leurs pages Web s'affichent dans les résultats de recherche de Google. Cependant, il peut arriver que vous ne souhaitiez pas que certaines de vos pages soient indexées. Il peut s'agir, par exemple, de pages accessibles uniquement contre paiement (ou de registres (logs) de connexions).

Vous pouvez exclure ces pages de l'index de Google en créant un fichier robots.txt que vous enregistrez dans le répertoire racine de votre serveur Web. Ce fichier robots.txt vous permet alors de répertorier les pages que les moteurs de recherche ne doivent pas indexer. La création robots.txt d'un tel fichier est un jeu d'enfant et permet aux éditeurs Web de contrôler très précisément la manière dont les moteurs de recherche accèdent à leurs sites Web.

Par exemple, si un administrateur de sites Web ne souhaite pas que ses registres (logs) de connexions internes soient indexés, son fichier robots.txt doit contenir les informations suivantes :

`User-Agent: Googlebot` : la ligne User-Agent (Agent-utilisateur) indique que la section suivante contient un ensemble d'instructions destiné uniquement aux robots Googlebot.

`Disallow: /logs/` : La ligne Disallow (Interdire) indique aux robots Googlebot qu'ils ne doivent pas accéder aux fichiers situés dans le sous-répertoire contenant les registres de connexions de votre site.

Le propriétaire du site a ainsi clairement indiqué qu'aucune des pages contenues dans le répertoire des journaux de consignation ne devaient figurer dans les résultats de recherche de Google.

Tous les principaux moteurs de recherche liront et respecteront les instructions définies dans votre fichier robots.txt. Si vous le souhaitez, vous pouvez également définir des règles spécifiques pour chaque moteur de recherche.

Balises Méta : contrôle affiné

Outre le fichier robots.txt qui vous permet de définir de manière concise des instructions pour un grand nombre de fichiers de votre site Web, vous pouvez également utiliser les balises Méta afin de contrôler individuellement chaque page de votre site. Pour ce faire, il vous suffit d'ajouter des balises Méta au code HTML de la page Web souhaitée afin de contrôler les modalités d'indexation de cette page. Grâce à leur flexibilité, le fichier robots.txt et les balises Méta vous permettent de spécifier des règles complexes d'accès de manière relativement facile.

J'ai sur mon site un article d'actualité accessible uniquement aux utilisateurs inscrits. Que dois-je faire pour que cet article ne figure pas dans les résultats de recherche de Google ?

Pour ce faire, il vous suffit d'ajouter une balise Méta NOINDEX dans la première section <head> de cet article. Voici comment se présente l'insertion de cette balise dans le code HTML :

```
<html>
<head>
<meta name="googlebot" content="noindex">
[...]
```

L'insertion de cette balise Méta empêche alors que Google n'indexe votre fichier.

Cependant, n'oubliez pas qu'il peut arriver que vous souhaitiez que Google indexe ce type de page, par exemple la page d'un journal archivé accessible en ligne après paiement. Tandis que Google n'affichera pas ce type de contenu dans ses résultats de recherche, certains services de Google, tels que News Archive Search, l'indexeront en indiquant clairement aux internautes que l'accès à ce contenu est payant. Pour savoir comment permettre l'indexation sur certains services uniquement, consultez la Foire aux Questions (FAQ).

Fichier Robots.txt ou balises Méta ?

En général, le fichier robots.txt constitue une solution efficace pour contrôler l'ensemble des pages d'un site. Les balises Méta permettent quant à elles de définir des règles d'accès spécifiques pour chacune des pages de ce site. Elles sont particulièrement utiles si vous êtes autorisé à modifier des fichiers distincts du site mais pas l'intégralité de ce dernier. Elles vous permettent également de spécifier des règles de contrôle d'accès complexes distinctes pour chacune des pages de votre site.

Parfois, l'une ou l'autre de ces solutions peut vous permettre de régler le même problème.

Que dois-je faire pour m'assurer que le texte d'une page est indexé, mais pas les images qu'elle contient ?

Vous pouvez bloquer l'accès aux images de cette page en spécifiant leur extension dans le fichier robots.txt. La présence des lignes suivantes dans un fichier robots.txt indique à Google de ne pas indexer les fichiers ayant pour extension *.jpg ou *.jpeg :

```
User-agent:   Googlebot
Disallow:   /*.jpg$
Disallow:   /*.jpeg$
```

Si votre système de gestion de contenu stocke les images dans un répertoire distinct, vous pouvez également exclure du processus d'indexation la totalité de ce répertoire. Si vos images sont stockées dans un répertoire dénommé "images", vous pouvez exclure ce répertoire du processus d'indexation de tous les moteurs de recherche en spécifiant les lignes suivantes :

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

Vous pouvez aussi ajouter la balise Méta NOINDEX à chaque fichier comportant une image. Toutes ces solutions vous permettent d'empêcher que vos images ne soient indexées. L'utilisation de l'une ou l'autre dépend de la quantité d'images et des images que vous souhaitez exclure du processus d'indexation.

Contrôle des éléments en mémoire cache et des extraits de texte

Les résultats de recherche contiennent généralement un lien “En cache”, ainsi qu’un court extrait de texte (snippet, en anglais). Voici, par exemple, l’un des premiers résultats qui s’affichent lorsqu’on lance une recherche sur “canard colvert” :

Extrait : il s’agit d’un court extrait de texte provenant de la page Web.

Lien “En cache” : il s’agit du lien qui renvoie les utilisateurs à une copie de la page indexée et stockée sur l’un des serveurs de Google.

À quoi servent les extraits de texte ? Les internautes visiteront plus sûrement un site Web si les résultats de recherche contiennent un extrait de texte issu des pages de ce site. Ces extraits permettent en effet de voir facilement si les résultats sont pertinents ou non par rapport à la requête saisie. Lorsqu’il leur est impossible de déterminer rapidement si un résultat est pertinent ou non, les internautes passent habituellement à un autre résultat.

À quoi servent les liens “En cache” ? Les liens “En cache” sont utiles dans un grand nombre de cas, notamment lorsque les sites auxquels ils renvoient sont momentanément indisponibles, lorsque les sites d’informations arrivent à saturation après la survenue d’événements majeurs ou lorsque des sites sont supprimés par inadvertance. Ces copies en mémoire cache offrent également un autre avantage, puisqu’elles mettent en surbrillance les mots recherchés par ‘internautes lui permettant ainsi de rapidement évaluer le degré de pertinence des pages proposées.

La plupart des éditeurs Web souhaitent que Google affiche les deux. Cependant, il arrive que les propriétaires de site préfèrent que l’une ou les deux options soient désactivées.

> Le contenu de mon journal change plusieurs fois par jour. Les robots Googlebot ne semblent pas indexer ce contenu aussi rapidement que nous le mettons à jour et le lien “En cache” renvoie à une page qui n’est plus d’actualité. Que dois-je faire pour que Google ne génère plus de lien “En cache” ?

Pour que le lien “En cache” n’apparaisse plus dans les résultats de recherche, le propriétaire de ce site d’informations peut insérer une balise NOARCHIVE dans la page en question :

```
<META NAME="GOOGLEBOT" CONTENT="NOARCHIVE">
```

De la même façon, pour que les résultats de recherche n’affichent plus d’extrait de texte, il vous suffit d’insérer une balise NOSNIPPET :

```
<META NAME="GOOGLEBOT" CONTENT="NOSNIPPET">
```

Remarque : l’ajout d’une balise NOSNIPPET empêche également l’affichage d’un lien en mémoire cache. Par conséquent, l’ajout de la balise NOSNIPPET dispense de l’insertion d’une balise NOARCHIVE.

Le contenu de votre site est-il unique et pertinent ?

Une fois votre site repéré et indexé, la question que vous devez vous poser est la suivante : le contenu de mes pages Web est-il unique et pertinent ?

Jetez d’abord un coup d’œil à votre texte dans son ensemble. Vos titres et liens textuels ont-ils un caractère descriptif ? Votre texte se lit-il facilement, est-il clair et intuitif ?

De même qu’un livre est découpé en chapitres distincts traitant chacun d’un thème spécifique, les pages Web de votre site doivent chacune être consacrées à un sujet particulier. Mots clés et expressions ressortent naturellement de ce genre de pages. En outre, les internautes resteront plus sûrement sur une page Web offrant un contenu et des liens pertinents.

Toutefois, assurez-vous d’utiliser des expressions que les internautes sont eux-mêmes susceptibles d’utiliser dans de leur recherche. Par exemple, si votre site est celui d’un club de passionnés de MG, assurez-vous que les termes “voitures” et “MG” apparaissent sur vos pages. Des termes tels que “voitures britanniques” ne sont pas suffisamment précis.

Rendre votre site plus visible : les meilleures pratiques

Les propriétaires de site nous demandent souvent quelles sont les meilleures solutions à leur disposition pour améliorer la visibilité et le classement de leur site dans nos résultats de recherche. Notre réponse est simple : “Mettez-vous à la place des internautes”. C’est d’ailleurs ce que nous essayons nous-mêmes de faire.

Qu’est-ce que cela signifie dans la pratique ? Tout d’abord, assurez-vous de communiquer aux visiteurs les informations qu’ils recherchent. La pertinence des informations est en effet l’élément déterminant lorsqu’il s’agit d’attirer et de retenir l’attention d’un nombre croissant d’internautes.

De nombreux propriétaires de site sont obsédés par le classement PageRank de leurs pages Web respectives. Mais n’oubliez pas qu’en plus de ce dernier, plus de 200 autres critères entrent en ligne de compte lors du classement de votre site Web. Par conséquent, concentrez-vous sur la qualité de votre contenu et son accessibilité, plutôt que d’essayer de trouver des solutions pour ruser avec l’algorithme des moteurs de recherche. Si un site ne respecte pas nos directives en matière de qualité, son indexation peut être bloquée.

Ce qu’il faut faire :

1. Créez un contenu pertinent et accrocheur : les visiteurs accèdent à vos pages via différents liens. Par conséquent, assurez-vous que chaque page est susceptible de retenir leur attention.
2. Impliquez les internautes : vous pouvez peut-être ajouter une section Commentaires ou un blog à votre site. Créer une communauté contribuera à drainer des passages plus fréquents vers votre site. Impliquer vos visiteurs est une façon d’accroître leur fidélité et d’améliorer la visibilité de votre site.
3. Surveillez votre site : utilisez le Centre pour les webmasters (voir page 13) afin de savoir quelles requêtes conduisent les visiteurs à votre site ou pour connaître l’évolution de votre classement dans les résultats de recherche suite à d’importantes modifications que vous aurez apportées.
4. Visez à obtenir des liens entrants de la part de sites de haute qualité.
5. Fournissez des liens textuels clairs : choisissez avec soin l’emplacement des vos liens textuels sur votre site et assurez-vous qu’ils contiennent les termes correspondant exactement à la rubrique ou à la page à laquelle ils renvoient.

Ce qu’il faut éviter :

1. N’encombrez pas vos pages Web avec des listes de mots clés.
2. N’essayez pas de “dissimuler” vos pages en rédigeant du texte visible uniquement par les moteurs de recherche et non par les utilisateurs.
3. Ne créez pas des pages et des liens destinés uniquement à induire en erreur les robots d’exploration et les moteurs de recherche.
4. N’affichez pas les noms, liens ou contenus importants sous forme d’images. N’oubliez pas que les moteurs de recherche ne peuvent pas en “lire” le contenu.

5. Ne créez pas plusieurs copies d’une même page sous différents adresses URL dans le but d’induire en erreur les moteurs de recherche.

En cas de doute, consultez nos directives destinées aux administrateurs de sites Web disponibles à l’adresse suivante : *google.fr/webmasters/guidelines.html*

Outils Webmaster Central

Notre société s’efforçant de fournir les résultats de recherche les plus pertinents et utiles du Web, il est logique que nous cherchions également à offrir une assistance accessible au plus grand nombre et équitable à tous les administrateurs de sites Web, quelle que soit la taille des sites qu’ils administrent. C’est pourquoi nous avons crée le Centre pour les webmasters, disponible`à l’adresse suivante : *google.fr/webmasters*.

Ce centre constitue une source d’informations et d’outils très utile pour tous les éditeurs Web. Il contient des réponses complètes aux questions portant sur l’exploration, l’indexation et le classement. Il rassemble également les commentaires des utilisateurs et des informations sur les problèmes rencontrés et permet aux administrateurs de soumettre leurs propres commentaires. Enfin, il propose des outils de diagnostic conçus pour aider les administrateurs à résoudre les éventuels problèmes de référencement rencontrés.

Voici un avant-goût de ce que le Centre pour les webmasters vous propose :

- Diagnostic des éventuels problèmes lors de l’accès aux pages et éventail de solutions
- Demande de suppression de certaines pages de notre index
- Vérification de l’efficacité de votre fichier robots.txt (autorisation ou blocage de l’accès aux pages désignées.)
- Affichage des statistiques des pages et requêtes de votre site Web :
- Statistiques de requête : elles déterminent quelles requêtes de recherche drainent le plus grand nombre de visiteurs vers votre site et quels thèmes votre site pourrait développer afin d’attirer encore plus d’internautes.

- Analyse de page : cet outil vous permet de voir vos pages Web telles que Google les voit, d'afficher les termes apparaissant le plus fréquemment sur votre site, ainsi que les liens entrant y conduisant et les descriptions des autres sites qui renvoient au vôtre.
- Fréquence d'exploration : cet outil vous indique à quelle fréquence votre site est exploré par les robots Googlebot et vous permet de signaler à Google si vous souhaitez qu'il soit exploré plus ou moins souvent.

Protocole Sitemaps

Le Centre pour les webmasters propose également aux éditeurs Web d'utiliser le protocole Sitemaps pour les résultats de recherche Web, Google Mobile et Google News.

Sitemaps est un protocole que nous prenons en charge à l'instar d'autres moteurs de recherche, afin d'aider les administrateurs de sites Web à nous fournir plus d'informations sur leurs pages Web. Le protocole Sitemaps vient en complément d'autres mécanismes d'exploration Web standard. Les administrateurs de sites Web peuvent utiliser ce protocole afin de communiquer à Google de plus amples informations sur les pages de leur site et optimiser ainsi leur exploration et augmenter leur visibilité dans les résultats de recherche Google.

Outre le protocole Sitemaps pour la recherche Web, nous proposons également un protocole Mobile Sitemaps qui permet aux éditeurs Web de soumettre à notre index des adresses URL de pages destinées à être consultées sur des appareils mobiles (PDAs, smartphones, etc.).

Quant aux éditeurs de sites d'actualité référencés sur Google News, le protocole News Sitemaps leur fournit des statistiques sur leurs articles : requêtes, fréquences d'affichage, etc. En conjonction avec les outils de diagnostic du Centre pour les webmasters, le protocole News Sitemaps fournit également des rapports d'erreur qui permettent de mieux comprendre les problèmes rencontrés par Google lors de l'exploration ou de l'indexation d'articles provenant d'un site d'actualité. Enfin, le protocole News Sitemaps permet aux éditeurs de soumettre à l'index de Google News les adresses URL de leur choix. À la différence des protocoles Web et Mobile Sitemaps, le protocole News Sitemaps est pour l'instant uniquement disponible en anglais. Nous espérons qu'il sera bientôt disponible dans d'autres langues.

Foire aux questions (FAQ)

Pourquoi ne puis-je pas bénéficier d'une assistance personnelle pour mon site Web ?
Selon certaines estimations, il y a 100 millions de sites sur le Web. Chacun de ces sites est important à nos yeux car sans eux, quelle que soit leur taille ou leur importance, notre index serait moins complet et donc au final moins utile à nos utilisateurs.

Webmaster Central offre une assistance précieuse pour tous les types de sites Web. Nous publions les questions des éditeurs et y répondons de sorte que tout le monde puisse bénéficier des informations échangées. Webmaster Central, c'est aussi une communauté conviviale d'administrateurs Web prêts à partager leurs astuces et conseils et à vous aider à résoudre vos éventuels problèmes.

Les annonces que vous affichez influencent-elles vos classements ? Les annonces que vous affichez et les résultats de recherche sont-ils totalement distincts ?
Les annonces et les résultats de recherche sont absolument indépendants les uns des autres. En fait, des équipes distinctes travaillent sur chacun de ces secteurs afin d'éviter toutes interférences. Nous sommes convaincus que nos résultats de recherche doivent être absolument impartiaux et objectifs afin de garantir aux utilisateurs des services la meilleure qualité qui soit.

Comment dois-je procéder pour que mon site soit référencé dans l'index de recherche de Google ?
Le référencement des sites dans les résultats de recherche Google est gratuit et se fait automatiquement (vous n'avez pas besoin de les soumettre manuellement à Google). Google est un moteur de recherche entièrement automatisé qui explore le Web de manière régulière afin d'y détecter les sites non encore référencés et les ajouter à notre index. En fait, la grande majorité des sites répertoriés dans nos résultats ne sont pas soumis manuellement, mais trouvés et ajoutés automatiquement par nos robots qui explorent le Web.

En outre, les outils de Google disponibles sur Webmaster Central offrent aux administrateurs de sites Web un moyen simple de soumettre à l'index de Google un plan de leur site ou une liste de leurs adresses URL et d'obtenir des rapports détaillés sur la visibilité de leurs pages sur Google. Grâce à ces outils, les propriétaires de sites peuvent automatiquement tenir Google informé des toutes les pages dont se compose leur site et de toutes les mises à jour apportées à ces pages.

En moyenne, combien de temps faut-il à Google pour repérer un nouveau site sur le Web et à quelle fréquence Google explore-t-il le Web de manière général ?
Il n'y a pas de durée standard pour qu'un site soit référencé. Les robots Googlebot explorent régulièrement le Web afin de mettre à jour notre index. Grâce à Webmaster Central, les administrateurs de sites Web peuvent voir à quelle fréquence leur site est exploré par les robots Googlebot et indiquer s'ils souhaitent que leur site soit exploré plus ou moins souvent.

Que se passe-t-il si je souhaite que mon site Web figure dans les résultats de recherche Web, mais pas dans les résultats de services distincts tels que Google News ou Google Image ?
Google donne toujours la possibilité aux éditeurs Web de ne pas figurer sur certains de ses services. Pour ce faire, il leur suffit de contacter la ou les équipes d'assistance en charge du ou des services concernés.

Ainsi que mentionné précédemment dans le présent guide, le protocole d'exclusion des robots peut être utilisé afin de bloquer l'indexation de certaines images et pages Web. À cette fin, vous pouvez également utiliser la fonctionnalité de suppression d'adresses URL disponible sur Webmaster Central, prenant en charge la recherche Web et la recherche Google Image.

En outre, le groupe d'ordinateurs Googlebot utilisant différents robots, vous pouvez définir avec précision les pages dont vous souhaitez bloquer l'accès :

- Robots Googlebot : explorent les pages pour notre index Web et notre index référençant les pages consacrées à l'actualité
- Robots Googlebot-Mobile : explorent les pages pour notre index référençant les contenus destinées aux appareils mobiles
- Robots Googlebot-Image : explorent les pages pour notre index référençant les images
- Robots Mediapartners-Google : explorent les pages afin d'identifier le contenu AdSense. Nous utilisons uniquement ces robots pour explorer les sites contenant des annonces AdSense.
- Robots Adsbot-Google : explorent les pages Web afin d'évaluer la qualité des pages de renvoi AdWords. Nous utilisons uniquement ces robots lorsque vous faites appel à Google AdWords pour promouvoir votre site.

Si vous souhaitez, par exemple, bloquer l'accès de l'intégralité des robots Googlebot à vos pages, vous pouvez utiliser la syntaxe suivante :

```
User-agent: Googlebot
Disallow: /
```

Est-ce que je peux choisir le texte de l'extrait qui apparaît dans les résultats de recherche ?
Non. En effet, une telle possibilité ne serait pas une bonne idée, ni du point de vue des internautes, ni du point de vue des créateurs de contenu. Nous choisissons un extrait issu du site qui comporte la requête de l'internaute dans son contexte et qui permet donc d'attester de la pertinence du résultat.

Les études démontrent que les internautes se rendront plus sûrement sur un site Web si les résultats de recherche contiennent un extrait de texte. Grâce à ces extraits, les internautes peuvent plus facilement évaluer la pertinence des résultats par rapport à leur requête. Lorsque les internautes ne peuvent pas évaluer rapidement la pertinence d'un résultat, ils passent généralement au résultat suivant.

Lorsqu'il nous est impossible de générer à l'aide d'algorithmes des extraits de texte exploitables et satisfaisants à partir du contenu de leurs pages Web, les éditeurs Web peuvent y ajouter la balise Méta suivante afin de nous fournir des informations supplémentaires. Cette balise doit être ajoutée dans la section <head> des pages concernées comme dans l'exemple suivant :

```
<meta name="description" content="Pourquoi Anne
n'aime-t-elle pas les lapins ? Nous allons bientôt le
savoir.">
```

Si vous ne souhaitez pas qu'un extrait de texte soit généré à partir du contenu de vos pages, il vous suffit d'ajouter la balise Méta NOSNIPPET comme suit :

```
<meta name="robots" content="nosnippet">
```

Enfin, nous utilisons parfois les descriptions de site contenues dans le répertoire ODP (Open Directory Project) pour les extraits de texte que nous affichons dans les résultats. Si vous ne souhaitez pas que nous utilisions cette description, il vous suffit d'ajouter la balise Méta suivante :

```
<meta name="robots" content="noodp">
```

Sur mon site, les articles consacrés aux dernières nouvelles s'affichent pendant quelques heures avant d'être mis à jour, puis sont déplacés vers la section des articles standard. Je souhaite que tous les articles de mon site soient référencés dans l'index de Google sauf les articles consacrés aux dernières nouvelles.
Pour ce faire, vous pouvez placer tous les articles consacrés aux dernières nouvelles dans un même répertoire, puis empêcher l'accès de ce dernier aux robots Googlebot à l'aide du fichier robots.txt.

Vous pouvez également, à cette fin, ajouter la balise Méta NOFOLLOW dans la section <HEAD> de la page HTML consacrée aux dernières nouvelles. Cette balise indique aux robots Googlebot qu'ils ne doivent suivre aucun des liens trouvés sur cette page. N'oubliez pas cependant, que cette balise empêche uniquement les robots Googlebot de suivre les liens allant d'une page à une autre. Si une autre page Web renvoie à ces articles, Google pourra les trouver et donc les indexer.

Si je dispose de plusieurs noms de domaine et si je diffuse le même contenu à partir de ces noms différents, mes sites courent-ils le risque d'être exclus de vos résultats de recherche ?
Bien que certains éditeurs Web tentent parfois d'utiliser le clonage de site et les sites miroirs pour induire en erreur les moteurs de recherche, il y a des cas où le clonage de contenu est tout à fait justifié et légitime. Nous ne souhaitons pas pénaliser les sites se trouvant dans ce cas. Par exemple, nous ne considérons pas que le même contenu exprimé dans deux langues différentes (par exemple, une version anglaise et une version française) est du contenu cloné.

Diffuser le même contenu sur plusieurs sites Web différents, notamment dans le cadre d'une syndication d'articles, ne signifie pas nécessairement que l'un ou plusieurs de ces sites seront entièrement supprimés des résultats de recherche. Toutefois, n'oubliez pas que les articles apparaissant sur plusieurs sites sont moins bien classés que les articles ne figurant que sur un seul, car les premiers bénéficient uniquement d'une fraction des liens entrants dont les articles publiés en un seul exemplaire bénéficient. En principe, un article publié sur un seul site sera mieux classé qu'un article publié sur plusieurs. Il sera donc consulté par un plus grand nombre d'internautes.

En outre, afin de garantir la qualité de nos résultats de recherche, nous n'y faisons pas figurer plusieurs exemplaires d'une même page. Nous préférons généralement y afficher une seule version de cette page. Les administrateurs de sites Web peuvent néanmoins nous indiquer à l'aide du fichier robots.txt quelle version de la page ils souhaitent que nous affichions dans nos résultats ou bloquer à l'aide des balises Méta appropriées l'affichage de toutes versions non désirées.

Pourquoi l'accès de mon site à l'index Google est-il bloqué ?

Tout d'abord, il se peut que l'indexation de votre site ne soit pas bloquée. De nombreuses autres raisons peuvent en effet expliquer pourquoi votre site n'apparaît pas dans nos résultats de recherche (voir pages 5 à 11).

Si aucun obstacle n'empêche le repérage ou l'indexation de votre site, alors il se peut effectivement que son accès à l'index soit bloqué. Cette situation se produit notamment lorsque les sites concernés ne respectent pas les normes de qualité définies par les directives que nous avons rédigées à l'attention des administrateurs de sites Web (disponibles sur le Centre pour les webmasters). Il s'agit le plus souvent de sites Web qui utilisent des méthodes déloyales pour obtenir un classement plus élevé dans nos résultats de recherche. Ces infractions à nos directives incluent notamment la dissimulation (cloaking, en anglais) qui consiste à rédiger du texte de sorte qu'il soit visible par les moteurs de recherche mais pas par les utilisateurs) ainsi que la configuration de pages et / ou de liens dans le seul but d'induire en erreur les moteur de recherche et de manipuler les résultats de recherche.

Lorsqu'un administrateur de sites Web pense que l'un de ses sites est en infraction avec nos directives en matière de qualité, il peut modifier ce site afin qu'il s'y conforme, puis cliquer sur le lien "Request re-inclusion" (Demander une réindexation) accessible à partir de l'interface des outils du Centre pour les webmasters afin que nous procédions à une réévaluation du site.

Glossaire

Adresse URL (Uniform Resource Locator)
Adresse d'un site Web sur Internet qui se compose des éléments suivants : http (protocole d'accès), nom de domaine (www.google.fr) et dans certains cas de l'emplacement d'un autre fichier (www.google.fr/webmaster).

Balises Méta
Balises dans le code HTML permettant de décrire le contenu d'une page Web. Les balises Méta peuvent être utilisées afin de définir des modalités d'indexation spécifiques pour chacune des pages d'un site.

Contenu dynamique
Contenu tel que des images, des animations ou des vidéos qui utilisent le langage Flash ou Javascript, des cadres ou des adresses URL générées dynamiquement.

Dissimulation (Cloaking)
Technique qui consiste à montrer aux moteurs de recherche un contenu différent de celui visible par les utilisateurs.

Exploration (Crawling)
Processus utilisé par les moteurs de recherche pour collecter des pages sur le Web.

Extension de fichier
Nom attribué aux fichiers informatiques (.doc, .txt, .pdf, etc.) indiquant généralement la nature des données contenues dans le fichier.

HTML (Hypertext Markup Language)
Langage de marquage utilisé sur le Web afin de structurer le texte.

Indexer
Processus consistant à référencer le contenu d'un site dans un moteur de recherche.

Lien "En cache"
Image d'une page Web capturée par les robots Googlebot lors de leur dernière visite. Une copie en cache permet aux utilisateurs d'afficher une page même lorsque sa version en ligne n'est pas disponible. Le contenu de cette copie peut toutefois varier légèrement de la version en ligne. Pour afficher la copie en cache d'une page, cliquez sur le lien "En cache" affiché sous le résultat de la recherche.

Mot clé
Terme saisi dans la zone de recherche d'un moteur de recherche, lequel lance une recherche afin de trouver des pages contenant ces termes.

Protocole d'exclusion des robots
Norme technique indiquant aux moteurs de recherche quels sites ou parties de site doivent être ou non référencés dans les résultats de recherche.

PageRank
Fonctionnalité proposée par Google contribuant à déterminer le classement d'un site dans nos résultats de recherche. Ce classement est établi en respectant le caractère profondément démocratique du Web, puisqu'il utilise son organisation sous forme de liens pour déterminer la valeur individuelle de chaque page. Le score PageRank des sites importants et de grande qualité est plus élevé. Cet élément est pris en compte par Google lors des recherches. Google associe ce classement à des techniques élaborées de recherche de texte correspondant aux critères saisis afin de trouver des pages pertinentes et importantes par rapport aux requêtes des internautes.

Répertoire racine
Répertoire principal dans un système de fichiers informatiques.

Robot d'exploration (Crawler)
Logiciel utilisé pour repérer, puis indexer les adresses URL du Web ou d'un réseau intranet.

Robots.txt
Fichier texte permettant aux éditeurs Web de contrôler l'accès de leur site à plusieurs niveaux : intégralité du site, répertoires individuels, pages d'un type particulier ou même pages individuelles. Ce fichier signale aux robots d'exploration quels répertoires peuvent être ou non explorés.

Système de gestion de contenu
Logiciel qui permet de gérer différents types de contenus : fichiers informatiques, images, fichiers audio, contenus Web, etc.

Site miroir
Version clonée d'un site Web, parfois utilisée pour induire en erreur les moteurs de recherche et ainsi optimiser l'indexation et le classement d'un site Web donné.

Pour en savoir plus sur le Centre pour les webmasters: veuillez consulter :

google.fr/webmasters/

Google™