Engineer

plan de site

Liste des pages d’un site destinée aux machines

Ce que c’est

Un plan de site est un fichier qu’un site publie à son propre sujet : les adresses que son propriétaire souhaite faire connaître aux moteurs de recherche, avec une entrée par page et quelques informations facultatives. Sans lui, un robot découvre le site plus lentement, en suivant les liens des pages déjà connues. Certaines pages se trouvent au bout d’une longue chaîne de liens ou n’ont presque aucun lien entrant.

Le format partagé par les grands moteurs est le protocole Sitemaps, publié sur sitemaps.org. Sa version XML place la liste dans un élément `<urlset>`, avec un `<url>` par page. À l’intérieur, `<loc>` contient l’adresse complète et constitue le seul élément enfant obligatoire. `<lastmod>` indique la date de dernière modification de la page, `<changefreq>` estime sa fréquence de modification et `<priority>` est une valeur de 0,0 à 1,0 qui la classe par rapport aux autres pages du site, jamais à celles des autres. Le protocole accepte aussi un fichier texte avec une adresse par ligne, ou un flux RSS ou Atom.

Le protocole fixe des limites strictes. Un fichier contient au maximum 50 000 adresses et 50 Mo non compressés. Il peut être compressé avec gzip, doit être en UTF-8 et ses valeurs doivent être échappées sous forme d’entités. Chaque adresse doit compter moins de 2 048 caractères. Un site plus grand répartit sa liste et publie un index de plans de site, un second type de fichier qui peut désigner jusqu’à 50 000 plans de site sur le même site. Un plan de site ne peut par ailleurs lister que des adresses situées à son emplacement ou en dessous : un fichier `/catalog/sitemap.xml` couvre les pages de `/catalog/` et ne sort pas de ce dossier.

Un moteur découvre l'emplacement du fichier de deux façons. Une ligne `Sitemap:` dans robots.txt donne son adresse complète, indépendamment de tout groupe user-agent ; le propriétaire peut aussi le soumettre dans les outils du moteur, comme Google Search Console.

Le même terme désigne également une page de liens pour les visiteurs humains. Lorsque la distinction importe, on précise « plan de site XML » pour celui destiné aux programmes.

Utilisez le concept pour prendre une décision

Utilisez « Qu'est-ce qu'un plan de site ? » 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 documentation de Google expose clairement les limites. Un site dont les pages sont correctement reliées peut généralement être découvert sans plan de site. Google situe la limite d'un « petit » site autour de 500 pages devant apparaître dans les résultats. Il précise aussi qu'un plan de site « ne garantit pas que tous ses éléments seront explorés et indexés » et que sa soumission « n'est qu'une indication ». Il est surtout utile aux grands sites, aux nouveaux sites vers lesquels pointent peu de liens depuis d'autres sites et aux sites riches en vidéos, images ou actualités.

Presque tous les générateurs de site en écrivent un sans travail supplémentaire. La véritable décision porte donc sur ce que le fichier affirme : chaque entrée est une déclaration que le moteur peut vérifier sur la page.

Quelles adresses inclure ? Google demande l'adresse canonique de chaque page. Une liste gonflée de redirections, de doublons ou de pages marquées `noindex` demande exactement ce que le reste du site invite à laisser de côté.

Quels champs ? Google indique qu'il ignore `<priority>` et `<changefreq>`. Il n'utilise `<lastmod>` que « si cette date est systématiquement exacte et vérifiable, par exemple en la comparant à la dernière modification de la page ». Le protocole précise que la date « doit être celle de la dernière modification de la page liée, pas celle de la génération du plan de site ». Pour Google, une modification du contenu principal, des données structurées ou des liens est significative ; une nouvelle année de copyright est mineure.

C'est là que se situe le coût. Une date issue du système de fichiers ne coûte rien et est généralement fausse : déploiements, exports et nouvelles copies de travail touchent des fichiers dont le texte n'a pas changé. Une date tirée de l'historique des versions est honnête, mais la construction doit alors disposer de cet historique et le lire dans le bon ordre. Une date saisie à la main reste exacte jusqu'à ce que quelqu'un oublie de la modifier.

À quoi cela ressemble en pratique

Le fichier robots.txt de ce site indique une seule adresse, `https://engineer.company/sitemap.xml`, et ce fichier est un index qui répertorie les plans de site des éditions linguistiques publiées. Chaque entrée correspondant à une page contient l'adresse et une date, ainsi que des liens vers cette même page dans ses autres traductions publiées. `<priority>` et `<changefreq>` sont omis, puisque Google dit les ignorer et que le protocole ne les impose pas. Lors de la rédaction de cet article, l'entrée de la page consacrée à la préparation à l'IA se présentait ainsi, liens de langue mis de côté :

```
<url>
  <loc>https://engineer.company/ai-ready/</loc>
  <lastmod>2026-10-15T14:25:16+02:00</lastmod>
</url>
```

La liste est volontairement sélective. Les sites satellites spécialisés de l'entreprise republient ces réalisations avec une URL canonique pointant vers l'original. Le même modèle exclut ces copies du plan de site de chaque satellite : une entrée demande au moteur d'indexer cette adresse précise, et demander une copie tout en désignant une autre page comme original serait contradictoire. Chaque hôte n'annonce que ses propres pages.

Ces dates ont leur propre histoire. À un moment, le plan de site annonçait aux moteurs que 172 de ses 176 pages avaient changé au même instant. La cause était ordinaire : la date provenait de l'heure de modification de chaque fichier, et l'export du contenu réécrit tous les fichiers à chaque exécution. La correction a déplacé la source vers l'historique des versions, avec une chaîne de repli qui essaie un champ explicite, puis l'historique des commits, puis le fichier. Elle cachait un piège utile à connaître : le nom littéral du champ doit figurer dans la liste, sinon le générateur ne le lit jamais. Ensuite, une resynchronisation après un mois de modifications n'a touché que huit fichiers sur 186, plutôt que tous, et le plan de site dit quelque chose de vrai.

Cette date sert désormais au-delà du plan de site. Comme l'explique la page consacrée à la découverte par les agents d'IA, chaque déploiement annonce les changements par IndexNow, un protocole auquel participent Bing et plusieurs autres moteurs. Le script de déploiement décide de ce qui a changé en comparant le `lastmod` de chaque page à son dernier envoi : un déploiement qui modifie une page n'annonce donc qu'une page. Google ne figure pas dans la liste des moteurs participants d'IndexNow ; il découvre donc la modification à sa prochaine lecture du plan de site. Un contrôle de la construction vérifie chaque date du plan de site par rapport à l'historique de la page et refuse la publication lorsqu'une date est fausse. Pour les personnes plutôt que les programmes, il existe aussi un index lisible de toutes les pages.

Où les problèmes surviennent

Toutes les pages prétendent avoir changé aujourd'hui. Le plan de site donne la même date, ou une date par déploiement, à des centaines de pages. Le moteur n'apprend rien sur ce qui est nouveau. La cause est une date issue du système de fichiers ou de l'horloge de construction. Google n'utilise `<lastmod>` que si la date est systématiquement exacte et vérifiable, ce qu'un horodatage partagé n'est manifestement pas. Prenez la date dans l'historique du contenu et comparez-y un échantillon d'entrées après toute modification de la construction.

La liste demande des pages que le site veut exclure. Redirections, doublons portant des paramètres de suivi, pages marquées `noindex`, hôtes de préproduction et copies dont l'URL canonique désigne une autre page apparaissent dans les plans de site des générateurs qui listent chaque fichier construit. Chaque entrée envoie le robot à une adresse que la page désavoue ensuite. Ne listez que l'adresse canonique de chaque page destinée aux résultats, et faites répondre le plan de site et la balise canonique selon la même règle, comme les modèles de ce site.

Les entrées dépassent le périmètre du fichier. Un plan de site situé à `/blog/sitemap.xml` liste des pages de boutique ou écrit des chemins relatifs comme `/page.html`. Le protocole limite un fichier aux adresses de son dossier ou de ses sous-dossiers, et Google explore les adresses telles qu'elles sont écrites. Publiez le plan de site à la racine, écrivez chaque `<loc>` comme une adresse `https://` complète du même hôte et passez à un index de plans de site avant d'approcher 50 000 entrées ou 50 Mo.

Des pages listées depuis des mois restent hors de l'index. Le propriétaire prend le plan de site pour une instruction et attend. Google parle d'une indication : une adresse listée peut toujours être ignorée ou regroupée avec un doublon pour des raisons extérieures au fichier. Regardez la page plutôt que la liste : d'autres pages créent-elles des liens vers elle ? Son URL canonique la désigne-t-elle elle-même ? robots.txt ou une règle `noindex` empêche-t-elle le robot d'y accéder ?

Pour aller plus loin

- Format XML Sitemaps — le protocole : chaque balise, les limites de taille, les fichiers d'index et les adresses qu'un fichier peut lister.
- Créer et soumettre un plan de site — les règles de Google, notamment les balises ignorées et les conditions auxquelles une date est jugée fiable.
- Comprendre les plans de site — les cas où Google recommande un plan de site et ceux où des liens corrects suffisent.
- Documentation IndexNow — le protocole qui informe les moteurs participants d'une modification le jour de sa livraison.
- RFC 9309 : protocole d'exclusion des robots — la norme robots.txt, qui permet au robot de lire une ligne `Sitemap:` à côté des règles.

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 qu'un plan de site ? » 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/sitemap/