Engineer

サイトマップ

サイトのページを機械向けに列挙した一覧

どのようなものか

サイトマップとは、ウェブサイトが自身について公開するファイルです。所有者が検索エンジンに知らせたいアドレスを、ページごとに1件ずつ列挙し、必要に応じて、それぞれにいくつかの補足情報を添えます。これが必要になるのは、クローラーが通常、すでに知っているページからリンクをたどるという遅い方法でサイトを知るためです。長いリンクの連鎖の先にあるページや、ほかからのリンクがほとんどないページもあります。

主要な検索エンジンが共通して使う形式は、sitemaps.orgで公開されているSitemapsプロトコルです。XML形式では、一覧を `<urlset>` 要素で囲み、各ページに一つの `<url>` を置きます。その中の `<loc>` には完全なアドレスを記述し、必須の子要素はこれだけです。 `<lastmod>` はページの最終更新日、 `<changefreq>` は今後の更新頻度の予測、 `<priority>` は0.0から1.0の値で、サイト内のほかのページに対する優先度を示します。他者のページとの比較ではありません。このプロトコルは、1行に一つのアドレスを書くプレーンテキストファイルや、RSS、Atomフィードも受け付けます。

プロトコルには明確な制限があります。一つのファイルに含められるアドレスは最大50,000件、非圧縮で50 MBまでで、gzipで圧縮でき、UTF-8で記述し、値をエンティティエスケープしなければなりません。各アドレスは2,048文字未満に収めます。より大きなサイトでは一覧を分割し、同じサイト内のサイトマップを最大50,000件列挙する、別形式のファイルであるサイトマップインデックスを公開します。また、サイトマップが列挙できるのは、自身と同じ階層か、その下にあるアドレスだけです。 `/catalog/sitemap.xml` のファイルは `/catalog/` のページを対象とし、そのフォルダーの外には及びません。

検索エンジンがファイルの場所を知る方法は二つあります。robots.txtの `Sitemap:` 行で、ユーザーエージェントのグループとは独立して完全なアドレスを示すか、所有者がGoogle Search Consoleなど、検索エンジン自身のツールか

ら送信します。

同じ言葉は、人間の訪問者に向けたリンク集のページも指します。区別が必要なときは、プログラム向けのものを「XMLサイトマップ」と呼びます。

概念を判断に生かす

「サイトマップとは？」を実際の判断に生かしましょう。その概念を、開発または購入するシステムに結び付けて考えるお手伝いをします。 お客様のAIアシスタントがサイトやアプリを見つけ、読み取り、操作できる状態を、最初の約束だけで済ませず、リリースごとに検証します。

サービスを見る お問い合わせ →

用語の背後にある前提を検証する

解説とリンク先の実績を参考に、プロジェクトについての質問を整理してください。ツールを選ぶ前に、前提を一緒に検討できます。 3言語の多言語Hugoサイトを構築 — 108 3言語・4形式を提供するテンプレートファイル

3言語の多言語Hugoサイトを構築 お問い合わせ →

なぜ重要なのか

Googleの文書は、その限界を率直に説明しています。ページが適切にリンクされたサイトは、通常サイトマップなしでも発見でき、Googleは検索結果に載せるべきページが約500件までのサイトを「小規模」としています。また、サイトマップは「含まれるすべての項目がクロールされ、インデックスに登録されることを保証しない」とし、送信は「単なるヒント」だと述べています。特に役立つのは、大規模なサイト、ほかのサイトからのリンクが少ない新しいサイト、動画、画像、ニュースを多く扱うサイトです。

ほぼすべてのサイトジェネレーターが追加の手間なく生成するため、本当に決めるべきなのはファイルで何を主張するかです。各項目は、検索エンジンが実際のページと照合できる記述になります。

どのアドレスを入れるのでしょうか。Googleは各ページの正規アドレスを求めています。リダイレクト、重複、`noindex` を指定したページで一覧を膨らませると、サイトのほかの部分では対象外にするよう伝えているものを、ここでは対象にするよう求めることになります。

どの項目を使うのでしょうか。Googleは `<priority>` と `<changefreq>` を無視すると述べています。 `<lastmod>` を利用するのは、「一貫して正確で、例えばページの最終変更と比較することで、その正確さを検証できる場合」に限ります。プロトコルも、日付は「サイトマップの生成日時ではなく、リンク先のページが最後

に変更された日付に設定しなければならない」と定めています。Googleは、ページの主要な内容、構造化データ、リンクの変更を重要とし、著作権表示の年を更新することは軽微だとしています。

ここに運用の負担があります。ファイルシステムから日付を取るのは簡単ですが、デプロイやエクスポート、新しいチェックアウトでも、文章が変わっていないファイルの日時が更新されるため、通常は誤った値になります。バージョン履歴から取る日付は内容に即していますが、ビルドにはその履歴を用意し、正しい順序で参照する必要があります。手入力の日付は、誰かが更新を忘れるまでは正確です。

実際の活用例

このサイトのrobots.txtは、`https://engineer.company/sitemap.xml`という一つのアドレスを示しています。このファイルは、公開している各言語版のサイトマップを列挙するサイトマップインデックスです。各ページの項目には、アドレスと日付に加え、公開済みの他の翻訳版にある同じページへのリンクが含まれます。`<priority>`と`<changefreq>`は省いています。Googleがどちらも無視すると明示しており、プロトコル上も必須ではないためです。この記事の執筆時点で、AI対応についてのページの項目は、言語間リンクを除くと次のようになっていました。

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

一覧の対象は意図的に選別しています。会社の専門分野別サテライトサイトは、正規URLをこのサイトへ向けて実績を再公開します。同じテンプレートが、それらのコピーを各サテライトのサイトマップから除外します。項目を載せることは、その正確なアドレスをインデックスに登録するよう検索エンジンへ求めることなので、別のページを原本として示しながらコピーの登録を求めると矛盾するためです。各ホストは自分自身のページだけを通知します。

日付にも、それ自体の経緯があります。かつてサイトマップは、176ページのうち172ページが同じ瞬間に変更されたと検索エンジンへ伝えていました。原因はありふれたものでした。日付を各ファイルの変更時刻から取っており、コンテンツのエクスポートが毎回すべてのファイルを書き直していたのです。修正では、情報源をバージョン履歴へ移し、明示的なフィールド、コミット履歴、ファイル

の順に試すフォールバックの連鎖を設けました。ここには知っておくべき落とし穴がありました。その一覧にフィールド名そのものが含まれていないと、ジェネレーターはそのフィールドを一度も読みません。修正後、1か月の編集を経て再同期したときに更新されたのは、186ファイルすべてではなく8ファイルだけで、サイトマップは内容に即したことを述べるようになりました。

その日付は今、サイトマップ以外にも使われています。AIエージェントに発見されるためのページで説明しているように、デプロイのたびに、Bingなど複数の検索エンジンが参加するIndexNowプロトコルで変更を通知します。デプロイスクリプトは、各ページの `lastmod` を前回送信した値と比較して変更を判断するため、1ページだけを編集するデプロイなら、通知も1ページになります。GoogleはIndexNow自身が挙げる参加検索エンジンの一覧に含まれないため、次にサイトマップを読んだときに変更を知ります。ビルド内の検査は、すべてのサイトマップの日付をページの履歴と照合し、一つでも誤っていれば公開を拒否して、日付の正確さを保ちます。プログラムではなく人間向けには、全ページを読める形でまとめた索引もあります。

問題が起きるところ

すべてのページが今日変わったと主張する。 数百ページにわたってサイトマップが一つの日付、またはデプロイごとの日付を持っていると、検索エンジンはどの内容が新しいのか分かりません。原因はファイルシステムやビルド時刻から取った日付です。Googleが `<lastmod>` を利用するのは、一貫していて正確だと検証できる場合だけで、共通のタイムスタンプは明らかにその条件を満たしません。日付はコンテンツ自身の履歴から取り、ビルドを変更した後は、項目をいくつか選んで履歴と照合してください。

対象外のページまで一覧に載せてしまう。 ビルドしたファイルをすべて列挙するジェネレーターのサイトマップには、リダイレクト、追跡パラメーターを持つ重複アドレス、 `noindex` を指定したページ、ステージング用ホスト、正規URLが別の場所を指すコピーまで現れます。それぞれが、ページ自身は対象外だと伝えるアドレスへクローラーを送ります。検索結果に載せたい各ページの正規アドレスだけを一覧にし、このサイトのテンプレートのように、サイトマップとcanonicalタグが同じルールに基づいて生成されるようにしてください。

項目がファイルの対象範囲から外れる。 `/blog/sitemap.xml` のサイトマップにショップのページを載せたり、 `/page.html` のような相対パスを書いたりします。プロトコルは、ファイルと同じフォルダーか、その下にあるアドレスだけを認め、Googleは記述されたとおりのアドレスをクロールします。サイトマップは

ルートに公開し、すべての `<loc>` を同じホスト上の完全な `https://` アドレスで記述し、一覧が50,000件または50 MBに近づく前に、サイトマップインデックスへ切り替えてください。

数か月載せているページがインデックスに入らない。 所有者はサイトマップを指示だと受け取り、待ち続けます。Googleはそれをヒントと呼んでいます。載せたアドレスでも、ファイル外の理由によって対象から外されたり、重複として統合されたりすることがあります。一覧ではなくページを確認してください。ほかのページからリンクされているか、正規URLが自身を指しているか、robots.txtや `noindex` のルールでクローラーを遠ざけていないかを調べます。

参考資料

- Sitemaps XML形式 — すべてのタグ、サイズ上限、インデックスファイル、ファイルに載せられるアドレスを定めたプロトコル本体です。
- サイトマップを作成して送信する — 無視するタグや、日付を信頼する条件を含むGoogleのルールです。
- サイトマップについて — サイトマップが必要な場合と、適切なリンクがあれば足りる場合についてのGoogleの説明です。
- IndexNowの文書 — 変更を公開した日に、参加する検索エンジンへ知らせるプロトコルです。
- RFC 9309：Robots Exclusion Protocol — ルールと並べて `Sitemap:` 行をクローラーに読ませられるrobots.txtの標準です。

用語集

実際の制約に合わせて選ぶ

業務の流れ、既存のインターフェース、変更を必要とする利用者をお知らせください。検証可能な納品内容と、お客様のチームが保守を続けられる引き継ぎ内容を定められます。

料金を見る お問い合わせ →

用語を実務に結び付ける

「サイトマップとは？」がお客様の組織にどう当てはまるかをご相談ください。現在のシステムの具体例が一つあれば、話を始められます。

サービスを見る

welcome@engineer.company +45 50 10 61 84 デンマーク、コペンハーゲン

このページを持ち歩く

連絡先カード

心を込めて丁寧に制作： D.

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

著作権 © 2025 – 現在 · Engineer ApS · すべての権利を留保します。

印刷元： https://engineer.company/ja/glossary/sitemap/