Engineer

robots.txt

クローラーに巡回しない場所を伝えるファイル

どのようなものか

ウェブ上の各ホストは、固定のパス `/robots.txt` に、巡回プログラムへ宛てた要望を並べるプレーンテキストファイルを公開できます。Martijn Kosterが1994年に慣例を定め、2022年9月にIETFがRobots Exclusion ProtocolとしてRFC 9309を公開し、規則の厳密な記述と、エラーやキャッシュへの対応を加えました。

ファイルはグループの一覧です。グループは、一つ以上の `user-agent` 行で始まり、各行に `ExampleBot` などの製品トークンでクローラーを指定するか、すべてのクローラーを表す `*` を記します。その下に、それぞれパスを持つ `allow` 行と `disallow` 行を置きます。クローラーは、大文字と小文字を区別せず、自分のトークンを指定したグループを探して従います。複数のグループが自分を指定していれば、その規則を一つに統合します。自分向けのグループがないクローラーだけが `*` グループを使い、どちらもなければ従う規則はありません。

グループ内の規則は、そのパスがアドレスの先頭と一致すると適用され、最も具体的な一致、つまりオクテット数が最も多いものが優先されます。 `allow` と `disallow` が同じ長さで競合すれば `allow` が優先され、どの規則にも一致しないアドレスは許可されます。パスの照合範囲を拡張する文字が二つあり、 `*` は任意の文字の並びを表し、 `$` はアドレスの末尾を指定します。 `#` はコメントの開始を表し、このファイル自身のパスは常に許可されます。

RFCは、ファイルの取得方法も定めています。サービスの最上位にある、小文字のそのパスに置くため、各ホストが自分のファイルで応答し、あるサブドメインのファイルは、別のサブドメインには何も指示しません。形式はUTF-8のプレーンテキストです。クローラーは、少なくとも5回まではリダイレクトをたどるべきで、キャッシュしたコピーを24時間を超えて使うべきではありません。また、解析する量に上限を設けられますが、500 KiB未満にはできません。

プロトコルに含まれない行も許容されます。RFCは「その他のレコード」と呼び、`Sitemap`を例に挙げ、クローラーが解釈するかどうかを任意としています。ただし、その行をグループの終わりとして扱ってはならないという条件があります。

概念を判断に生かす

「robots.txt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

RFCは導入部で、「これらの規則は、アクセスを認可する仕組みではありません」と限界を示しています。ファイルは要望にすぎません。尊重するクローラーは立ち入りませんが、無視するクローラーは、サーバーが渡すものを読みます。セキュリティーの節は、ファイルにパスを記すと、それが公開され、見つけられるようになるとも説明しています。そのため、非公開にすべきものには、HTTP認証などの実際の制御が必要です。

Googleの文書は、目的をさらに限定しています。「主に、リクエストによってサイトに過負荷がかかるのを防ぐために使う」もので、「ウェブページをGoogleの検索結果から除外する仕組みではない」と説明しています。巡回を許可しないアドレスも、ほかのページからリンクされればインデックスに入り得ます。その場合、クローラーはページを読んでいないため、説明なしで検索結果に出ます。Googleは、結果から除外するには`noindex`の規則かパスワードを使うよう案内しています。ただし、`noindex`は、クローラーが取得できるページでしか機能しません。

このファイルは、本来の用途では力を発揮します。重要でないページやほぼ同じ内容のページなど、サーバーへのリクエストは増やしても、検索エンジンにはほとんど役立たない部分へ、クローラーが進まないようにします。影響の広さは危険にもつながります。`*`の下の一行は、ファイルを尊重するすべてのクローラーと、すべてのアドレスに適用され、変更が伝わるまで一日かかることもあります。そのため、誤ったグループに`Disallow: /`を入れると、サイト全体を閉じてしまいます。

実際の活用例

このサイトのファイルは長く、ほぼすべてがコメントです。クローラーが動作の根拠にするのは、次の4行です。

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

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

一つのグループですべてのクローラーを扱い、`Allow: /` がすべてのアドレスに一致します。どの規則にも一致しないアドレスは許可されるため、規則のないグループでも結果は同じです。それでも明記するのは、開放する方針を選んでいるからです。直前のコメントは、方針を「受け継いだのではなく、決めたもの」と表現し、AIエージェントがサイトに求めるものを扱うページも、巡回方針を、このサイトが意図的に下した判断の一つと説明しています。

`Content-Signal` は、RFC 9309が定義するレコードではありません。Content Signalsの語彙に由来し、ここでは、検索、推論時の質問への回答、学習のための収集に同意しています。RFC上は「その他のレコード」の一つです。慣例を知るクローラーは解釈してよく、それより古いクローラーは、ファイル自身のコメントにあるとおり、その行を読み飛ばします。どちらも、その行をグループの終わりと見なしてはなりません。`Sitemap:` 行も同種のレコードで、サイトのサイトマップインデックスを指します。クレジットのページも、ファイルは「検索クローラーもAIシステムも含め、すべてを許可する」と同じようにまとめ、サイトマップも挙げています。

何を記さないかも、意図的に決めています。AI対応のページは、このサイトが採用しなかった五つの推奨事項の中に、エージェントごとの規則を挙げています。「あるクローラー向けに書いた規則は、そのクローラーにとって、気付かれないまま共通の規則を置き換える」ため、一つのrobots規則をすべてのエージェントに使います。一つのAIクローラーに専用グループを用意していたら、そのクローラーはそのグループだけを読み、後から加えた `Disallow` が `*` の下にあっても、そのクローラーには決して届きません。

残りは、ファイルを開いた人のためのコメントです。方針の理由に続いて、`llms.txt`、`agents.txt`、`humans.txt`、`security.txt` などの機械可読ファイルのアドレスを記しています。「robots.txtには、それらのためのフィールドがない」ため、コメントにしています。そのため、コメントも公開される内容の一部で

す。テンプレートの `#` 行は、そのままのバイト列ですべてのクローラーに配信されます。そこに二度、非公開リポジトリーの計画文書を指す、保守担当者向けのメモが入っていました。現在は、生成済みサイトの検査が、配信するファイルにリポジトリー内へのパスがあると失敗します。また、テンプレートの保守メモは、サイト生成時に除去されるテンプレートコメントに置きます。

問題が起きるところ

全員向けの規則が、一つのクローラーには適用されなくなる。 一つのフォルダーに入らせないためなどに、単一のボット用グループを追加すると、その時からそのボットは、 `*` の下にあるすべての `Disallow` を含む、すべての規則を無視します。RFC 9309は、自分を指定するグループがないときに限り、 `*` グループを使うよう定めています。Googleの文書も、個別のグループと `*` グループは「結合しない」と説明しています。名前付きグループごとに共通規則をすべて繰り返すか、このサイトのようにグループを一つにします。

巡回を拒否したページが検索結果に残る。 アドレスが説明なしで表示され、リンク元のサイトが使った言葉を伴う場合もあります。Googleはページを取得していないため、アドレスが存在することと、ほかのサイトが何と呼ぶかしか知りません。 `Disallow` は巡回を制御するもので、インデックスへの登録は制御しません。規則を外してページに `noindex` を付けるか、パスワードで保護します。 `noindex` を `Disallow` の奥に隠しても、決して読まれません。

隠したかったものを、ファイルが宣伝する。 `Disallow: /admin/` のような行は、礼儀正しいクローラーかどうかに関係なく、誰にでも見る場所を正確に伝えます。ファイルは設計上、公開されます。RFCのセキュリティーの節も、パスを記すと見つけられるようになるとして、代わりにアクセス制御を求めています。非公開の資料は認証で保護し、このファイルにはまったく記さないようにします。

サーバーエラーで、サイト全体の巡回が止まる。 ページは開けても、 `/robots.txt` へのリクエストが5xxを返すかタイムアウトすると、RFCに従うクローラーは、全面的な拒否として扱わなければなりません。Googleは最初の12時間、巡回を停止し、その後は最後に正常に取得できたコピーを最大30日間使います。逆の誤りは、目立ちません。Googleは4xxの応答を、429を除き、401や403も含め、ファイルが存在しないものとして扱い、すべてのアドレスが開放されます。ファイルを静的なテキストとして200で配信し、ホームページだけでなく、 `/robots.txt` 自体を監視します。

参考資料

- RFC 9309：Robots Exclusion Protocol — グループ、照合、エラー処理、キャッシュ、セキュリティーの注記を定める標準そのものです。
- robots.txtの概要 — Googleによるファイルの目的と、ページを検索結果から隠せない理由です。
- Googleによるrobots.txt仕様の解釈 — ステータスコード、キャッシュ、Googleが無視するフィールド、グループの選び方を説明します。
- noindexで検索インデックスへの登録をブロックする — ページを検索結果から除外する規則と、それがページを巡回できなければ機能しない理由です。

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

「robots.txt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/robots-txt/