SEO Technical Guide

Sitemap & robots.txt

Complete Optimisation Guide for Website SEO

■■ Sitemap XML	■ robots.txt	■ Speed Tips
Boost crawl efficiency	Control crawler access	Submit & verify fast

Version	1.0	**Audience**	Web Developers / SEO Teams
Date	April 2026	**Scope**	All websites

Table of Contents

1. Introduction

Two technical files are essential for every well-optimised website: the **XML Sitemap** and the **robots.txt** file. Together they form the communication channel between your site and search-engine crawlers. The sitemap tells crawlers *what* to index, while robots.txt tells them *where* they are (and are not) allowed to go.

Correctly configuring both files accelerates indexing of new content, prevents crawl budget being wasted on irrelevant pages, and can measurably improve search-ranking velocity for large or frequently updated sites.

File	Format	Primary Purpose	Required?
sitemap.xml	XML	Tell crawlers which URLs exist	Strongly recommended
robots.txt	Plain text	Control crawler access	Yes (even if empty)

2. XML Sitemap

An XML sitemap is a structured list of all the URLs on your website you want search engines to crawl and index. It supports metadata such as last-modified date, change frequency, and priority, giving crawlers additional signals about your content freshness.

2.1 Basic XML Sitemap Template

```
<?xml version="1.0" encoding="UTF-8"?>
<urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9">

<!-- Homepage -->
<url>
<loc>https://www.example.com/</loc>
<lastmod>2026-04-01</lastmod>
<changefreq>daily</changefreq>
<priority>1.0</priority>
</url>
```

```
  <!-- Blog post -->
  <url>
  <loc>https://www.example.com/blog/seo-tips</loc>
  <lastmod>2026-03-28</lastmod>
  <changefreq>weekly</changefreq>
  <priority>0.8</priority>
  </url>

  <!-- Product page -->
  <url>
  <loc>https://www.example.com/products/widget</loc>
  <lastmod>2026-04-10</lastmod>
  <changefreq>monthly</changefreq>
  <priority>0.7</priority>
  </url>

  </urlset>
```

XML Tag Reference

Tag	Required	Values / Notes
<loc>	Yes	Full canonical URL (https://)
<lastmod>	Optional	ISO 8601 date: YYYY-MM-DD
<changefreq>	Optional	always | hourly | daily | weekly | monthly | yearly | never
<priority>	Optional	0.0 – 1.0 (default 0.5); homepage typically 1.0

■■ **Keep each sitemap under 50,000 URLs and 50 MB uncompressed. For larger sites, use a Sitemap Index file (Section 2.4).**

2.2 Image Sitemap Extension

Extend your sitemap with image metadata to help Google index your images and surface them in Google Image Search.

```
<url>
<loc>https://www.example.com/recipes/chocolate-cake</loc>
<image:image>
<image:loc>https://www.example.com/images/choc-cake.jpg</image:loc>
<image:caption>Decadent triple-layer chocolate cake</image:caption>
<image:title>Chocolate Cake Recipe</image:title>
</image:image>
</url>
```

2.3 Video Sitemap Extension

```
<url>
<loc>https://www.example.com/video/intro</loc>
<video:video>
<video:thumbnail_loc>https://example.com/thumb.jpg</video:thumbnail_loc>
<video:title>Product Introduction</video:title>
<video:description>Watch how our product works.</video:description>
<video:content_loc>https://example.com/video/intro.mp4</video:content_loc>
<video:duration>180</video:duration>
<video:publication_date>2026-01-15</video:publication_date>
</video:video>
</url>
```

2.4 Sitemap Index File

When your site has more than 50,000 URLs, split sitemaps into groups and reference them from a single Sitemap Index file.

```
<?xml version="1.0" encoding="UTF-8"?>
<sitemapindex xmlns="http://www.sitemaps.org/schemas/sitemap/0.9">

<sitemap>
<loc>https://www.example.com/sitemap-pages.xml</loc>
<lastmod>2026-04-15</lastmod>
</sitemap>

<sitemap>
<loc>https://www.example.com/sitemap-blog.xml</loc>
<lastmod>2026-04-20</lastmod>
</sitemap>

<sitemap>
<loc>https://www.example.com/sitemap-products.xml</loc>
<lastmod>2026-04-18</lastmod>
</sitemap>

</sitemapindex>
```

■ **Place sitemap-index.xml in your root domain, then submit only this single file to Google Search Console.**

3. robots.txt — Configuration

The robots.txt file lives at **https://www.example.com/robots.txt** and is the first file search-engine bots request when they visit your site. It instructs crawlers which paths they may and may not access.

> **■■ robots.txt is advisory, not enforceable. Malicious bots ignore it. Use server-level authentication or noindex meta tags for true security.**

3.1 Core Directives

Directive	Description	Example
User-agent	Specifies which bot the rules apply to (* = all)	User-agent: *
Disallow	Paths the bot must NOT crawl	Disallow: /admin/
Allow	Overrides a Disallow for a sub-path	Allow: /public/
Sitemap	Absolute URL of your sitemap	Sitemap: https://example.com/sitemap.xml
Crawl-delay	Seconds to wait between requests (not Google)	Crawl-delay: 10

3.2 Common Configuration Patterns

General website (recommended baseline)

```
# ■■ Allow all bots ■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■
User-agent: *
Disallow: /admin/
Disallow: /login/
Disallow: /checkout/
Disallow: /cart/
Disallow: /private/
Disallow: /tmp/
```

```
Disallow: /*?*sessionid= # block session URLs
Allow: /

# ■■ Sitemap location ■■■■■■■■■■■■■■■■■■■■■■■■■■■■
Sitemap: https://www.example.com/sitemap.xml
```

Block AI training crawlers

```
User-agent: GPTBot
Disallow: /

User-agent: Google-Extended
Disallow: /

User-agent: CCBot
Disallow: /

User-agent: anthropic-ai
Disallow: /
```

Block all bots except Googlebot

```
User-agent: *
Disallow: /

User-agent: Googlebot
Allow: /
Sitemap: https://www.example.com/sitemap.xml
```

3.3 robots.txt for Popular CMS Platforms

WordPress

```
User-agent: *
Disallow: /wp-admin/
Allow: /wp-admin/admin-ajax.php
Disallow: /wp-includes/
Disallow: /wp-content/plugins/
Disallow: /wp-content/cache/
Sitemap: https://www.example.com/sitemap_index.xml
```

Shopify

```
User-agent: *
Disallow: /admin
Disallow: /cart
Disallow: /orders
Disallow: /checkouts/
Disallow: /checkout
Disallow: /account
Disallow: /collections/*sort_by*
Disallow: /*/collections/*sort_by*
Sitemap: https://www.example.com/sitemap.xml
```

4. Submit & Verify in Google Search Console

After creating both files, submit them to Google Search Console (GSC) to accelerate indexing and monitor any errors.

Step 1	**Open GSC** Go to search.google.com/search-console and select your property.
Step 2	**Add Sitemap** Navigate to Indexing › Sitemaps and paste your sitemap URL, then click Submit.
Step 3	**Check Status** Refresh after a few minutes. Status should read 'Success'. Note how many URLs were discovered vs indexed.
Step 4	**robots.txt Tester** Under Settings › robots.txt, paste your robots.txt content and test individual URLs to verify allow/disallow rules.
Step 5	**Monitor Coverage** Check Indexing › Pages weekly. Fix 'Excluded' URLs caused by incorrect robots.txt Disallow rules.

■ **Ping Google directly after publishing new content: https://www.google.com/ping?sitemap=https://www.example.com/sitemap.xml**

5. Checklist & Quick Reference

Area	Action Item	Done?
Sitemap	sitemap.xml exists at domain root	■
Sitemap	Only canonical, indexable URLs included	■
Sitemap	lastmod dates are accurate and up to date	■
Sitemap	File under 50,000 URLs / 50 MB	■
Sitemap	Submitted to Google Search Console	■
Sitemap	Auto-regenerated on content publish (via plugin/hook)	■

robots.txt	File exists at https://domain.com/robots.txt	■
robots.txt	Sitemap URL declared inside robots.txt	■
robots.txt	Admin/login paths blocked for all bots	■
robots.txt	No accidental block of CSS/JS (breaks rendering)	■
robots.txt	Tested with GSC robots.txt Tester	■
Monitoring	Coverage report checked monthly in GSC	■
Monitoring	Sitemap errors resolved within 7 days	■

Quick-link your sitemap in robots.txt

Always add `Sitemap: https://www.example.com/sitemap.xml` at the bottom of robots.txt. This ensures every crawler — including those that skip Search Console — can discover your sitemap automatically.