Brandella Journal

FREE SEO AND AEO GROWTH ASSET 13 OF 36

Crawlability Checks Kit: URL Checker, Robots.txt Review, Sitemap and AI Bot Audit

Free crawlability checklist in Excel: test robots.txt, noindex, canonicals, sitemaps and AI bot access URL by URL, and see exactly why a page is not indexed.

Updated 23 November 2026
Online version and downloads: https://brandella.in/resources/seo-aeo-assets/crawlability-checks/
Free for personal and client work. By Santaji Gadepatil, Founder of Brandella Journal.

What is in this kit

Item	What it is
Excel workbook, 11 tabs	URL Checker (500 ready rows with 7 measures, 13 weighted checks, score, grade, fixes and a verdict), Robots.txt Review (150 lines), Sitemap Files, AI Bot Access for 10 bots, Dashboard with two charts, Monitoring Plan, Settings, AI Prompts, Checklist and Sources. Opens in Excel and Google Sheets.
Worked demo inside	27 URLs, a 23 line robots.txt, 6 sitemap files and 12 monitoring tasks for a fictional accounting software company, so you see every conflict and verdict working before you add your own site.
Full guide (PDF)	This complete guide as a printable PDF, to share with your developers, hosting team or clients.
22 AI prompts	Prompts to review robots.txt, plan AI bot rules, explain Page indexing reasons, clean sitemaps and decide between noindex, canonical and robots.txt, including a 5 prompt chained workflow.
Printable checklist	The whole crawlability checklist on one page, to tick off after every release, migration or CDN change.

Most crawlability checklists are a list of things to look at: robots.txt, sitemaps, canonicals, noindex. This free **crawlability checklist** is a working kit instead. An Excel URL Checker takes each page, asks whether you want it indexed, and checks every signal against that goal with 13 weighted checks, so it can say exactly what is wrong: blocked by robots.txt, noindex on a page you want, a canonical that points elsewhere, an orphan, a sitemap full of redirects, or a noindex Google can never see.

It is built for the AI search era. Your robots.txt now decides who can read your site for search, for AI answers and for AI training, and those are different bots with different rules. The kit reads your robots.txt line by line the way Google's specification describes, flags the traps (like a Googlebot group that silently drops every * rule), checks each sitemap file against Google's limits, and compares your policy for Googlebot, Google-Extended and the OpenAI, Anthropic and Perplexity bots with what your file really does.

How to use this crawlability checks kit

1. **Set the date and rules.** On Settings, set the audit date. Check the click depth rule, the lastmod allowance and the check weights; they are starter assumptions.
2. **List your URLs and your intent.** On the URL Checker, add your important pages from every template and the URLs you want kept out (filters, internal search, tags, old URLs, staging). Choose Yes or No in Should be indexed? for each.
3. **Add crawler and Search Console facts.** From your crawler: status code, blocked by robots.txt, noindex, canonical, in sitemap, internal links in and click depth. From Search Console: the closest Page indexing reason, from the report or URL Inspection.

4. **Paste robots.txt.** Paste your robots.txt into Robots.txt Review, one line per row, and add the file size. Read every High and Medium flag.
5. **Check sitemap files and bots.** Add each sitemap file with its URL count, size and Search Console status. On AI Bot Access, choose Allow or Block for each bot and add log hits if you have them.
6. **Fix by verdict.** Sort the URL Checker by Verdict. Fix every Fix now and Conflict row first, then canonicals, orphans, pages Google has not indexed and sitemap clean up.
7. **Monitor every week.** Use the Monitoring Plan: compare robots.txt each week, read the Page indexing and Sitemaps reports, and recheck after every release or CDN change.

What crawlability means in the AI search era

Crawlability is whether a crawler can reach a page and read it. Indexability is whether the page is allowed and fit to be stored in the index. A page can fail at either step, and the signals often contradict each other: a page in the sitemap that robots.txt blocks, a page you want indexed with a canonical that points somewhere else, a noindex that Google cannot see because the page is blocked.

Google's technical requirements start here: Googlebot is not blocked, the page returns HTTP 200 and it has indexable content. For AI search, Google's AI features page adds no new requirements: a page must be indexed and eligible to show a snippet, and crawling must be allowed in robots.txt and by any CDN or hosting infrastructure.

What is new is the number of bots. Each AI company now runs separate user agents for search, for model training and for fetches a user asks for, and they treat robots.txt differently:

- **Search and AI search bots** such as Googlebot, OAI-SearchBot, Claude-SearchBot and PerplexityBot decide whether you can appear and be linked in those answers. OpenAI says sites opted out of OAI-SearchBot will not be shown in ChatGPT search answers.
- **Training bots and tokens** such as GPTBot, ClaudeBot and Google-Extended control whether content is used to train models. Google says Google-Extended does not affect inclusion or ranking in Google Search.
- **User requested fetchers** such as ChatGPT-User, Claude-User and Perplexity-User visit a page because a person asked. OpenAI says robots.txt rules may not apply to ChatGPT-User, and Perplexity says Perplexity-User generally ignores robots.txt.

JavaScript matters too. Google renders JavaScript, but a Vercel and MERJ study from December 2024 found that none of the major AI crawlers it measured rendered JavaScript. The kit checks whether your main content is in the raw HTML.

What is inside the crawlability workbook

Tab	What it does	What you type
URL Checker	One row per URL. Works out parameter URLs, canonical state, the canonical target's status and indexability, what Google can do with the page, the lastmod gap and whether Google says it is indexed. Runs 13 weighted checks and gives a score, grade, fixes in order and a verdict.	URL, page type, should be indexed, status, redirect target, blocked, noindex, canonical, sitemap, lastmod, last change, links in, click depth, raw HTML, Search Console status, impressions
Robots.txt Review	Splits every line into field and value, removes comments, tracks which user agent group each rule belongs to, says whether Google supports the field and flags risky rules with a severity. Checks the 500 KiB limit.	The robots.txt lines and the file size
Sitemap Files	Checks each file for status, absolute URL, the 50,000 URL and 50 MB limits, Search Console errors, discoverability and how long since Google read it.	Sitemap URL, type, status, URLs, size, listed, submitted, status, last read
AI Bot Access	10 bots with what each vendor says they do. Finds whether each bot has its own group or uses the * group, whether it is blocked from the whole site, compares that with your policy and writes the lines to add.	Your policy and log hits per bot
Dashboard	19 key numbers, URLs by verdict and URLs failing each check, with two charts.	Nothing
Monitoring Plan	12 recurring checks with next due date and status: OK, Due soon, Overdue, Never done or follow up.	Last done date and result
Settings, AI Prompts, Checklist, Sources	Thresholds, weights, lists, 22 prompts, the checklist with progress, and where every fact comes from.	Optional tuning

The 13 URL checks and why each one matters

Each check scores 1 (pass), 0.5 (partly) or 0 (fix). A check is blank when it does not apply: most checks only run for pages you want indexed, and two run only for pages you want kept out. The weights are starter assumptions, editable on Settings.

Check	Weight	Rule	Why it matters
Not blocked by robots.txt	10	Wanted pages are allowed for Googlebot.	Googlebot not being blocked is the first technical requirement.
No noindex	10	Wanted pages have no noindex in meta robots or the X-Robots-Tag header.	A noindex keeps the page out of Search and its AI features.
Returns 200	9	Wanted pages return 200.	Google does not index 4xx URLs, and 5xx URLs that keep failing are dropped.
Canonical points to itself	7	Self = pass, missing = partly, another URL = fix.	Google recommends a self referencing canonical; a canonical elsewhere asks Google to index the other URL.
Canonical target is live	6	The canonical target returns 200, is not noindex and not blocked.	A canonical pointing to a redirect or 404 sends mixed signals.
Noindex can be seen	6	A noindex page is not blocked by robots.txt.	Google cannot read a noindex on a page it is not allowed to crawl, and a blocked URL may still be indexed without its content.
Sitemap lists only indexable URLs	5	Sitemap URLs are indexable; wanted pages missing from it = partly.	Google says to include the URLs you want to see in search results: the canonical ones.
lastmod is accurate	2	Not in the future, not older than the last real change, not more than 7 days after it.	Google uses lastmod only if it is consistently and verifiably accurate.
Linked internally	7	At least 1 internal link in.	Google finds most pages through links; an orphan relies on the sitemap alone.
Shallow click depth	4	3 clicks or less pass, 4 partly, 5 or more or unreachable fix.	Deep pages are found and refreshed less. The 3 click rule is common advice, not a Google rule.
Content in the raw HTML	4	Main content in the initial HTML = pass, only after JavaScript = partly.	Google renders JavaScript; many AI crawlers do not.
Unwanted URL kept out	5	Pages you do not want: indexed = fix, still indexable = partly.	Filters, internal search and staging pages waste crawling and can show up in search instead of your real pages.
Google has indexed it	6	The Page indexing status for a wanted page is Indexed.	The final proof. Any other status needs URL Inspection.

How the score and verdict work (with a worked example)

The score is the weighted share of checks a URL passes: the sum of weight x result for every check that applies, divided by the total weight of those checks, times 100. A is 85 or more, B 70 or more, C 55 or more, anything lower is D. The verdict is separate from the score: it names the first blocking problem in a fixed order, because one noindex matters more than five small passes.

Worked example. In the demo, BrightLedger (a fictional accounting software company) published "TDS rates 2026" at `/blog/tds-rates-2026/`. It returns 200, has a self canonical, 6 internal links, a click depth of 2, its content is in the raw HTML and it is in the blog sitemap with an accurate lastmod. But the draft's `noindex` meta tag went live with it, and the Page indexing status is Excluded by noindex.

- **Checks that apply:** 11 of 13, with a total weight of 70. Canonical target is live (no other canonical) and Unwanted URL kept out (it is a wanted page) do not apply.
- **Passes (49 points):** not blocked (10), returns 200 (9), canonical points to itself (7), noindex can be seen (6), lastmod is accurate (2), linked internally (7), shallow click depth (4) and content in the raw HTML (4).
- **Fails (0 points):** no noindex (10), sitemap lists only indexable URLs (5, because the sitemap lists a noindex page) and Google has indexed it (6).

49 of 70 is a score of **70**, grade **B**, with 3 fixes. The verdict is **Fix now: noindex on a page you want**. A B grade on a page that is completely out of Google is the reason the verdict exists. One fix solves all three: remove the noindex, then request indexing in URL Inspection.

The Googlebot group trap (second worked example)

Google's robots.txt specification says only one group is valid for a crawler: the group with the most specific user agent that matches. User agent groups and the * group are not combined.

BrightLedger's robots.txt has a sensible * group (block `/app/`, `/search` and `?sort=` URLs). Then someone added a Googlebot group with `Noindex: /blog/tag/`, `Disallow: /staging/` and `Disallow: /integrations/`. From that moment Googlebot ignores the whole * group. The Robots.txt Review flags the Googlebot line as **High** and the Noindex line as ignored, because Google supports only user-agent, allow, disallow and sitemap.

The URL Checker shows the damage:

- `/search?q=invoice` is no longer blocked for Googlebot and is now indexed: **Clean up: indexed but you do not want it**, score 0.
- `/integrations/tally/`, a landing page with 2,300 impressions, is blocked: **Fix now: blocked by robots.txt**.
- `/staging/new-homepage/` has a noindex but is blocked, so Google cannot see it: **Conflict: noindex hidden by robots.txt**.

The fix: delete the Googlebot group or copy every * rule it needs into it, drop the Noindex and `/integrations/` lines, and protect staging with a password instead of robots.txt.

Verdicts and what to do

Verdict	When	What to do
Fix now: blocked by robots.txt	A wanted page is disallowed for Googlebot.	Narrow or remove the rule; test it with a robots.txt tester.
Fix now: noindex on a page you want	Meta robots or X-Robots-Tag noindex.	Remove it in the template, plugin or server header; request indexing.
Fix now: not returning 200	A wanted page returns 3xx, 4xx or 5xx.	Fix the error, or link and list the final URL instead.
Fix: canonical points elsewhere	A wanted page canonicalises to another URL.	Make the canonical self referencing, or stop wanting this URL indexed.
Link it: orphan page	A wanted page has no internal links in.	Link it from its hub and related pages.
Investigate: Google has not indexed it	Every signal is fine but the status is not Indexed.	Use URL Inspection; improve content, merge duplicates, add links.
Conflict: noindex hidden by robots.txt	A noindex page is blocked.	Unblock it so Google can read the noindex.
Clean up: indexed but you do not want it	An unwanted URL is indexed.	Add noindex and keep it crawlable, or canonicalise or remove it.
Clean up: remove from the sitemap	An unwanted, non indexable URL is in a sitemap.	Remove redirects, 404s, noindex and canonicalised URLs from sitemaps.
Fix: canonical target not live	The canonical target is a redirect, error or noindex page.	Point the canonical at the live version.
Decide: still open to indexing	You do not want it, but nothing stops Google indexing it.	Choose noindex, a canonical or removal.
Healthy, Improve when you can, OK: kept out as planned	No blocking problem.	Work through the fixes column when you have time.

robots.txt rules the review checks

Rule	Flag in the kit
Google supports user-agent, allow, disallow and sitemap. Other fields such as crawl-delay are not supported.	Google supports it? says No for crawl-delay, noindex and unknown fields (Low).
Only the most specific matching group applies; user agent groups and * are not combined.	Googlebot group warning (High).
Paths start with / and are case sensitive; * matches any characters and $ marks the end of the URL.	Path should start with / or * (Medium).
A rule without a path is ignored.	Empty disallow: this group may crawl everything (Info).
Google enforces a 500 KiB limit and ignores the rest.	Size check at the top of the tab.
Sitemap lines use a full URL.	Sitemap URL must be absolute (Medium).
Pages need their CSS, JavaScript and fonts to render the way visitors see them.	May block files needed to render (Medium), using the path list on Settings.
Disallow: / blocks everything for that group.	High for * or Googlebot, Info for other bots (check your AI bot policy).

The first six rules come from Google's specification; the CSS and JavaScript list is an editable assumption of this kit. Google also caches robots.txt for up to 24 hours, treats a 4xx robots.txt (except 429) as no restrictions, and stops crawling for 12 hours when robots.txt returns a 5xx. A robots.txt applies only to its own host, protocol and port, so each subdomain needs its own file. Our complete robots.txt guide has examples for every rule.

Sitemaps, canonicals and noindex together

Google's sitemap guide limits each file to 50 MB uncompressed or 50,000 URLs, asks for absolute URLs, ignores `priority` and `changefreq`, and uses `lastmod` only if it is consistently and verifiably accurate. It also calls submitting a sitemap merely a hint.

Its canonical guide ranks the signals: redirects and rel canonical are strong, sitemap inclusion is weak. It says not to use robots.txt for canonicalisation, not to use noindex to choose a canonical within a site, and not to send different canonicals through different methods.

In the demo, the docs sitemap has 52,400 URLs and must be split. The old `/blog-sitemap.xml` returns 404 but is still named in robots.txt with a relative URL. A landing page sitemap is not in robots.txt, an index or Search Console, so Google may never find it. And 7 URLs in sitemaps are not indexable: a 503 page, a blocked page, the noindex blog post, a canonicalised docs page, two redirects and a 404.

Choosing the right tool: use **noindex** for pages that should stay crawlable but out of Search, a **canonical** for duplicates of a page you want indexed, a **404 or 410** for pages that are gone, and

robots.txt only to stop crawling of URL spaces that waste crawl time. Read our canonical tag implementation guide for HTML examples.

Column by column guide to the URL Checker

Column	What goes in it	Tip
URL, Page type, Should be indexed?	The full URL, one of 11 page types and your intent.	Intent is the most important column: every check depends on it.
Status code, Redirects to	From the crawler.	Use the status Googlebot gets, not your logged in browser.
Blocked by robots.txt?	Yes or No for Googlebot.	Test with a robots.txt tester or your crawler's robots report.
Noindex	No, Meta robots, X-Robots-Tag or Both.	Check the HTTP headers too; CDNs and servers add X-Robots-Tag.
Canonical URL	The full canonical URL, or blank if there is none.	Must match the URL exactly to count as Self, including the trailing slash.
In XML sitemap?, Sitemap lastmod, Last real content change	From the sitemap and your CMS.	The last real change is when the main content changed, not the footer.
Internal links in, Click depth	From the crawler.	Leave click depth blank when the crawler cannot reach the page.
Main content in raw HTML?	Yes or No.	Compare view source with the rendered page.
Search Console status	The closest Page indexing reason.	URL Inspection shows it per URL.
Grey columns	Measures, checks, score, grade, fixes and verdict.	Sort or filter by Verdict.

Your first 30 days with this kit

- **Week 1: map intent.** List 50 to 200 URLs: every template, your top pages and every kind of URL you want kept out. Mark Should be indexed? for each.
- **Week 2: fix blockers.** Clear every Fix now and Conflict row. Review robots.txt flags and fix the High ones on the same day.
- **Week 3: clean signals.** Fix canonicals, remove non indexable URLs from sitemaps, split oversized files and link every orphan.
- **Week 4: bots and monitoring.** Set your AI bot policy, add the missing lines, check your CDN rules and logs, and start the weekly Monitoring Plan.

Then run the full technical SEO audit kit for speed, mobile, HTTPS and structured data. Our guide to the URL Inspection tool shows how to read each indexing status.

The chained prompt workflow (from robots.txt to fixes)

Prompts 18 to 22 below are built to run in order in one AI chat:

1. **Brief the assistant** on your site, your URL types and which ones you want indexed.
2. **Paste robots.txt** and get the effective rules for each bot that matters.
3. **Paste the URL Checker rows** with verdicts and get the root causes grouped.
4. **Pressure test the fixes** so nothing you want gets blocked by accident.
5. **Get the change list**: the new robots.txt, sitemap changes and rows for your tracker.

AI assistants cannot fetch your robots.txt reliably or see your Search Console. Paste the real files and exports, and test every changed rule with a robots.txt tester before you publish it.

Crawlability mistakes to avoid

- **Using robots.txt to remove pages.** A blocked URL can still be indexed without its content, and Google cannot see a noindex on it.
- **A bot group that forgets the * rules.** Any named group replaces the * group for that bot.
- **Noindex or crawl-delay in robots.txt.** Google does not support them there.
- **Sitemaps full of redirects and 404s.** List only the canonical, indexable URLs you want in Search.
- **lastmod on every URL set to today.** If lastmod is not accurate, Google stops trusting it.
- **Copying an AI block list.** Block training bots if you want, but check you are not blocking AI search bots you want citations from.
- **Forgetting the CDN.** Bot protection can block crawlers your robots.txt allows. Check logs for 403s to bots.
- **Staging sites protected only by robots.txt.** Use a password. Our guide to XML sitemaps in the AI search era covers what else belongs in your crawl setup.

AI prompts

Replace the parts in [square brackets]. Always check the output against real data.

1. Explain my robots.txt

```
Here is my robots.txt: [paste]. For each group, list the user agents and the rules.
Then tell me, for Googlebot, Bingbot, GPTBot,
OAI-SearchBot, ClaudeBot and PerplexityBot, which group each one follows and what it
may not crawl. Remember that only the most
specific matching group applies.
```

2. Test URLs against robots.txt

```
Here is my robots.txt: [paste] and these URLs: [paste list]. For Googlebot, say whether each URL is allowed or blocked and which rule
decides it. Use the longest matching path, and the least restrictive rule when an allow and a disallow are the same length.
```

3. Write robots.txt for my policy

```
My site: [domain]. Keep these out of crawling: [paths]. Bot policy: allow search and AI search bots, block AI training bots
[or your policy]. Write a clean robots.txt with only user-agent, allow, disallow and sitemap lines, and explain each group.
```

4. Find the Googlebot group trap

```
Here is my robots.txt: [paste]. Does any named group (for example Googlebot) lose rules that only exist in the * group? List the
rules each named group is missing and give me corrected groups.
```

5. Clean my sitemap

```
Here are the URLs in my sitemap with status code, noindex and canonical: [paste]. List every URL that should not be in the sitemap and
why, and any important URL from this list of pages I want indexed that is missing: [paste].
```

6. Fix lastmod logic

```
My CMS sets lastmod like this: [describe]. Here are 20 URLs with lastmod and the real last content change: [paste]. Is our lastmod
accurate? Write the rule a developer should implement so lastmod only changes when the main content changes.
```

7. Split a large sitemap

```
My sitemap [name] has [number] URLs and is [size] MB. Here is our URL structure: [describe]. Propose how to split it into files under
50,000 URLs and 50 MB, how to name them and what the sitemap index should look like.
```

8. Explain a Page indexing reason

```
Search Console shows [reason, for example Crawled, currently not indexed] for these URLs: [paste with page type and a short summary].
Explain what the reason means, the 5 most likely causes for pages like these, and what to check in URL Inspection.
```

9. Crawled or discovered but not indexed

```
These pages are not indexed: [paste URL, title, word count, internal links in,
impressions]. Which look like a content quality problem,
which like a discovery problem and which like duplicates? Suggest improve, merge, link,
noindex or remove for each.
```

10. Noindex, canonical or robots.txt?

```
I want to handle these URL types: [paste list, for example filter URLs, internal
search, tag pages, print versions, old campaign pages].
For each, should I use noindex, a canonical, a 404 or 410, a redirect, or a robots.txt
disallow? Explain each choice in one line.
```

11. Parameter URL plan

```
My site creates URLs with these parameters: [paste examples with ?sort=, ?color=, ?
page= and so on]. Which create real unique content,
which are duplicates and which waste crawling? Give me a rule per parameter.
```

12. Canonical conflicts

```
Here are URLs with their canonical, status code, sitemap status and how internal links
point to them: [paste]. Find every conflict
and tell me which URL should be the canonical for each group.
```

13. JavaScript content check

```
Here is the raw HTML of [URL]: [paste view source] and the visible text of the rendered
page: [paste]. Which important content, links,
title or meta tags exist only after JavaScript? What should be server rendered first?
```

14. AI bot policy

```
Our goals: [for example appear and be cited in AI search, but not be used for
training]. Explain the difference between search bots,
training bots and user requested fetchers for OpenAI, Anthropic, Perplexity and Google,
and which robots.txt tokens match our goals.
Tell me to confirm every token on each vendor's own crawler page.
```

15. Read bot traffic in my logs

```
Here is a summary of my server logs for 30 days: user agent, hits, status codes and top
paths: [paste]. Which bots get errors, which
crawl pages they should not, and which important sections they never reach? Remind me
to verify bots by IP or reverse DNS.
```

16. CDN and firewall check

```
Our CDN or firewall is [provider] with these bot settings: [describe]. Our logs show
[403 or challenge counts by user agent]. Could we
be blocking Googlebot or AI search bots we want? What should we change and how do we
test it safely?
```

17. Migration crawl checks

```
We are moving from [old setup] to [new setup] on [date]. Write the crawlability checks
for before launch, launch day and the first
4 weeks: robots.txt, sitemaps, redirects, canonicals, noindex, status codes and Search
Console reports.
```

18. Workflow 1 of 5: brief the assistant

```
You are my crawlability reviewer. My site: [domain] on [CMS or framework]. URL types I
want indexed: [list]. URL types I want kept out:
[list]. Bot policy: [policy]. Only use Google's robots.txt rules: one group per
crawler, longest path wins. Confirm, then wait.
```

19. Workflow 2 of 5: robots.txt

```
Here is my robots.txt: [paste]. For Googlebot and each bot in my policy, give the group
it follows and its effective rules.
List every rule that conflicts with my lists of URLs to index and to keep out.
```

20. Workflow 3 of 5: the URL Checker rows

```
Here are my URL Checker rows with URL, should be indexed, indexability, Search Console
status, fixes and verdict: [paste]. Group the
problems by root cause (one robots rule, one template, one plugin) and rank the groups
by how many important URLs they affect.
```

21. Workflow 4 of 5: pressure test

```
For each fix you suggest, tell me what else it could change. Would a robots.txt edit
block anything I want indexed? Would a noindex hit
pages that get traffic? Remove or adjust any risky fix.
```

22. Workflow 5 of 5: the change list

```
Give me the final change list: the full corrected robots.txt, the sitemap changes, and
one row per remaining fix with tabs between
columns: URL or pattern, Problem, Fix, Owner (SEO, Developer, Content or Hosting), How
to validate.
```

Checklist

- [ ] robots.txt returns 200, is under 500 KiB and uses only user-agent, allow, disallow and sitemap

- [ ] No page you want indexed is blocked by robots.txt, for Googlebot or for the AI search bots you allow
- [ ] Any named bot group repeats every * rule that bot still needs
- [ ] CSS, JavaScript and fonts needed to render pages are not blocked
- [ ] No noindex in meta robots or X-Robots-Tag on pages you want indexed
- [ ] Pages with noindex are not blocked, so Google can see the noindex
- [ ] Every page you want indexed returns 200 and has a self referencing canonical
- [ ] Canonicals point to live, indexable URLs, never to redirects or 404s
- [ ] Sitemaps list only 200, indexable, canonical URLs, each file under 50,000 URLs and 50 MB
- [ ] Sitemaps use absolute URLs, are listed in robots.txt or submitted, and lastmod is accurate
- [ ] Every page you want indexed has internal links and sits within 3 clicks of the home page
- [ ] Main content is in the raw HTML, not only after JavaScript runs
- [ ] Filter, search, tag, staging and old URLs are handled with noindex, canonicals, 404s or passwords on purpose
- [ ] Your policy for every search and AI bot matches what robots.txt and your CDN really do
- [ ] The Page indexing and Sitemaps reports are checked weekly and robots.txt changes are compared

Frequently asked questions

Is this crawlability checklist free?

Yes. The Excel workbook, PDF guide, prompts and checklist are free to download and use for your own sites and for clients, with no sign up.

What is the difference between crawlability and indexability?

Crawlability is whether a crawler can reach and read a page. Indexability is whether the page is allowed and fit to be stored in the index. A page can be crawlable but carry a noindex, or be blocked by robots.txt and still appear in Google without its content.

Should I block a page with robots.txt or noindex?

To keep a page out of Google Search, use noindex and leave the page crawlable so Google can read it. Use robots.txt to stop crawling of URL spaces that waste crawl time. Google says a URL disallowed in robots.txt may still be indexed without its content.

Does Google follow crawl-delay or noindex in robots.txt?

No. Google supports user-agent, allow, disallow and sitemap, and says other fields such as crawl-delay are not supported. Some other crawlers, such as Anthropic's, say they support Crawl-delay.

Should I block AI bots in robots.txt?

That is a business choice. Training bots (GPTBot, ClaudeBot, the Google-Extended token) and AI search bots (OAI-SearchBot, Claude-SearchBot, PerplexityBot) are separate, so you can block training and still appear in AI search. Google says Google-Extended does not affect Google Search.

Why is my page not indexed when everything looks fine?

If the page is crawlable, returns 200, has no noindex and a self canonical, the reason is usually quality, duplication or discovery. The kit shows this as Investigate: Google has not indexed it. Check URL Inspection, improve the content, add internal links and merge near duplicates.

Sources

- Google Crawling Infrastructure: How Google interprets the robots.txt specification
  https://developers.google.com/crawling/docs/robots-txt/robots-txt-spec
- Google Search Central: Build and submit a sitemap
  https://developers.google.com/search/docs/crawling-indexing/sitemaps/build-sitemap
- Google Search Central: How to specify a canonical URL
  https://developers.google.com/search/docs/crawling-indexing/consolidate-duplicate-urls
- Google Crawling Infrastructure: Google's common crawlers (including Google-Extended)
  https://developers.google.com/crawling/docs/crawlers-fetchers/google-common-crawlers
- OpenAI: Overview of OpenAI crawlers
  https://developers.openai.com/api/docs/bots
- Anthropic: Does Anthropic crawl data from the web, and how can site owners block the crawler?
  https://support.claude.com/en/articles/8896518-does-anthropic-crawl-data-from-the-web-and-how-can-site-owners-block-the-crawler

More free templates, checklists and prompts: https://brandella.in/resources/seo-aeo-assets/