TECHNICAL SEO AUDIT - SAMPLE REPORT

books.toscrape.com

Crawl, speed and structure audit with a prioritized fix list

Prepared by theaio (AIO, theaio.co) - crawled 2026-10-30

This is a sample, and the site is not a client. books.toscrape.com is a public practice site ("Books to Scrape - Sandbox") built so people can practice crawling it. It is not meant to rank. We used it because it is safe to crawl and shows what the audit surfaces. Findings describe a sandbox, not anyone's business. The crawl was read-only, one request at a time with a delay, with a named user agent. Nothing was changed.

Executive summary

120	120	9	99
URLS CRAWLED (SAMPLE CAP)	RETURNED 200	ISSUE TYPES FOUND	LIGHTHOUSE PERF (HOME, MOBILE)

The crawler found the following. The five that matter most for a real site, in priority order:

1. **No canonical tag** - 120 of 120 pages (100%). Add a self-referencing canonical to every indexable page, then check which URL Google actually chose in Search Console.
2. **Missing meta description** - 56 of 120 pages (47%). Write a unique 70-160 character description per page template.
3. **Duplicate titles** - 7 of 120 pages (6%). Make each title unique; add the product or category name.
4. **No structured data detected** - 120 of 120 pages (100%). Add Product, BreadcrumbList and Organization markup (JSON-LD is Google's recommended format) to the matching templates, then validate it.
5. **HTML not compressed** - 119 of 120 pages (99%). Enable gzip or Brotli at the server or CDN.

Site-level checks

Check	Result	Status	Meaning
robots.txt	HTTP 404	High	Crawlers get no rules and no sitemap pointer.
sitemap.xml	HTTP 404	High	No sitemap, so new and deep pages rely on link discovery.
Missing URL returns 404	HTTP 404	Good	Correct: a real 404 is returned.
http redirects to https	HTTP 200	High	No permanent redirect from http to https; two versions of the site can be indexed.
HSTS header	max-age=0; includeSubDomains; preload	High	HSTS is missing or switched off (max-age=0), so browsers are not told to prefer https.

Page-level issues

Priority	Issue	Pages affected	Why it matters
High	No canonical tag	120 of 120 (100%)	Consolidation risk: without one, Google picks the preferred URL itself and duplicate variants (see the repeated home-page title) may be split.
High	Missing meta description	56 of 120 (47%)	Google may write its own snippet instead, which can be weaker for click-through.
High	Duplicate titles	7 of 120 (6%)	Pages compete with each other and look identical in results.
Medium	No structured data detected	120 of 120 (100%)	No JSON-LD, Microdata or RDFa was found (presence only; the markup is not validated). Without it a page is not eligible for some rich results.
Medium	HTML not compressed	119 of 120 (99%)	Larger downloads, slower first load on mobile.
Medium	Pages under 150 words (triage heuristic)	15 of 120 (12%)	Little unique text to rank on. This is a screening rule, not a ranking requirement.
Medium	Slow responses (over 1 second)	8 of 120 (7%)	Slow servers hurt crawl rate and user experience.
Low	Description too long (over 160 characters)	64 of 120 (53%)	Very long descriptions are usually truncated or rewritten by Google (about 160 characters is a triage guide).
Low	Titles over 60 characters	28 of 120 (23%)	Long titles may be truncated in results (about 60 characters is a triage guide, not a rule).

120 of 120 crawled URLs have at least one issue. Percentages use the number of HTML pages parsed (120) except status, redirect and speed checks, which use all 120 URLs.

Checked and clean: Broken pages (4xx/5xx); Pages that failed to load; Missing title tag; No mobile viewport tag; Missing H1; Images without alt text; No html lang attribute; Description too short (under 70 characters); Duplicate descriptions; More than one H1; Pages set to noindex; Pages more than 3 clicks deep; Redirected internal links.

Evidence: example URLs

Each finding in a real audit lists its URLs. Here is the first few for the top issues; the full list is in `crawl.csv`.

No canonical tag

URL path	What the crawler saw
/	(no canonical)
/index.html	(no canonical)
/catalogue/category/books_1/index.html	(no canonical)
/catalogue/category/books/travel_2/index.html	(no canonical)

Missing meta description

URL path	What the crawler saw
/	(no meta description)
/index.html	(no meta description)
/catalogue/category/books_1/index.html	(no meta description)
/catalogue/category/books/travel_2/index.html	(no meta description)

Duplicate titles

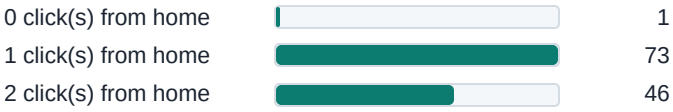
URL path	What the crawler saw
/	All products | Books to Scrape - Sandbox
/index.html	All products | Books to Scrape - Sandbox
/catalogue/category/books_1/index.html	Books | Books to Scrape - Sandbox
/catalogue/category/books/mystery_3/index.html	Mystery | Books to Scrape - Sandbox

No structured data detected

URL path	What the crawler saw
/	(no structured data detected)
/index.html	(no structured data detected)
/catalogue/category/books_1/index.html	(no structured data detected)
/catalogue/category/books/travel_2/index.html	(no structured data detected)

Crawl depth and response time

Pages by click depth



Server response

Median 579 ms, slowest 3108 ms across 120 successful fetches. 8 URLs took over one second.

Timings are measured from this crawler's location and include network time; they indicate relative problems, not user experience.

Speed and Core Web Vitals (lab test, home page)

Lighthouse 12.8.2, mobile profile with default throttling, run 2026-10-30. Lab data is a repeatable test, not what real visitors experienced; a real audit adds your Search Console field data when you share it.

Performance	Accessibility	Best practices	SEO	LCP	Total Blocking Time	CLS
99	87	75	91	1.9 s	60 ms	0

What pulled the other scores down

- **Accessibility:** Background and foreground colors do not have a sufficient contrast ratio.; Heading elements are not in a sequentially-descending order; Touch targets do not have sufficient size or spacing.
- **Best Practices:** Does not use HTTPS; Browser errors were logged to the console; Issues were logged in the `Issues` panel in Chrome Devtools
- **Seo:** Document does not have a meta description

Good thresholds for the three vitals: LCP 2.5 s or less, CLS 0.1 or less; Google replaced FID with INP (200 ms or less), which needs real-user data and is not a lab metric.

Biggest speed opportunities

Opportunity	Estimated saving
Reduce unused CSS	0.3 s

Prioritized fix list

Priority	Issue	Fix	Effort
High	No canonical tag	Add a self-referencing canonical to every indexable page, then check which URL Google actually chose in Search Console.	1 hour
High	Missing meta description	Write a unique 70-160 character description per page template.	2-4 hours
High	Duplicate titles	Make each title unique; add the product or category name.	2-4 hours
Medium	No structured data detected	Add Product, BreadcrumbList and Organization markup (JSON-LD is Google's recommended format) to the matching templates, then validate it.	1 day

Priority	Issue	Fix	Effort
Medium	HTML not compressed	Enable gzip or Brotli at the server or CDN.	30 minutes
Medium	Pages under 150 words (triage heuristic)	Add useful unique copy where a page needs to rank; do not pad pages that already serve the visitor.	varies
Medium	Slow responses (over 1 second)	Add caching or a CDN in front of the site.	1-4 hours
Low	Description too long (over 160 characters)	Trim to the first 150-160 characters that sell the click, not a pasted product blurb.	1-2 hours
Low	Titles over 60 characters	Shorten to about 60 characters with the key words first.	1-2 hours

Suggested order of work

1. **Week 1:** every Critical and High item that is a template change (title, description, viewport, H1, canonical, lang).
2. **Week 2:** redirects and broken links, compression and caching, sitemap and robots.txt.
3. **Weeks 3-4:** structured data, thin pages and internal linking; re-crawl to confirm.

Method, limits and what a paid audit adds

- **Method:** read-only crawler (`seo_audit.py`, standard library only, included), robots.txt respected, one request at a time, named user agent. Crawl capped at 120 URLs for this sample; a paid audit covers up to 100, 500 or 2,000 URLs by tier.
- **Crawl cap:** 481 more URLs were discovered but not crawled once the 120-URL cap was reached, so counts are for the crawled sample, and click depth is the shortest path this crawler found (pages beyond the cap are not counted).
- **Not covered here:** JavaScript-rendered content, validation of structured data (the crawler checks presence of JSON-LD, Microdata or RDFa only), Search Console field data, backlinks, content strategy and competitor gaps (Premium tier). Lab speed data is one page on one run.
- **What you get in a paid audit:** this same structure for your site, a CSV of the crawl, a prioritized fix list you can hand to a developer, and (Premium) a 90-day roadmap and a re-crawl after your fixes.

Scope. We do not promise rankings or traffic.