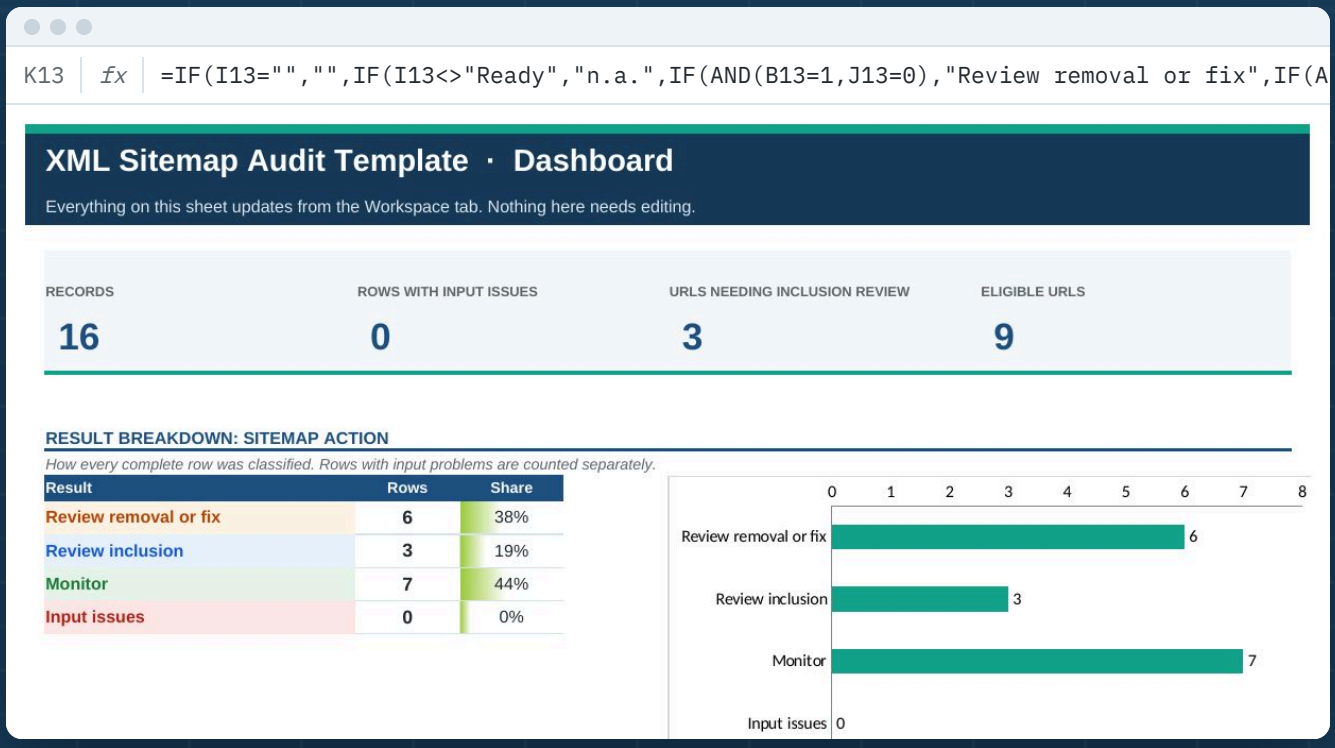
OptimizationTheory.com

Technical SEO Toolkit, resource 17

XML Sitemap Audit Template

Reconcile sitemap URLs against status, canonical, robots and noindex to see what to add, remove or fix.

Built for the DataForSEO API: On-Page API



Field guide, v3.0, November 2026

Jourdan Rombough, optimizationtheory.com

What this template decides

Which URLs should be removed from your XML sitemap (or fixed), which eligible URLs are missing from it, and which are fine.

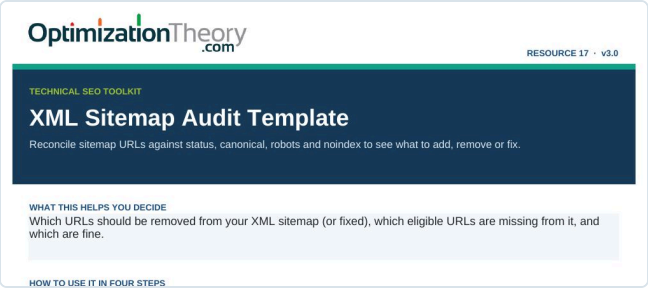
An XML sitemap is a list of URLs you're asking Google to treat as canonical and indexable. When it includes redirects, 404s, noindexed pages and parameter duplicates, you're sending mixed signals and making Search Console's sitemap reports useless.

This XML sitemap audit template reconciles each URL's sitemap membership with what a crawl observed. A URL is eligible when it returns 200, is self-canonical, crawlable, not noindexed and intended for indexing. The template flags sitemap URLs that aren't eligible (Review removal or fix) and eligible URLs missing from the sitemap (Review inclusion).

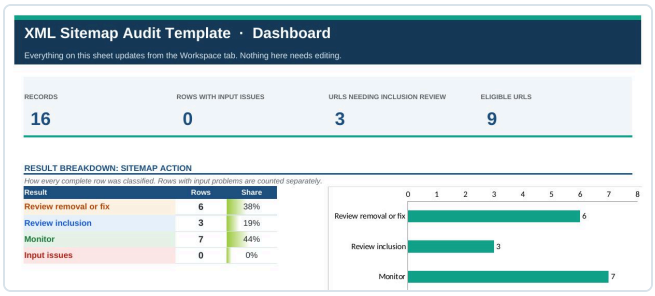
Who it's for

- Technical SEOs cleaning up sitemaps
- Developers generating sitemaps from a CMS
- Migration teams validating the new sitemap
- Sites with indexing gaps in Search Console

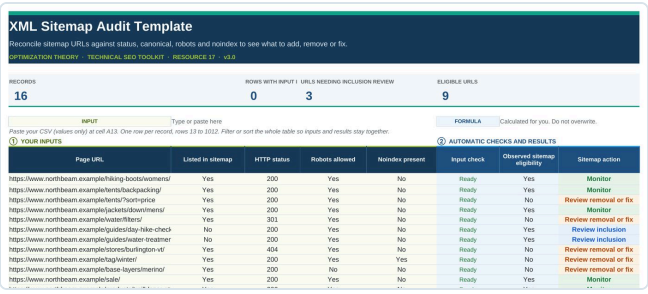
What's in the workbook



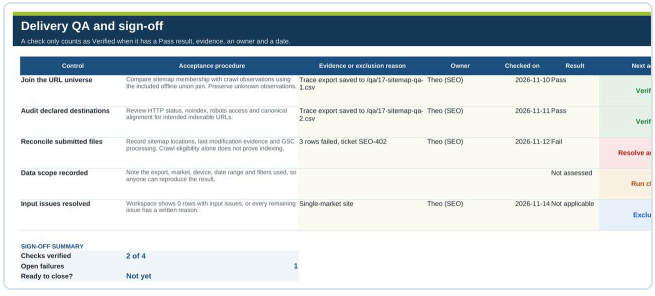
Start Here. Purpose, four-step setup and colour key.



Dashboard. KPIs, charts, Top 10 and an automatic action list.



Workspace. The working table, room for 1,000 rows.



Delivery QA. Evidence, owner and date for each validation step.

How the calculation works

Eligible = 1 when status is 200, canonical equals URL, robots allows, noindex is 0 and intended indexable is 1. In sitemap but not eligible = Review removal or fix. Eligible but not in sitemap = Review inclusion. Otherwise Monitor.

The key formula, as written in the workbook

```
=IF(I13="","",IF(I13<>"Ready","n.a.",IF(AND(B13=1,J13=0),"Review removal or fix",IF(AND(B13=0,J13=1),"Review inclusion","Monitor"))))
```

Result column	Plain-English formula	What it tells you
Observed sitemap eligibility	IF(AND(HTTP status=200,Canonical URL=Page URL,Robots allowed=1,Noindex present=0,Intended indexable=1),1,0)	Observed condition for inclusion in this sitemap plan.
Sitemap action	IF(AND(Listed in sitemap=1,Observed sitemap eligibility=0),"Review removal or fix",IF(AND(Listed in sitemap=0,Observed sitemap eligibility=1),"Review inclusion","Monitor"))	Suggests reconciliation based on intention and crawl evidence.

The results you'll see

Review removal or fix | Review inclusion | Monitor

Green means the row meets the checks, red means fix it first, amber needs a human review and blue marks an opportunity. Rows with missing or out-of-range inputs show in the Input check column and are left out of the results until fixed.

How to use it, step by step

1

Parse your sitemaps

Extract every URL from the sitemap index and child sitemaps. Mark in_sitemap = 1.

2

Add crawled URLs not in the sitemap

From a full crawl, add indexable URLs that aren't listed. Mark in_sitemap = 0.

3

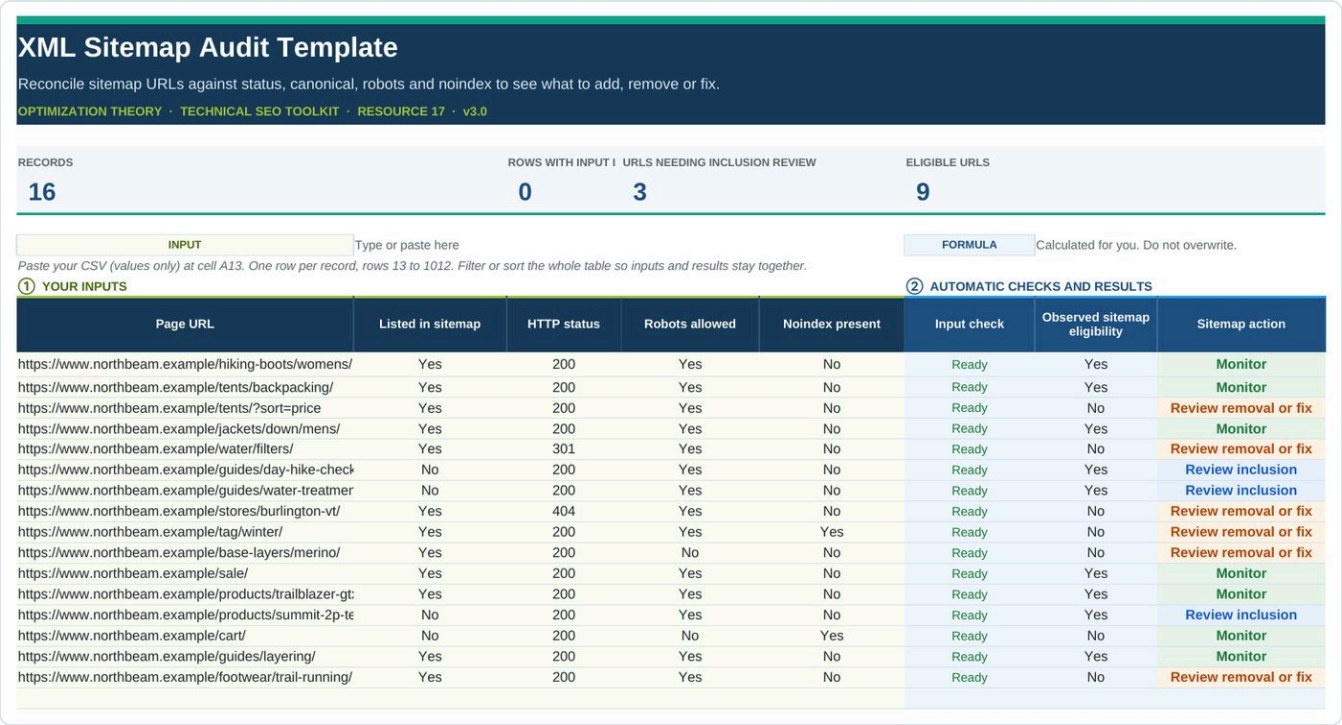
Record observed signals

Status code, canonical, robots allowed, noindex and whether you intend the page to be indexed.

4

Fix at the generator

Most sitemap errors come from the generator logic. Fix the rule, not the file.



XML Sitemap Audit Template

Reconcile sitemap URLs against status, canonical, robots and noindex to see what to add, remove or fix.

OPTIMIZATION THEORY · TECHNICAL SEO TOOLKIT · RESOURCE 17 · v3.0

RECORDS	ROWS WITH INPUT I	URLS NEEDING INCLUSION REVIEW	ELIGIBLE URLS
16	0	3	9

INPUT Type or paste here — FORMULA Calculated for you. Do not overwrite.

Paste your CSV (values only) at cell A13. One row per record, rows 13 to 1012. Filter or sort the whole table so inputs and results stay together.

① YOUR INPUTS — ② AUTOMATIC CHECKS AND RESULTS

Page URL	Listed in sitemap	HTTP status	Robots allowed	Noindex present	Input check	Observed sitemap eligibility	Sitemap action
https://www.northbeam.example/hiking-boots/womens/	Yes	200	Yes	No	Ready	Yes	Monitor
https://www.northbeam.example/tents/backpacking/	Yes	200	Yes	No	Ready	Yes	Monitor
https://www.northbeam.example/tents/?sort=price	Yes	200	Yes	No	Ready	No	Review removal or fix
https://www.northbeam.example/jackets/down/mens/	Yes	200	Yes	No	Ready	Yes	Monitor
https://www.northbeam.example/water/filters/	Yes	301	Yes	No	Ready	No	Review removal or fix
https://www.northbeam.example/guides/day-hike-check	No	200	Yes	No	Ready	Yes	Review inclusion
https://www.northbeam.example/guides/water-treatmer	No	200	Yes	No	Ready	Yes	Review inclusion
https://www.northbeam.example/stores/burlington-vt/	Yes	404	Yes	No	Ready	No	Review removal or fix
https://www.northbeam.example/tag/winter/	Yes	200	Yes	Yes	Ready	No	Review removal or fix
https://www.northbeam.example/base-layers/merino/	Yes	200	No	No	Ready	No	Review removal or fix
https://www.northbeam.example/sale/	Yes	200	Yes	No	Ready	Yes	Monitor
https://www.northbeam.example/products/trailblazer-gt	Yes	200	Yes	No	Ready	Yes	Monitor
https://www.northbeam.example/products/summit-2p-te	No	200	Yes	No	Ready	Yes	Review inclusion
https://www.northbeam.example/cart/	No	200	No	Yes	Ready	No	Monitor
https://www.northbeam.example/guides/layering/	Yes	200	Yes	No	Ready	Yes	Monitor
https://www.northbeam.example/footwear/trail-running/	Yes	200	Yes	No	Ready	No	Review removal or fix

The Workspace tab with the worked example. Lime cells are inputs; blue cells are formulas.

Fill it with DataForSEO

This template is built to run on **DataForSEO On-Page API**: a crawl of your site: URLs, status codes, canonicals, inlinks and sitemap membership.

Most of the data columns fill automatically. Add the business inputs listed below, then paste.

1. Request the data

```
POST https://api.dataforseo.com/v3/on_page/task_post
[
  {
    "target": "yourdomain.com",
    "max_crawl_pages": 1000
  }
]
```

Post the crawl task, wait for it to finish, then request on_page/pages with {"id": "<task id>", "limit": 1000} and save that response.

2. Save the full JSON response

Save the complete response, not just the task ID, next to the template folder.

3. Convert it to the input CSV

The script in the Automation folder runs offline: no API call, no credentials.

```
python3 Automation/dataforseo_import.py import --template 17 --input response.json --output my-inputs.csv
```

Fills automatically	You add
Page URL, Listed in sitemap, HTTP status, Canonical URL	Robots allowed, Noindex present, Intended indexable

Source reference fills automatically with the task ID, endpoint, location, language, device and snapshot time, so every row is traceable. Paste the finished CSV at Workspace cell A13. Estimate the cost first with the DataForSEO Cost Calculator (template 38).

Input dictionary

Keep the CSV columns in this order and paste values at Workspace cell A13. A missing observation stays blank; use 0 only for a measured zero or a verified No.

CSV field	Column	Meaning and rule	Typical source
`url`	Page URL	Exact absolute URL. Keep case and query parameters unless a documented normalization rule applies.	DataForSEO: url
`in_sitemap`	Listed in sitemap	1 when the exact URL is present in the sitemap under review. 1 = Yes, 0 = No.	DataForSEO: from_sitemap
`status_code`	HTTP status	Observed response for the sitemap URL.	DataForSEO: status_code
`canonical_url`	Canonical URL	Resolved declared canonical. Enter MISSING when absent.	DataForSEO: canonical_url
`robots_allowed`	Robots allowed	1 when the intended bot is allowed to fetch the page. 1 = Yes, 0 = No.	Your records or manual review
`noindex`	Noindex present	1 if a blocking noindex is present. 1 = Yes, 0 = No.	Your records or manual review
`intended_indexable`	Intended indexable	1 for URLs intentionally eligible for indexing. 1 = Yes, 0 = No.	Your records or manual review
`source_ref`	Source reference	Task ID, export filename and observation scope. Retain the untouched source response separately. Optional.	DataForSEO: source_ref

The worked example

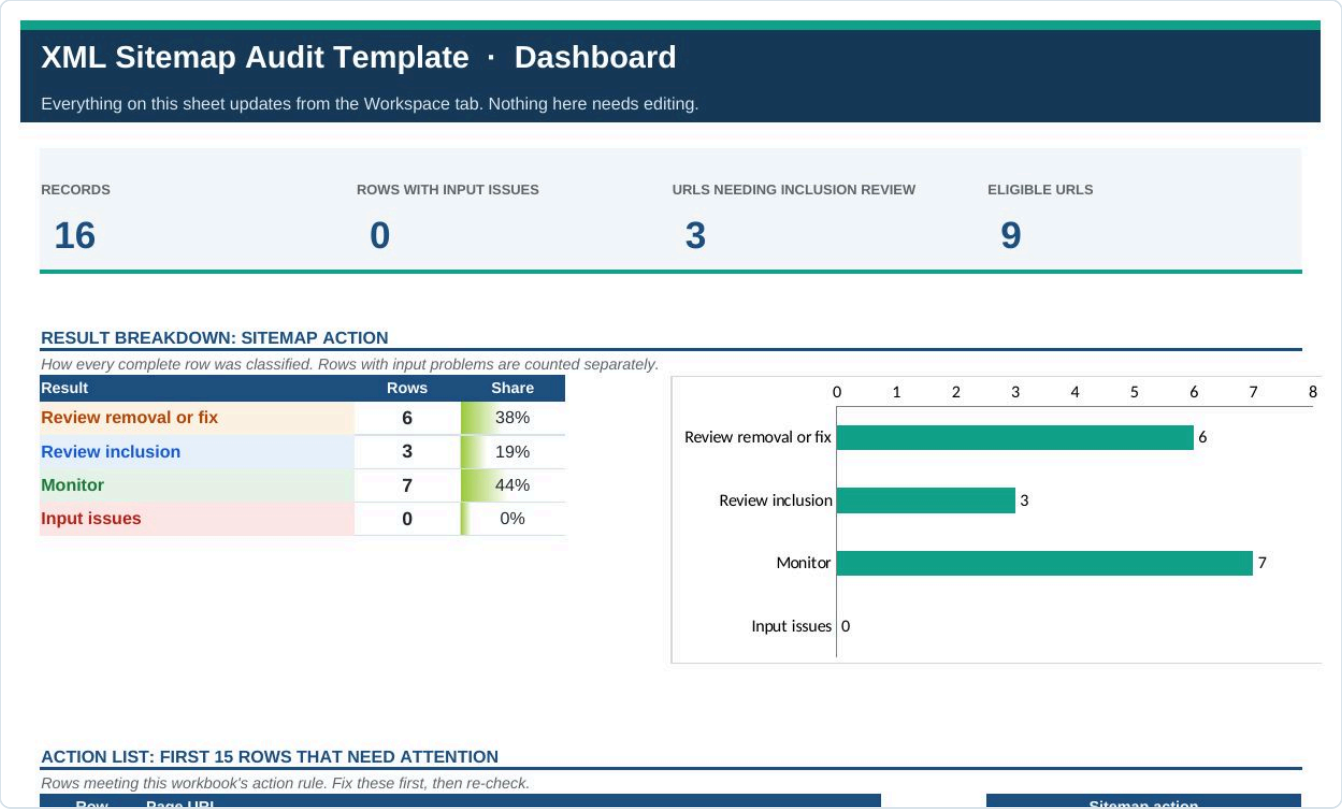
Every workbook ships filled with data from Northbeam Outdoor, a fictional outdoor-gear retailer. This one has 16 example rows.

16	0	3	9
Records	Rows with input issues	URLs needing inclusion review	Eligible URLs

The example sitemap contains a sort-parameter URL, a 301, a closed store page that 404s, a noindexed tag page and a robots-blocked category. Two valuable guides and a product page are missing entirely.

Result breakdown

Result	Rows	Share
Review removal or fix	6	38%
Review inclusion	3	19%
Monitor	7	44%
Input issues	0	0%



The Dashboard updates automatically from the Workspace tab.

Reading the result honestly

Use a union of sitemap and crawl inventories. A crawl that never discovers a URL cannot establish that it is absent from a sitemap.

Limits. DataForSEO from_sitemap records crawler discovery context. Verify actual XML membership when auditing a specific sitemap.

Mistakes to avoid

Listing redirected URLs. Sitemaps should contain final destination URLs only.

Including noindex pages. Asking Google to index a page that says noindex is a contradiction.

Updating lastmod on every build. Only change lastmod when content meaningfully changes, or Google will learn to ignore it.

Tips from the field

Split sitemaps by type (products, categories, articles) to read indexing rates per type in Search Console.

Re-run after every migration and major release.

Use the indexability audit to investigate URLs that are eligible but not indexed.

Questions people ask

What should be in an XML sitemap?

Only canonical, indexable URLs that return 200 and that you want in search results.

Should noindex pages be in the sitemap?

No. Remove them, or remove the noindex if you want them indexed.

How many URLs can a sitemap have?

Up to 50,000 URLs or 50 MB uncompressed per file. Use a sitemap index for more.

Does a sitemap guarantee indexing?

No. It helps discovery and signals preference, but Google still decides what to index.

Templates that work well with this one

Indexability Audit Template
Check status, robots, noindex and canonical on every page you want indexed, then see exactly what to fix first.

Crawl Budget Analysis Template
Group verified Googlebot requests by URL pattern, measure crawl waste and pick the right fix for each pattern.

Redirect Mapping Template
Map every old URL to its new home and verify each redirect: one permanent hop, a 200 page, no noindex, matching canonical.

Hreflang Audit Template
Validate every hreflang relationship: alternate status, self-reference, return tags, canonical match and locale codes.

Programmatic SEO Checklist
QA every generated page for verified data, unique value, indexability, links and rendering before release.

Schema Markup Audit Template
Check presence, syntax, required fields, visible-content match and policy for every structured data type you ship.

Read the full tutorial online: **optimizationtheory.com/download-view/xml-sitemap-audit-template/**
Get all six templates in the **Technical SEO Toolkit**, or all 40 in the **Complete SEO & GEO Toolkit**.