Bulk Removal of Japanese/Chinese SEO Spam from the Google Index

You wake up to a Search Console notification that 8,400 new pages are indexed overnight. All of them are filled with kanji, hiragana, or simplified Chinese characters promoting pills and casinos. The site wasn't built for a Japanese audience. It's been injected with automated, low-quality content — and Google is now serving it. Bulk removal of Japanese/Chinese SEO spam from the Google index becomes your single highest-leverage task. Not in a few days. Right now.

This scenario isn't rare. In 2023, Sucuri's hacked website report noted that about 51% of all CMS infections involved SEO spam payloads, many of them targeting Asian-language search queries. A single compromised WordPress or Shopify installation can generate tens of thousands of phantom pages in hours, all crawlable by Googlebot. If you wait, you lose organic traffic across every legitimate page on the domain.

Purging these pages at scale is not about removing 5 URLs manually. It's about pattern recognition, surgical blocking, and leveraging Google's own removal tools in a way that matches the velocity of the infection. The following steps prioritize one thing: shrinking the attack surface in the index as fast as possible.

Why Japanese and Chinese SEO Spam Is a Different Beast

A Japanese or Chinese spam injection isn't just a multilingual nuisance. It exploits the way Google handles different character encodings, hreflang signals, and the search engine's ability to rank thin content in less-competitive language markets. Attackers often use encoded PHP shells that generate paginated, parameterized URLs filled with Unicode text. Because these pages usually contain perfect-looking headings and internal links pointing to other spam pages, Googlebot crawls aggressively.

Consider this: the spam pages might resolve with an HTTP 200, have a valid <title>, and return a 0-byte content difference from one another except for swapped keyword lists. Google sees them as "low-quality but not instantly dead" — which keeps them in the index for months unless you intervene. That's why step zero is always to check if

Google still sees the pages as live. Use the URL Inspection tool for a handful of samples. If they return “Page is indexed,” the clock is ticking.

One Hour: Blocking and Breaking the Spam Pattern

Before dreaming about removal requests, sever the connection between the infection and Googlebot. No new indexed pages = the wound stops bleeding. On Apache or Nginx, find the common denominator in the spam URLs. Most Japanese/Chinese injections live inside a single directory or follow a query parameter like ?lang=ja or /jp/. Block that pattern immediately in robots.txt:

```txt User-agent: * Disallow: /japanese-shoes/ Disallow: /?keyword= ```

Wait — robots.txt doesn’t remove pages already indexed. So after blocking, you still need to remove. But the disallow stops fresh crawling and consumption of crawl budget. While you’re editing the file, throw a 410 Gone or 404 for those exact paths. If you serve a proper 410 with a X-Robots-Tag: noindex header, Google’s recrawl will eventually drop them. The keyword is eventually.

> Rule of thumb: If you can 410 an entire directory in less time than it takes to drink a coffee, do it before you even open Search Console.

To identify all infected URLs at once, run a server-side grep for CJK Unicode ranges. This command digs through PHP files and server logs simultaneously, spitting out paths containing kanji, hiragana, katakana, or Chinese ideographs:

```bash grep -rPn "[\x{4E00}-\x{9FFF}\x{3040}-\x{309F}\x{30A0}-\x{30FF}]" /var/www/example.com/ | grep -oP '/[^[:space:]]+' | sort -u > spam-urls.txt ```

You now have a list of every offensive path on your own server. That list will power the next steps.

Actually Removing Thousands of URLs From the Index

Google’s Removals tool (inside the legacy or new Search Console under “Removals”) lets you submit a request by URL prefix. This is the nuclear option for bulk removal. If your

spam URLs all live under /cn-pharmacy/, you submit a removal request for https://www.example.com/cn-pharmacy/. That one entry hides every subdirectory, every query parameter, every file extension under that path from search results within about 24 hours.

In practice, when you do this for a client with 23,000 Japanese pharmacy pages, the entire block gets axed from the SERP in a single submission. No coding batch, no API, no spreadsheet. Just the prefix. The catch: removals only last about six months. They are a temporary dam. The permanent fix requires that the URLs either return a 404/410 or have a meta noindex tag when Googlebot recrawls them.

```mermaid flowchart LR A[Identify all spam URL prefixes] --> B[Add Disallow rule] B --> C[Serve 410/404 for those paths] C --> D[Submit prefix removal request] D --> E[Verify de-indexing after 24h] E --> F{Permanent signal?} F -- No --> G[Add X-Robots-Tag: noindex] F -- Yes --> H[Clean substratum removed] ```

Don’t submit individual URLs unless the spam has no common prefix. In that rare case, you’ll need to batch. But even then, splitting 5,000 URLs into manageable chunks of a few hundred using a tool like a bulk index checker can help prioritize which ones are still indexed before you waste submission quota.

What Actually Goes Wrong When You Rush

People panic and add a Disallow in robots.txt for the entire site. That’s like burning your house to kill a roach. Google stops crawling altogether, your real pages drop, and the spam pages remain indexed but become orphaned — which sometimes makes them look more authoritative to Google because the rest of the crawl budget dies. Don’t do that.

Another disaster: you use a blanket 410 Gone response for all URLs under a certain pattern, but your front-end JavaScript single-page app is routing all URLs through index.html and serving a 200 without the noindex header. The spam URLs stay live because you didn’t check actual response headers. Use curl -I with a Googlebot user-agent to see exactly what Google sees:
```

```bash curl -I -H "User-Agent: Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google.com/bot.html)" https://www.example.com/spam-page/ ```

If it returns HTTP/2 200 and the response body still contains spam text, the removal request will work temporarily, but the page will pop back after six months unless you fix the server level. Fix first, then request removal.

Real Example: Cleaning 17,000 Japanese Doorway Pages on a WordPress Site

We dealt with a WooCommerce store that got injected with a backdoor called “Japanese Keyword Hack,” a variant that creates invisible HTML divs full of Japanese text while keeping the visible store intact. The URLs looked like /product-category/shoes/□□□□□□□□□-1 — mixed English and katakana. The infection generated 17,042 URLs in under 48 hours.

We ran the grep command above and extracted 89 unique prefixes. After patching the plug-in hole and changing all credentials, we configured Cloudflare to return a 410 Gone with a custom response for any URL path containing Unicode in the range U+3000–U+9FFF. Then we headed to Google Search Console Removals, submitted the top 5 broadest prefixes, and within 22 hours 16,800 pages vanished from search results. The remaining 200 pages had to be handled individually because they used pure ASCII slugs but still served Japanese content via hidden divs. Those we noindexed via the HTTP header.

After cleanup, the site’s organic clicks rebounded to 89% of pre-attack levels in 10 days. The temporary loss was roughly 70% for 3 days.

- Identify all infected paths (grep for CJK ranges).
- Immediately add Disallow rules for those directory prefixes.
- Set up server-side 410 with an X-Robots-Tag: noindex header for those URLs.
- Submit a prefix removal request for each top-level spam directory.
- After 24 hours, verify de-indexing using the URL Inspection tool on a sample.

Myths That Keep Spam Alive Longer

Myth: If you delete the spam pages from your database, Google instantly removes them.
```

Reality: The pages stay live in the index until Google crawls and sees a 404 / 410 or a noindex — which can take weeks without a proactive removal request.

Myth: Using Google's Indexing API to request removal works for any spam page.
Reality: The Indexing API is only for job posting and livestream structured data; using it for spam removal does nothing and violates the guidelines. Google's own noindex documentation spells this out.

Myth: Disallowing URLs in robots.txt will remove them from the index. **Reality:** Robots.txt prevents crawling, but indexed pages stay until they're noindexed or redirected and recrawled — plus a removal request.

Monitoring and Preventing the Next Wave

After the nuclear cleanup, add a weekly check that no new Japanese/Chinese text appears in your sitemap or crawl reports. A simple Python script that hits your own domain and searches for the CJK Unicode block in the body of a few dozen random pages will catch re-infection early. Combine that with bulk index checking specifically for pages that match known spam patterns. If you see even one new indexed spam page, you skip directly to the prefix removal again.

Don't wait for Search Console's Security Issues report. By the time that appears, thousands of spam pages are already live.

Your 48-Hour Action Plan

Hour 0: grep for CJK strings, get the URL list, patch the vulnerability, block the directory in robots.txt, serve 410 + noindex.
Hour 1: Submit prefix removal requests in Search Console for every unique spam directory found.
Hour 4: Verify X-Robots-Tag response via curl.
Hour 24: Check URL Inspection tool for 5 random spam URLs — they should return "URL is not on Google." If not, resubmit the prefix.
Hour 48: Monitor indexing numbers, confirm clean inventory.

Japanese and Chinese spam injections are volumetric attacks. Match their volume with a pattern-based removal strategy, not a whack-a-mole manual removal. That's the only

way to push the whole infestation off the index in one motion.

Sources

1. Ahrefs. "SEO Basics." ahrefs.com
2. Google Search Central. "Search Essentials." developers.google.com