Website Scraping Risk Audit Report

AI-assisted and manually reviewed scraping exposure review for public or authorized website data.

CLIENT	Demo Client	COMPANY	Example Retail Co.
WEBSITE URL	https://www.example-store.example		
AUDIT DATE	2026-06-22	PREPARED BY	DataCrawlPro
REVIEW METHOD	AI-assisted analysis and manual review		

OVERALL SCRAPING EXPOSURE SCORE

64/100

HIGH RISK

MEDIUM CONFIDENCE

MODERATE DIFFICULTY

High Risk means public website data appears easier to collect due to repeated patterns, visible data, predictable URLs, or unclear crawler policy.

EXECUTIVE SUMMARY

The demo website shows high scraping exposure because product data is visible in repeated templates and public page paths are predictable.

The demo website exposes repeated product listing pages, predictable pagination, visible product prices, public category paths, and structured product metadata. This makes basic product data collection relatively easy for ordinary crawlers. No private account data was reviewed. This audit focuses only on public scraping exposure.

Business impact: Product names, prices, category paths, and structured metadata could be collected or monitored if the same patterns existed on a real website.

TOP 3 IMMEDIATE CONCERNS

1. **Repeated public product listing patterns**
2. **Visible product/pricing data**
3. **Unclear AI crawler policy**

PLAIN-ENGLISH SUMMARY

Your public product and pricing pages appear easier to collect because they follow repeated patterns. This does not mean your website is hacked, but it means bots or competitors may be able to monitor public data at scale unless practical controls are added.

INCLUDED CHECKS

- Repeated product listing pages
- Predictable pagination
- Visible product prices

EXCLUDED CHECKS

- Private account data
- Authenticated areas
- Payment systems

PERMISSION NOTE

Sample/demo only. No real website was reviewed.

Disclaimer: This is a scraping exposure review, not a full cybersecurity penetration test. It does not guarantee complete protection from bots, scraping, AI crawlers, or data collection attempts.

Exposure Summary & Scraping Difficulty

WEBSITE PROFILE

TYPE	Fictional ecommerce store
INDUSTRY	Demo retail
PAGES	Product listing pages, Category pages, Product detail pages, robots.txt

Product names | Visible product prices | Product URLs | Availability labels | Category paths

SCRAPING DIFFICULTY

MODERATE 64/100

Basic product data collection appears relatively easy for ordinary crawlers because of repeated listing pages, predictable pagination, visible prices, public category

AI CRAWLER / CRAWLER POLICY

AI crawler policy is unclear. Review robots.txt and define practical crawler guidance. Robots.txt is advisory and not a security control.

PUBLIC DATA EXPOSED	LIKELY SCRAPABLE DATA	HIGH-VALUE DATA	COMPETITOR RELEVANCE	AI CRAWLER RELEVANCE
5	6	4	Medium	Medium
visible patterns	data types	commercial signals	business context	policy context

EXPOSURE HEATMAP

CATEGORY	LEVEL	SIGNAL	SHORT NOTE
Product data	HIGH		Product names and prices are visible in repeated page templates
Pricing data	HIGH		Product prices
Contact/listing data	MEDIUM		Public listing fields may be collectable.
Structured data	MEDIUM		Product prices
Sitemap/URL discovery	MEDIUM		Product listing pages
AI crawler visibility	MEDIUM		AI crawler policy is unclear.
Rate-limit visibility	UNKNOWN		Public review cannot confirm hidden server-side rate limits.

PUBLIC DATA CATEGORY BARS

PRODUCT DATA 74/100

PRICING DATA 72/100

STRUCTURED DATA 62/100

RATE-LIMIT VISIBILITY 49/100

Unknown from public review. Verify logs, CDN, WAF, or application controls.

FACTORS THAT MAKE SCRAPING EASIER

- Repeated product listing templates
- Predictable pagination
- Visible product prices
- Public category paths
- Structured product metadata

FACTORS THAT MAKE SCRAPING HARDER

- No private account data reviewed
- Public review cannot confirm hidden server-side rate limits. Verify logs, CDN, WAF, or application controls.

Key Findings

TOTAL FINDINGS	CRITICAL	HIGH	MEDIUM	LOW	TOP PRIORITY
6	0	0	6	0	2

Critical 0 / High 0 /
Medium 6 / Low 0

F-001

MEDIUM P1

Product listing pages use repeated visible patterns

Observed: The product listing pages use consistent HTML structure across categories.

Business risk: Ordinary crawlers may be able to collect product and pricing data with relatively low effort.

Recommended fix: Add server-side rate limits and monitor high-frequency category crawling.

F-002

MEDIUM P1

Pricing data is easy to identify from public pages

Observed: Visible product prices appear in repeated locations across public listing and product detail pages.

Business risk: A crawler can repeatedly collect public pricing signals if no practical throttling or monitoring exists.

Recommended fix: Monitor repeated price collection and remove nonessential pricing metadata.

F-003

MEDIUM P2

Sitemap and internal links may expose important public URLs

Observed: Public navigation, category paths, and sitemap-style discovery can reveal important product and listing URLs.

Business risk: Important public catalog URLs may be collected, monitored, or revisited frequently.

Recommended fix: Review sitemap visibility and preserve only search-critical public URLs.

F-004

MEDIUM P2

AI crawler policy appears incomplete or unclear

Observed: Robots.txt does not clearly define policy for major AI crawlers.

Business risk: AI crawlers or commercial scraping bots may access public product pages without a clearly stated policy

Recommended fix: Review robots.txt and add practical AI crawler guidance.

F-005

MEDIUM P2

Public listing/contact data may be collectable at scale

Observed: The demo website exposes repeated public listing/contact patterns that would be easy to enumerate if present on

Business risk: Contact, listing, or availability signals could be copied, monitored, or republished at scale.

Recommended fix: Reduce unnecessary public contact/listing fields and monitor bulk access.

F-006

LOW TO MEDIUM P3

Server-side rate limiting cannot be confirmed from public review

Observed: No visible public messaging confirmed server-side throttling for repeated listing or product-detail access.

Business risk: If rate controls are weak or absent, repeated public page requests may be easier to sustain.

Recommended fix: Verify CDN, WAF, application, and server logs; add limits for repeated public data requests.

Developer Fix Checklist & Next Steps

FIRST 24 HOURS
Review robots.txt for search crawlers and AI crawlers

THIS WEEK
Add monitoring and rate-limit review for repeated listing, product detail, category, and pagination

THIS MONTH
Review public APIs, feeds, sitemap exposure, and structured metadata.

RE-AUDIT TIMING
Re-audit after public exposure and monitoring changes are deployed.

CHECKLIST PRIORITY MIX
6 recommended actions

P1 - FIX FIRST	P2 - IMPROVE NEXT	P3 - MONITOR LATER
3 actions	3 actions	0 actions

P1 - FIX FIRST — 3 actions

DEVELOPER
Review robots.txt for search crawlers and AI crawlers
Add practical crawler guidance while treating robots.txt as advisory, not a security control.

DEVELOPER
Add clear AI crawler policy
Document practical crawler guidance in robots.txt and supporting policy language, while treating robots.txt as advisory.

DEVELOPER
Add rate limiting for repeated listing and product detail requests
Throttle abnormal high-frequency category, product detail, and pagination access.

P2 - IMPROVE NEXT — 3 actions

DEVELOPER
Monitor abnormal pagination and category crawling patterns
Log sequential page traversal and high-volume category access.

DEVELOPER
Avoid exposing internal IDs or unnecessary metadata in public HTML
Keep SEO-required structured data, but remove nonessential commercial signals.

DEVELOPER
Review public APIs, feeds, and sitemap exposure
Inventory public feeds, APIs, sitemap entries, and structured data sources.

P3 - MONITOR LATER — 0 actions

No P3 items in this basic report.
Monitor after priority fixes are deployed.

QUICK WIN 1
Clarify AI crawler guidance in robots.txt and public policy language
Effort and impact depend on current logs, CDN/WAF controls, and developer workflow.

QUICK WIN 2
Add monitoring for high-volume product/category page requests
Effort and impact depend on current logs, CDN/WAF controls, and developer workflow.

DELIVERY NOTE
AI-assisted and manually reviewed
This report is reviewed before client delivery and does not claim full security coverage.

LIMITATIONS

- This is a fictional demo report. It is provided only to show report format and does not represent a real client website audit.
- Demo only. Prepared for a fictional website.
- No private account data was reviewed.
- This audit focuses only on public scraping exposure for public pages / public or authorized data.

PRACTICAL NEXT STEPS

- Review robots.txt and AI crawler policy first.
- Add rate limiting and logging for repeated listing, product detail, category, and pagination requests.
- Review public APIs, feeds, sitemap exposure, and structured metadata.
- Re-audit after public exposure and monitoring changes are deployed.

Disclaimer: This is a scraping exposure review, not a full cybersecurity penetration test. It does not guarantee complete protection from bots, scraping, AI crawlers, or data collection attempts.