# Web Scraping Feasibility Analysis - Soko Glam (sokoglam.com)

## Executive Summary

**Difficulty Score: 3/10 (EASY)**

Soko Glam presents an EASY scraping target with a 100% HTTP success rate using real browser headers. The site is built on Shopify with standard Cloudflare protection that does not impede legitimate requests. HTTP scraping with authentic browser headers provides complete access to rich product data through JSON-LD structured markup, eliminating the need for browser automation.

**Key Findings:**

* \*\*HTTP Success Rate\*\*: 100% with real browser headers
* \*\*Total Products\*\*: 781 products available
* \*\*Data Quality\*\*: Complete product information in JSON-LD format
* \*\*Anti-Bot Protection\*\*: Standard Cloudflare, no aggressive blocking
* \*\*Recommended Approach\*\*: HTTP requests with authentic headers
* \*\*Proxy Requirements\*\*: Datacenter proxies sufficient

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

## Technical Reconnaissance

### HTTP Testing Methodology - Two-Phase Approach

**Phase 1: Browser Header Extraction**

Used Playwright MCP to extract authentic browser session data from https://sokoglam.com:

**Extracted Headers:**

`

User-Agent: Mozilla/5.0 (Macintosh; Intel Mac OS X 10\_15\_7) AppleWebKit/537.36 (KHTML, like Gecko) HeadlessChrome/139.0.7258.5 Safari/537.36

Accept: text/html,application/xhtml+xml,application/xml;q=0.9,\*/\*;q=0.8

Accept-Language: en-US,en;q=0.9

Connection: keep-alive

`

**Session Cookies Captured:**

* `localization=US`
* `cart\_currency=USD`
* `\_tracking\_consent` (Shopify analytics)
* `\_\_cf\_bm` (Cloudflare bot management)
* `\_shopify\_essential` (Shopify session)
* Multiple marketing cookies (Attentive, Klaviyo, Yotpo)

**Phase 2: HTTP Request Testing with Real Headers**

Applied extracted headers to HTTP requests achieving 100% success rate across all tested products.

### Site Architecture Analysis

Platform: Shopify-based e-commerce

CDN: Cloudflare with standard bot management

Response Size: ~14,347 lines of HTML per page

Server Location: Multiple regions (gcp-asia-southeast1)

### Product Data Structure

Primary Data Source: JSON-LD Structured Data

`json

{

"@context": "http://schema.org/",

"@type": "Product",

"name": "Perfect Cover BB Cream SPF 42 PA+++",

"price": "22.00",

"priceCurrency": "USD",

"brand": {"@type": "Brand", "name": "MISSHA"},

"availability": "http://schema.org/InStock",

"description": "Complete product description...",

"aggregateRating": {

"@type": "AggregateRating",

"ratingValue": "4.3",

"reviewCount": "434"

}

}

`

**Available Data Fields:**

* Product name, brand, description
* Pricing (USD) and currency
* Variant information (colors, sizes)
* Stock availability status
* Customer ratings and review count
* Product images and metadata
* SKU and GTIN identifiers

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

## Protection System Analysis

### Anti-Bot Detection Mechanisms

**1. Cloudflare Bot Management**

* \*\*Status\*\*: Present but permissive
* \*\*Challenge Rate\*\*: 0% during testing
* \*\*Behavior\*\*: Allows legitimate requests with proper headers
* \*\*Ray ID Tracking\*\*: Present in responses (`cf-ray: 98acc707dca7a5d3-DEL`)

**2. Shopify Native Protection**

* \*\*Rate Limiting\*\*: Not triggered during consecutive requests
* \*\*Session Management\*\*: Standard Shopify cookie validation
* \*\*Request Validation\*\*: Basic header verification

**3. JavaScript Challenges**

* \*\*Type\*\*: Standard Cloudflare JS challenge capability
* \*\*Trigger Rate\*\*: 0% with real browser headers
* \*\*Bypass\*\*: Not required with proper HTTP headers

### Performance Metrics

**HTTP Requests with Real Browser Headers:**

* \*\*Success Rate\*\*: 100% (8/8 test products)
* \*\*Average Response Time\*\*: 2.5 seconds
* \*\*HTTP Status\*\*: Consistent 200 responses
* \*\*Content Completeness\*\*: 100% - Full JSON-LD data present
* \*\*Rate Limiting\*\*: None observed up to 5 consecutive requests

**Rate Limiting Thresholds:**

* \*\*Rapid Requests\*\*: 5 consecutive requests successful
* \*\*No Blocking Observed\*\*: Zero blocking during 30-second test window
* \*\*Recommended Rate\*\*: 2-3 requests/second for stealth operation

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

## Difficulty Scoring Analysis

### Scoring Criteria Applied:

* \*\*HTTP Success Rate with Real Headers\*\*: 100% ✓
* \*\*Data Completeness\*\*: 100% ✓
* \*\*Anti-Bot Bypass Required\*\*: No ✓
* \*\*Rate Limiting Impact\*\*: Minimal ✓

**Result: 3/10 (EASY) - HTTP Approach Recommended**

The perfect HTTP success rate with authentic browser headers places this firmly in the EASY category, requiring no browser automation or sophisticated evasion techniques.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

## Traffic Analysis and Stealth Recommendations

### Estimated Daily Traffic

Based on domain authority and Korean beauty market position:

* \*\*Estimated Daily Visitors\*\*: 15,000-25,000
* \*\*Daily Page Views\*\*: 75,000-125,000
* \*\*Recommended Scraping Rate\*\*: <2,500 requests/day (<2% of traffic)

### Optimal Scraping Schedule

* \*\*Business Hours (EST)\*\*: Avoid 9 AM - 6 PM
* \*\*Peak Traffic\*\*: Weekends and evenings
* \*\*Recommended Windows\*\*: 2 AM - 6 AM EST
* \*\*Rate Limit\*\*: 2-3 requests/second maximum

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

## Product Inventory Analysis

### Sitemap Analysis Results

Total Products: 781 (verified via sitemap\_products\_1.xml)

URL Structure: https://sokoglam.com/products/[product-handle]

Update Frequency: Daily (as indicated in sitemap)

**Product Categories Identified:**

* Korean Skincare (cleansers, toners, serums, moisturizers)
* K-Beauty Makeup (BB creams, lip products)
* Brand Collections (MISSHA, ETUDE, Banila Co, Benton)
* Gift Sets and bundles

### Data Extraction Estimates

**Complete Inventory Scraping Time:**

* \*\*Sequential HTTP\*\*: ~32 minutes (781 products × 2.5s average)
* \*\*Parallel (5 threads)\*\*: ~6-8 minutes
* \*\*Parallel (10 threads)\*\*: ~4-5 minutes
* \*\*Browser Automation\*\*: ~45-60 minutes (estimated, not required)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

## Proxy Requirements and Recommendations

### HTTP-First Analysis Results

Based on 100% HTTP success rate with real browser headers:

**Datacenter Proxies - RECOMMENDED**

* \*\*Sufficient for this target\*\* based on testing results
* \*\*Cost-effective\*\* solution for 781 products
* \*\*No residential proxies required\*\*
* \*\*Providers\*\*: OxyLabs, Bright Data datacenter pools

**Why Datacenter Proxies Work:**

1. Real browser headers bypass basic bot detection
2. No aggressive fingerprinting observed
3. Standard Cloudflare protection allows legitimate requests
4. Cost savings significant vs residential proxies

### Proxy Configuration Recommendations

`

Rotation: Every 50-100 requests

Geographic Distribution: US-based preferred

Concurrent Connections: 5-10 maximum

Session Persistence: Not required

`

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

## Technical Implementation Recommendations

### Primary Approach: HTTP Requests with Real Headers

**Recommended Implementation:**

`python

headers = {

'User-Agent': 'Mozilla/5.0 (Macintosh; Intel Mac OS X 10\_15\_7) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/139.0.7258.5 Safari/537.36',

'Accept': 'text/html,application/xhtml+xml,application/xml;q=0.9,\*/\*;q=0.8',

'Accept-Language': 'en-US,en;q=0.9',

'Connection': 'keep-alive'

}

`

**Data Extraction Strategy:**

1. Parse JSON-LD structured data (primary source)
2. Extract OpenGraph meta tags (backup)
3. Fallback to HTML parsing if needed (rarely required)

### Rate Limiting and Error Handling

* \*\*Base Delay\*\*: 500ms between requests
* \*\*Exponential Backoff\*\*: On 429/503 responses
* \*\*Retry Logic\*\*: 3 attempts with increasing delays
* \*\*Circuit Breaker\*\*: Pause scraping on consistent failures

### Monitoring and Maintenance

* \*\*Success Rate Monitoring\*\*: Alert if <95%
* \*\*Response Time Tracking\*\*: Baseline 2.5s average
* \*\*Daily Sitemap Checks\*\*: Monitor for new products
* \*\*Header Refresh\*\*: Update browser headers monthly

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

## Risk Assessment and Mitigation

### Low-Risk Indicators

✓ 100% HTTP success rate with standard headers

✓ No CAPTCHA challenges encountered

✓ Standard Shopify/Cloudflare protection only

✓ Rich structured data available server-side

✓ Predictable URL patterns in sitemap

### Potential Risk Factors

⚠️ Cloudflare updates could change protection level

⚠️ Rate limiting may activate under high load

⚠️ IP blocking possible with aggressive scraping

### Mitigation Strategies

1. \*\*Maintain polite rate limiting\*\* (2-3 RPS maximum)
2. \*\*Regular header updates\*\* from real browser sessions
3. \*\*Monitor for protection changes\*\* in responses
4. \*\*Rotate datacenter proxies\*\* every 50-100 requests
5. \*\*Implement graceful degradation\*\* to browser automation if needed

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

## Competitive Analysis Context

### E-commerce Scraping Difficulty Comparison

* \*\*Easy (1-4/10)\*\*: Soko Glam, basic Shopify stores
* \*\*Moderate (5-7/10)\*\*: Target, Best Buy, major retailers
* \*\*Hard (8-10/10)\*\*: Amazon, eBay, heavily protected platforms

Soko Glam's protection level aligns with typical mid-tier e-commerce sites using standard Shopify + Cloudflare configuration.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

## Conclusion and Final Recommendations

### Primary Recommendation: HTTP Scraping

Soko Glam is perfectly suited for HTTP-based scraping with real browser headers, achieving 100% success rate and complete data access. Browser automation is unnecessary and would only add complexity and cost.

### Implementation Priorities

1. \*\*Use extracted browser headers\*\* for all HTTP requests
2. \*\*Deploy datacenter proxies\*\* for IP rotation
3. \*\*Implement 2-3 RPS rate limiting\*\* for stealth operation
4. \*\*Focus on JSON-LD parsing\*\* for data extraction
5. \*\*Monitor daily for sitemap updates\*\* and new products

### Success Factors

* ✅ Proven 100% HTTP success rate
* ✅ Complete product data available
* ✅ Cost-effective datacenter proxy solution
* ✅ Predictable Shopify architecture
* ✅ No browser automation complexity required

Expected Timeline: Complete 781-product scraping in under 10 minutes with proper parallel processing and rate limiting.