# Brooks Running - Web Scraping Feasibility Analysis

## Executive Summary

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

Brooks Running (https://www.brooksrunning.com/en\_us) represents an **EASY** scraping target due to excellent HTTP request compatibility with real browser headers, minimal anti-bot protection, and server-side rendered product data. HTTP requests using authentic browser headers achieve 100% success rate and complete data extraction, making this an ideal candidate for efficient, cost-effective scraping operations.

## Methodology Enhancement

**Two-Phase HTTP Testing Approach:**

1. **Phase 1: Browser Header Extraction** - Used Playwright MCP to extract authentic browser headers, user-agent strings, and session tokens from live website interaction
2. **Phase 2: HTTP Request Validation** - Used extracted real browser headers for HTTP requests, achieving significantly more accurate feasibility results than generic headers

**Extracted Browser Headers:**

User-Agent: Mozilla/5.0 (Macintosh; Intel Mac OS X 10\_15\_7) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/141.0.0.0 Safari/537.36  
Accept: text/html,application/xhtml+xml,application/xml;q=0.9,image/avif,image/webp,image/apng,\*/\*;q=0.8  
Accept-Language: en-US,en;q=0.9  
Accept-Encoding: identity (for testing; gzip/deflate supported)  
Cache-Control: no-cache  
Pragma: no-cache  
Sec-Fetch-Dest: document  
Sec-Fetch-Mode: navigate  
Sec-Fetch-Site: none  
Sec-Fetch-User: ?1  
Upgrade-Insecure-Requests: 1

## HTTP-First Viability Assessment

### ✅ HTTP Request Success Metrics

* **Success Rate with Real Browser Headers:** 100% (10/10 requests successful)
* **Data Completeness:** 100% - Full product information embedded in HTML
* **Average Response Time:** 1.6 seconds
* **Content Delivery:** Server-side rendered, no JavaScript dependency for core data
* **Status Code Consistency:** All requests returned HTTP 200

### Product Data Structure

**Complete product information available in server-rendered HTML:** - Product names, descriptions, pricing - Color variants and availability - Size options and inventory status - Product specifications (weight, drop, support type) - Customer reviews and ratings - Multiple product images - Related product recommendations - Breadcrumb navigation

### Sample Successful Requests

All tested product URLs returned complete data: - Women’s Ghost 17: https://www.brooksrunning.com/en\_us/womens/shoes/road-running-shoes/ghost-17/120431.html - Men’s Ghost 17: https://www.brooksrunning.com/en\_us/mens/shoes/road-running-shoes/ghost-17/110442.html  
- Women’s Glycerin 22: https://www.brooksrunning.com/en\_us/womens/shoes/road-running-shoes/glycerin-22/120434.html

## Website Structure Analysis

### Site Architecture

* **Platform:** Salesforce Commerce Cloud (Demandware)
* **Content Delivery:** Server-side rendering with progressive enhancement
* **URL Structure:** Clean, SEO-friendly paths
  + Pattern: /en\_us/{gender}/{category}/{subcategory}/{product-name}/{sku}.html
  + Example: /en\_us/womens/shoes/road-running-shoes/ghost-17/120431.html

### Sitemap Analysis

* **Total Products Identified:** 196 products in main sitemap
* **Sitemap Location:** https://www.brooksrunning.com/sitemap\_index.xml
* **Product Sitemap:** https://www.brooksrunning.com/en\_us/sitemap\_0-product.xml
* **Last Updated:** October 9, 2025 (current date)
* **Update Frequency:** Monthly/Weekly based on product type

### Product Categories

* Men’s and Women’s running shoes
* Men’s and Women’s apparel (tops, bottoms, outerwear)
* Sports bras and accessories
* Limited edition and collaboration products
* Unisex specialty running gear

## Anti-Bot Protection Analysis

### ✅ Minimal Protection Detected

**Low-Level Security Measures:** - Basic cookie tracking for session management - Standard server-side request validation - No aggressive fingerprinting detected - No JavaScript challenges or CAPTCHAs encountered - No IP blocking observed during testing

### Rate Limiting Assessment

**Tested with 10 consecutive requests (0.5s interval):** - No rate limiting triggers detected - Consistent HTTP 200 responses - Stable response times (1.4-2.5 seconds) - No degradation in content quality - No blocking mechanisms activated

### Content Delivery Network

* **CDN Provider:** Akamai (detected from service worker scripts)
* **Bot Detection:** Minimal/Basic level implementation
* **Content Availability:** Full product data served without restrictions

## Robots.txt Compliance

### Allowed Paths

✅ **Product pages are fully accessible** ✅ **Main category pages allowed** ✅ **Brand and content pages allowed**

### Restricted Paths (per robots.txt)

* Cart and checkout processes (/\*shopping-cart\*, /\*checkout\*)
* User accounts (/\*account\*)
* Search results with parameters (/\*prefn\*, /\*prefv\*)
* Pagination parameters (/\*start=)
* View parameters (\*view=grid\*, \*view=list\*)

**Scraping Impact:** Restrictions only affect non-product pages, no impact on product data extraction.

## Technical Specifications

### Server Response Analysis

* **HTTP Version:** HTTP/2
* **SSL/TLS:** TLS 1.3 (secure connection required)
* **Content-Type:** text/html;charset=UTF-8
* **Server:** Not disclosed (security practice)
* **Caching:** Standard browser caching headers present

### Performance Metrics

* **Time to First Byte:** ~0.8-1.2 seconds
* **Full Page Load:** ~1.4-2.5 seconds
* **Content Size:** ~150-300KB per product page
* **Image Assets:** Served from CDN (fast delivery)

## Recommended Scraping Strategy

### 🏆 PRIMARY RECOMMENDATION: HTTP Requests with Real Browser Headers

**Approach:** Direct HTTP requests using extracted authentic browser headers

**Implementation:**

headers = {  
 'User-Agent': 'Mozilla/5.0 (Macintosh; Intel Mac OS X 10\_15\_7) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/141.0.0.0 Safari/537.36',  
 'Accept': 'text/html,application/xhtml+xml,application/xml;q=0.9,image/avif,image/webp,image/apng,\*/\*;q=0.8',  
 'Accept-Language': 'en-US,en;q=0.9',  
 'Accept-Encoding': 'gzip, deflate, br',  
 'Cache-Control': 'no-cache',  
 'Pragma': 'no-cache',  
 'Sec-Fetch-Dest': 'document',  
 'Sec-Fetch-Mode': 'navigate',  
 'Sec-Fetch-Site': 'none',  
 'Sec-Fetch-User': '?1',  
 'Upgrade-Insecure-Requests': '1'  
}

**Advantages:** - ⚡ **10-50x more efficient** than browser automation - 💰 **Significantly lower infrastructure costs** - 📊 **100% data extraction success rate** - 🔒 **No blocking risks with proper headers** - ⏱️ **Fast processing** (1.6s avg response time)

**Recommended Request Pattern:** - **Rate:** 1-2 requests per second (well below trigger threshold) - **Session Management:** Rotate user sessions every 100-200 requests - **Error Handling:** Standard HTTP retry logic for 5xx errors

## Data-Driven Proxy Recommendations

### 🏆 DATACENTER PROXIES (RECOMMENDED)

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

* **Rationale:** HTTP requests with authentic headers show no blocking
* **Cost-Effectiveness:** 80-90% lower cost than residential proxies
* **Performance:** Faster response times and higher throughput
* **Providers:** OxyLabs, Bright Data datacenter pools
* **Proxy Rotation:** Optional but recommended for large-scale operations

**Testing Results Support:** Zero blocking observed with datacenter IP during testing phase.

### Residential Proxies (If Needed)

Only recommended if datacenter proxies show blocking in production: - **Trigger Condition:** >5% request failure rate with datacenter proxies - **Providers:** Bright Data, OxyLabs, Nimble - **Use Case:** Backup strategy only

## Product Count Estimation

### Total Available Products

**Current Product Count:** 196 products (as of October 9, 2025)

**Breakdown Analysis:** - **Active Products:** 196 in current sitemap - **Categories Covered:** - Running shoes (men’s and women’s) - Apparel (tops, bottoms, outerwear) - Sports bras and accessories - Limited editions and collaborations

**Seasonal Variations:** - Product count likely fluctuates ±20-30 products seasonally - New releases typically quarterly - Limited editions added monthly

### URL Pattern Analysis

**Consistent Structure Identified:** - Base: https://www.brooksrunning.com/en\_us/ - Categories: {mens|womens|featured|unisex}/ - Subcategories: {shoes|apparel}/{product-type}/ - Product: {product-name}/{sku}.html

## Traffic Analysis & Ethical Considerations

### Estimated Site Traffic

Based on brand size and e-commerce presence: - **Estimated Daily Visitors:** 50,000-100,000 unique visitors - **Peak Traffic Hours:** 9 AM - 6 PM EST - **Recommended Scraping Schedule:** Outside peak hours (6 PM - 9 AM EST)

### 10% Traffic Rule Compliance

**Scraping Volume Recommendations:** - **Maximum Daily Requests:** 5,000-10,000 requests (well under 10% threshold) - **Hourly Rate Limit:** 200-400 requests/hour during off-peak - **Respectful Crawling:** 1-2 second delays between requests

## Legal & Ethical Considerations

### Terms of Service Review

* **Public Data:** Product information, pricing, descriptions publicly displayed
* **No Login Required:** All product data accessible without registration
* **Commercial Use:** Standard e-commerce data aggregation practices
* **Robots.txt Compliance:** Full compliance with allowed paths

### Best Practices

* ✅ Respect robots.txt disallow rules
* ✅ Implement proper rate limiting (1-2 req/sec)
* ✅ Use authentic browser headers
* ✅ Monitor for blocking and adjust accordingly
* ✅ Focus only on publicly available product data

## Maintenance & Monitoring

### Regular Monitoring Points

1. **Success Rate Tracking:** Maintain >95% success rate
2. **Response Time Monitoring:** Alert if >3 second average
3. **Content Structure Changes:** Monitor HTML structure changes
4. **Blocking Detection:** Watch for unusual response patterns
5. **Sitemap Updates:** Check for new products weekly

### Evasion Strategies (If Needed)

**Current Status:** Not required based on testing **Contingency Plans:** 1. **User-Agent Rotation:** Multiple browser profiles 2. **IP Rotation:** Switch to residential proxies if needed 3. **Session Management:** Implement cookie rotation 4. **Request Timing:** Randomize request intervals

## Scalability Assessment

### Throughput Projections

**With Current Settings (Datacenter Proxies):** - **Products per Hour:** 1,800-3,600 (0.5-1 req/sec) - **Full Catalog Scrape:** 1-2 hours for all 196 products - **Daily Capacity:** Multiple full catalog refreshes possible

**Scaling Considerations:** - Multiple proxy pools for higher throughput - Parallel processing with proper rate limiting - Geographic distribution for global access

## Conclusion

Brooks Running represents an **EASY** scraping target with excellent compatibility for HTTP-based data extraction. The combination of server-side rendered content, minimal anti-bot protection, and 100% success rate with real browser headers makes this an ideal candidate for efficient, cost-effective scraping operations.

**Key Success Factors:** 1. ✅ **100% HTTP Success Rate** with authentic browser headers 2. ✅ **Complete Server-Side Data Rendering** - no JavaScript dependency 3. ✅ **Minimal Anti-Bot Protection** - basic security only 4. ✅ **Clean URL Structure** - predictable and crawlable 5. ✅ **Comprehensive Sitemap** - 196 products well-documented

**Recommended Implementation:** - **Primary Strategy:** HTTP requests with real browser headers - **Proxy Type:** Datacenter proxies (cost-effective and sufficient) - **Rate Limiting:** 1-2 requests per second - **Expected Success:** 95-100% data extraction rate - **Infrastructure Cost:** Low (HTTP-based approach)

This analysis demonstrates the effectiveness of the enhanced two-phase methodology, where authentic browser headers extracted via Playwright MCP enable highly accurate HTTP feasibility assessments, leading to optimal scraping strategies and cost savings.