SOP: Technical SEO Data Extraction Standards

*Generated: 03/09/2025 14:16*

# SOP: Technical SEO Data Extraction Standards

|  |  |
| --- | --- |
| **Document ID:** | **DWS-SOP-TECHNICAL-001** |
| \*\*Version:\*\* | 1.0 |
| \*\*Status:\*\* | Final |
| \*\*Approved By:\*\* | Craig Cottle |
| \*\*Date of Issue:\*\* | 26-Aug-2025 |
| \*\*Next Review Date:\*\* | 26-Feb-2026 |

---

## 1.0 Purpose

This Standard Operating Procedure (SOP) establishes mandatory standards for technical SEO data extraction to achieve 99% accuracy in bulk website analysis whilst maintaining ethical compliance and preventing hallucinated data. This SOP addresses specific extraction challenges with page titles, meta descriptions, and comprehensive technical SEO elements through validated methodologies that ensure reliable, auditable, and ethically compliant data collection processes.

## 2.0 Scope

This SOP applies to all technical SEO data extraction activities within the Autonomous Agentic Marketing System, including:

* Page title extraction with multiple fallback strategies
* Meta description validation and extraction
* Technical SEO elements (canonical URLs, robots meta, schema markup)
* Bulk site analysis capabilities similar to Screaming Frog
* All crawler implementations (enhanced\_seo\_crawler.py, advertools\_seo\_integration.py)
* Quality assurance protocols for extracted data validation
* Ethical web scraping compliance and rate limiting protocols

## 3.0 Definitions

* \*\*Technical SEO Data:\*\* Structured information extracted from web pages including titles, meta descriptions, heading tags, canonical URLs, robots directives, schema markup, and other elements that impact search engine optimisation.
* \*\*Fallback Strategy:\*\* A hierarchical extraction method that attempts multiple techniques to ensure data capture when primary methods fail.
* \*\*Extraction Confidence Score:\*\* A numerical rating (0-100) indicating the reliability and accuracy of extracted data based on method success and validation results.
* \*\*Rate Limiting Protocol:\*\* Controlled request frequency mechanisms that ensure respectful crawling behaviour and server load management.
* \*\*Bulk Analysis:\*\* Systematic extraction and analysis of technical SEO data across multiple pages or entire websites in a single operation.
* \*\*Anti-Hallucination Validation:\*\* Cross-referencing and verification processes that prevent fabricated or inaccurate data from entering the analysis pipeline.

## 4.0 Technical SEO Data Extraction Standards

### 4.1 Page Title Extraction Standards

### \*\*Primary Extraction Method Hierarchy:\*\*

1. \*\*Standard HTML Title Tag (Priority 1):\*\*

CSS Selector: title::text  
 XPath: //title/text()  
 Validation: Non-empty, trimmed string  
 Confidence Score: 95-100

1. \*\*Open Graph Title Fallback (Priority 2):\*\*

CSS Selector: meta[property="og:title"]::attr(content)  
 XPath: //meta[@property='og:title']/@content  
 Validation: Content length >0, <150 characters  
 Confidence Score: 85-90

1. \*\*Twitter Card Title Fallback (Priority 3):\*\*

CSS Selector: meta[name="twitter:title"]::attr(content)  
 XPath: //meta[@name='twitter:title']/@content  
 Validation: Content length >0, <70 characters  
 Confidence Score: 80-85

1. \*\*Primary H1 Tag Fallback (Priority 4):\*\*

CSS Selector: h1::text  
 XPath: //h1[1]/text()  
 Validation: First H1 only, trimmed content  
 Confidence Score: 70-75

1. \*\*Semantic Title Elements Fallback (Priority 5):\*\*

CSS Selectors: .page-title::text, .entry-title::text, .post-title::text  
 Validation: Class-based title extraction  
 Confidence Score: 60-65

1. \*\*URL-Derived Title (Last Resort):\*\*

Method: URL path parsing with formatting  
 Process: Extract path → Remove hyphens/underscores → Title case  
 Confidence Score: 30-40

### \*\*Title Validation Requirements:\*\*

* \*\*Length Validation:\*\* 10-150 characters (flag outliers)
* \*\*Encoding Validation:\*\* UTF-8 compliance, special character handling
* \*\*Duplication Check:\*\* Cross-reference against previously extracted titles
* \*\*Quality Score:\*\* Calculate based on length, uniqueness, and relevance indicators

### 4.2 Meta Description Extraction Standards

### \*\*Meta Description Extraction Hierarchy:\*\*

1. \*\*Standard Meta Description (Priority 1):\*\*

CSS Selector: meta[name="description"]::attr(content)  
 XPath: //meta[@name='description']/@content  
 Validation: 50-320 characters, meaningful content  
 Confidence Score: 95-100

1. \*\*Open Graph Description (Priority 2):\*\*

CSS Selector: meta[property="og:description"]::attr(content)  
 XPath: //meta[@property='og:description']/@content  
 Validation: Content coherence, appropriate length  
 Confidence Score: 85-90

1. \*\*Twitter Description (Priority 3):\*\*

CSS Selector: meta[name="twitter:description"]::attr(content)  
 XPath: //meta[@name='twitter:description']/@content  
 Validation: Platform-specific formatting compliance  
 Confidence Score: 80-85

1. \*\*First Paragraph Extraction (Priority 4):\*\*

CSS Selector: p::text  
 Processing: Extract first substantial paragraph (>50 chars)  
 Truncation: 160 characters + ellipsis if needed  
 Confidence Score: 60-70

1. \*\*Article Excerpt Elements (Priority 5):\*\*

CSS Selectors: .excerpt::text, .summary::text, .intro::text  
 Processing: Content summarisation from semantic elements  
 Confidence Score: 50-60

### \*\*Meta Description Validation Protocols:\*\*

* \*\*Content Quality Assessment:\*\* Language detection, readability scoring
* \*\*Length Optimisation:\*\* Flag descriptions <120 or >160 characters
* \*\*Duplication Detection:\*\* Identify identical descriptions across pages
* \*\*Relevance Scoring:\*\* Content alignment with page topic analysis

### 4.3 Technical SEO Elements Extraction Standards

### \*\*Canonical URL Standards:\*\*

Extraction Method: CSS: link[rel="canonical"]::attr(href)  
Validation: Valid URL format, same-domain verification  
Fallback: Self-referencing canonical if none specified  
Quality Check: Canonical chain analysis, loop detection  
Confidence Score: Based on validity and implementation correctness

### \*\*Meta Robots Directives:\*\*

Primary: meta[name="robots"]::attr(content)  
Secondary: meta[name="googlebot"]::attr(content)  
Processing: Parse directive values (index, noindex, follow, nofollow)  
Validation: Directive syntax compliance, conflicting instruction detection  
Documentation: Flag non-standard or problematic directives

### \*\*Schema Markup Analysis:\*\*

JSON-LD: script[type="application/ld+json"]::text  
Microdata: Elements with itemscope, itemprop attributes  
RDFa: Elements with vocab, property attributes  
Validation: Schema.org compliance, structured data integrity  
Processing: Extract and categorise schema types, validate syntax

### \*\*Header Analysis Standards:\*\*

H1 Tags: Count, content extraction, uniqueness verification  
H2-H6 Tags: Hierarchical structure analysis, content outlining  
Validation: Proper heading hierarchy, multiple H1 detection  
Content Analysis: Heading relevance to page topic

### \*\*Link Analysis Protocol:\*\*

Internal Links: Same-domain link counting and analysis  
External Links: Cross-domain link identification and categorisation  
Link Attributes: rel="nofollow", target="\_blank" detection  
Anchor Text: Link text extraction and analysis  
Link Quality: Broken link detection, redirect chain analysis

## 5.0 Quality Validation Procedures (99% Accuracy Target)

### 5.1 Multi-Method Validation Protocol

### \*\*Cross-Validation Requirements:\*\*

1. \*\*Dual-Method Extraction:\*\* Run extraction using both Scrapy and advertools methods
2. \*\*Consistency Verification:\*\* ≥95% agreement between extraction methods required
3. \*\*Discrepancy Resolution:\*\* Manual review for conflicting results
4. \*\*Method Reliability Scoring:\*\* Track success rates per extraction method

### \*\*Data Integrity Checkpoints:\*\*

1. \*\*Format Validation:\*\*

* HTML structure compliance
* Character encoding verification (UTF-8)
* Special character handling validation
* Data type consistency checks

1. \*\*Content Quality Assessment:\*\*

* Language detection and validation
* Content coherence scoring
* Spam/placeholder content detection
* Relevance scoring against page content

1. \*\*Statistical Anomaly Detection:\*\*

* Outlier identification in data distributions
* Pattern recognition for potential extraction errors
* Historical comparison for consistency validation
* Industry benchmark comparison

### 5.2 Automated Quality Gates

### \*\*Stage 1: Pre-Processing Validation\*\*

* URL accessibility verification (200 status codes)
* robots.txt compliance checking
* SSL certificate validation
* Content-Type header verification

### \*\*Stage 2: Extraction Quality Scoring\*\*

* Method success rate calculation
* Confidence score assignment (0-100 scale)
* Fallback method tracking and analysis
* Error pattern identification and logging

### \*\*Stage 3: Post-Processing Validation\*\*

* Duplicate detection across extracted dataset
* Completeness scoring (% of successful extractions)
* Quality metrics calculation (length, format, content appropriateness)
* Cross-referencing with industry standards

### 5.3 Human Quality Assurance Protocol

### \*\*Review Triggers:\*\*

* Confidence scores below 85%
* Extraction method failures exceeding 5%
* Statistical anomalies in extracted data
* New website patterns or technologies encountered

### \*\*Review Process:\*\*

1. \*\*Sample Size:\*\* Minimum 10% of low-confidence extractions
2. \*\*Expert Validation:\*\* Subject matter expert review and verification
3. \*\*Improvement Documentation:\*\* Process refinement based on review findings
4. \*\*Threshold Adjustment:\*\* Confidence score calibration based on review outcomes

## 6.0 Ethical Web Scraping Guidelines and 2025 Compliance

### 6.1 Legal and Ethical Framework

### \*\*2025 Compliance Requirements:\*\*

1. \*\*Data Protection Compliance:\*\*

* GDPR compliance for EU-accessible websites
* CCPA compliance for California-based websites
* Australian Privacy Principles (APP) adherence
* Cookie policy respect and data minimisation

1. \*\*Website Terms of Service Respect:\*\*

* Terms of Service review before crawling
* Copyright and intellectual property respect
* Commercial use restrictions compliance
* User-generated content handling protocols

1. \*\*Industry Best Practices:\*\*

* Search engine ethical guidelines adherence
* Professional SEO industry standards
* Academic and research ethical protocols
* Open source community guidelines compliance

### 6.2 Technical Compliance Implementation

### \*\*robots.txt Compliance Protocol:\*\*

# Mandatory robots.txt checking before crawling  
robots\_parser = urllib.robotparser.RobotFileParser()  
robots\_parser.set\_url(f"{domain}/robots.txt")  
robots\_parser.read()  
  
# Verify crawling permission for each URL  
if not robots\_parser.can\_fetch(USER\_AGENT, url):  
 skip\_crawling(url, reason="robots.txt\_disallow")

### \*\*Crawl-Delay Respect:\*\*

* Minimum 1-second delay between requests (default)
* robots.txt crawl-delay directive compliance
* Dynamic delay adjustment based on server response times
* Peak hours avoidance (optional server-specific configuration)

### \*\*User Agent Declaration:\*\*

Standard User Agent: "DWS-SEO-Analyzer/1.0 (+https://discoverwebsolutions.com.au/bot)"  
Purpose Declaration: Technical SEO analysis for website improvement  
Contact Information: Include valid contact details for webmaster inquiries  
Transparency: Clear identification of crawling purpose and organisation

## 7.0 Rate Limiting and Respectful Crawling Protocols

### 7.1 Request Rate Management

### \*\*Base Rate Limiting Configuration:\*\*

DOWNLOAD\_DELAY = 1 # Minimum 1 second between requests  
RANDOMIZE\_DOWNLOAD\_DELAY = True # Add randomisation (0.5-1.5x base delay)  
CONCURRENT\_REQUESTS = 1 # Conservative concurrent request limit  
CONCURRENT\_REQUESTS\_PER\_DOMAIN = 1 # Strict per-domain limiting

### \*\*Adaptive Rate Limiting:\*\*

AUTOTHROTTLE\_ENABLED = True  
AUTOTHROTTLE\_START\_DELAY = 1  
AUTOTHROTTLE\_MAX\_DELAY = 10  
AUTOTHROTTLE\_TARGET\_CONCURRENCY = 2.0  
AUTOTHROTTLE\_DEBUG = True # Enable for monitoring

### \*\*Server Health Monitoring:\*\*

* Response time tracking and delay adjustment
* HTTP status code monitoring (429, 503, 500 handling)
* Server load consideration during peak hours
* Graceful degradation for overloaded servers

### 7.2 Crawling Behaviour Standards

### \*\*Session Management:\*\*

* HTTP session persistence for reduced server load
* Cookie handling for session-based websites
* Keep-alive connection utilisation
* Connection pooling for efficiency

### \*\*Error Handling Protocol:\*\*

# Retry configuration for transient failures  
RETRY\_ENABLED = True  
RETRY\_TIMES = 3  
RETRY\_HTTP\_CODES = [500, 502, 503, 504, 408, 429]  
RETRY\_BACKOFF\_MULTIPLIER = 2 # Exponential backoff

### \*\*Resource Limits:\*\*

* Maximum pages per site: 1000 (configurable)
* Maximum crawl duration: 2 hours per site
* Bandwidth consumption monitoring
* Storage space management for crawl data

## 8.0 Error Handling and Retry Mechanisms

### 8.1 Comprehensive Error Classification

### \*\*Network-Level Errors:\*\*

* \*\*DNS Resolution Failures:\*\* Retry with alternative DNS servers
* \*\*Connection Timeouts:\*\* Progressive timeout extension (5s → 10s → 20s)
* \*\*SSL/TLS Errors:\*\* Certificate validation bypassing with security logging
* \*\*Network Unreachable:\*\* Scheduled retry after connectivity verification

### \*\*HTTP Status Code Handling:\*\*

HTTP\_ERROR\_HANDLING = {  
 '4xx\_errors': {  
 404: 'log\_and\_continue', # Page not found - expected  
 403: 'retry\_with\_different\_ua', # Forbidden - try different approach  
 429: 'exponential\_backoff', # Rate limited - respect server limits  
 401: 'skip\_with\_auth\_note' # Unauthorised - requires credentials  
 },  
 '5xx\_errors': {  
 500: 'retry\_3\_times', # Internal server error - transient  
 502: 'retry\_with\_delay', # Bad gateway - network issue  
 503: 'retry\_after\_delay', # Service unavailable - temporary  
 504: 'retry\_with\_timeout\_increase' # Gateway timeout - slow server  
 }  
}

### \*\*Content-Level Errors:\*\*

* \*\*Invalid HTML Structure:\*\* Fallback to lenient parsing modes
* \*\*Character Encoding Issues:\*\* Multiple encoding detection attempts
* \*\*JavaScript-Required Content:\*\* Selective browser automation for critical pages
* \*\*CAPTCHA/Anti-Bot Protection:\*\* Graceful failure with human escalation option

### 8.2 Retry Strategy Implementation

### \*\*Exponential Backoff Algorithm:\*\*

def calculate\_retry\_delay(attempt\_number, base\_delay=1):  
 """Calculate retry delay with exponential backoff + jitter"""  
 delay = base\_delay \* (2 \*\* attempt\_number)  
 jitter = random.uniform(0.1, 0.3) \* delay # Add randomness  
 return min(delay + jitter, 300) # Cap at 5 minutes maximum

### \*\*Circuit Breaker Pattern:\*\*

class CircuitBreaker:  
 def \_\_init\_\_(self, failure\_threshold=5, recovery\_timeout=300):  
 self.failure\_threshold = failure\_threshold  
 self.recovery\_timeout = recovery\_timeout  
 self.failure\_count = 0  
 self.last\_failure\_time = None  
   
 def call\_protected\_function(self, func, \*args, \*\*kwargs):  
 if self.failure\_count >= self.failure\_threshold:  
 if time.time() - self.last\_failure\_time < self.recovery\_timeout:  
 raise CircuitBreakerOpenException("Circuit breaker is open")  
 else:  
 self.failure\_count = 0 # Reset after timeout

### 8.3 Data Recovery and Validation

### \*\*Partial Extraction Recovery:\*\*

* Save partial results before complete failure
* Resume crawling from last successful checkpoint
* Validate partial data integrity before processing
* Merge partial results with comprehensive validation

### \*\*Quality Degradation Handling:\*\*

* Define minimum acceptable data quality thresholds
* Implement graceful degradation when quality targets not met
* Provide quality warnings in output data
* Escalate to human review for critical quality failures

## 9.0 Integration Points

### 9.1 Enhanced SEO Crawler Integration (enhanced\_seo\_crawler.py)

### \*\*Technical Implementation Standards:\*\*

* \*\*Class Integration:\*\* `SEOSpider` class utilises all fallback strategies defined in this SOP
* \*\*Method Compliance:\*\* `\_extract\_title()` and `\_extract\_meta\_description()` methods implement full hierarchy
* \*\*Validation Integration:\*\* Quality scoring algorithms align with SOP confidence score requirements
* \*\*Error Handling:\*\* Scrapy-specific error handling protocols implement SOP retry mechanisms

### \*\*Configuration Alignment:\*\*

# SOP-compliant Scrapy settings  
SCRAPY\_SETTINGS = {  
 'ROBOTSTXT\_OBEY': True, # Mandatory robots.txt compliance  
 'DOWNLOAD\_DELAY': 1, # Minimum SOP delay requirement  
 'USER\_AGENT': 'DWS-SEO-Analyzer/1.0 (+https://discoverwebsolutions.com.au/bot)',  
 'CONCURRENT\_REQUESTS': 1, # Conservative crawling  
 'RETRY\_ENABLED': True, # SOP error handling  
 'RETRY\_TIMES': 3 # SOP retry limits  
}

### 9.2 Advertools Integration (advertools\_seo\_integration.py)

### \*\*Advanced Analysis Capabilities:\*\*

* \*\*Sitemap Analysis:\*\* Utilises `adv.sitemap\_to\_df()` for comprehensive sitemap evaluation
* \*\*Robots.txt Parsing:\*\* Implements `adv.robotstxt\_to\_df()` for directive analysis
* \*\*Enhanced Crawling:\*\* Uses `adv.crawl()` for robust multi-page analysis
* \*\*SERP Integration:\*\* Leverages advertools SERP analysis capabilities for competitive intelligence

### \*\*Quality Assurance Integration:\*\*

# SOP-compliant advertools configuration   
ADVERTOOLS\_CONFIG = {  
 'follow\_links': True,  
 'respect\_robots\_txt': True,  
 'custom\_settings': {  
 'USER\_AGENT': 'DWS-SEO-Analyzer/1.0 (+https://discoverwebsolutions.com.au/bot)',  
 'DOWNLOAD\_DELAY': 1,  
 'CONCURRENT\_REQUESTS': 1  
 }  
}

### 9.3 Quality Control Anti-Hallucination Protocol Alignment

### \*\*Multi-Layer Verification Implementation:\*\*

* \*\*Layer 1:\*\* Source grounding for all extracted technical data
* \*\*Layer 2:\*\* Cross-validation between Scrapy and advertools methods
* \*\*Layer 3:\*\* Confidence scoring alignment with DWS-SOP-QUALITY-001 thresholds

### \*\*British English Compliance Integration:\*\*

* \*\*Language Detection:\*\* Validate extracted content language
* \*\*Spelling Standards:\*\* Apply British English validation to content analysis
* \*\*Terminology Consistency:\*\* Ensure SEO terminology follows British conventions
* \*\*Date/Currency Formatting:\*\* Proper formatting in analysis outputs

## 10.0 Roles and Responsibilities

|  |  |
| --- | --- |
| **Role** | **Responsibilities** |
| \*\*SEO Technical Lead\*\* | Oversee extraction methodology implementation and quality standards |
| \*\*Python Developer\*\* | Implement and maintain crawler code according to SOP specifications |
| \*\*Quality Assurance Analyst\*\* | Monitor extraction accuracy, validate quality metrics, conduct sample reviews |
| \*\*Data Validation Specialist\*\* | Perform cross-validation between extraction methods, manage confidence scoring |
| \*\*Compliance Officer\*\* | Ensure ethical crawling practices, monitor legal compliance, manage robots.txt adherence |
| \*\*System Administrator\*\* | Configure rate limiting, monitor server resources, manage crawl scheduling |

## 11.0 Success Criteria and Performance Metrics

### 11.1 Primary Success Metrics

### \*\*Accuracy Targets:\*\*

* \*\*Title Extraction Success:\*\* ≥99% of pages yield non-empty titles
* \*\*Meta Description Extraction:\*\* ≥95% success rate with quality validation
* \*\*Technical SEO Elements:\*\* ≥98% successful extraction of canonical URLs, robots directives
* \*\*Cross-Method Validation:\*\* ≥95% consistency between Scrapy and advertools extraction

### \*\*Performance Standards:\*\*

* \*\*Processing Speed:\*\* Maximum 2 seconds per page average extraction time
* \*\*Error Rate:\*\* <1% unrecoverable extraction failures per crawl session
* \*\*Resource Efficiency:\*\* <100MB memory usage per 1000 pages crawled
* \*\*Compliance Rate:\*\* 100% robots.txt adherence, zero rate limit violations

### 11.2 Quality Assurance Metrics

### \*\*Monthly Quality Reviews:\*\*

* \*\*Sample Validation:\*\* Review 5% of extracted data monthly for accuracy
* \*\*Method Performance:\*\* Track success rates by extraction method and fallback usage
* \*\*Error Pattern Analysis:\*\* Identify and address recurring extraction issues
* \*\*Confidence Score Calibration:\*\* Adjust scoring algorithms based on review outcomes

### \*\*Quarterly Compliance Audits:\*\*

* \*\*Legal Compliance Review:\*\* Verify adherence to updated privacy regulations
* \*\*Ethical Guidelines Assessment:\*\* Ensure crawling practices remain industry-appropriate
* \*\*Technical Standards Update:\*\* Incorporate new web technologies and standards
* \*\*Performance Optimisation:\*\* Review and improve extraction efficiency

## 12.0 Risk Management

### 12.1 Critical Risk Mitigation

|  |  |  |  |
| --- | --- | --- | --- |
| **Risk** | **Impact** | **Probability** | **Mitigation Strategy** |
| \*\*Legal Compliance Violation\*\* | High | Low | Mandatory legal review, automated compliance checking |
| \*\*Data Extraction Failure\*\* | High | Medium | Multi-method fallbacks, circuit breaker patterns |
| \*\*Server Overload/Blocking\*\* | Medium | High | Aggressive rate limiting, respectful crawling protocols |
| \*\*Quality Degradation\*\* | Medium | Medium | Continuous monitoring, automated quality gates |
| \*\*Technology Changes\*\* | Medium | High | Regular SOP updates, flexible implementation architecture |

### 12.2 Operational Resilience

### \*\*Monitoring and Alerting:\*\*

* Real-time extraction success rate monitoring
* Automated alerts for quality threshold violations
* Server response time and error rate tracking
* Compliance violation immediate notification system

### \*\*Disaster Recovery:\*\*

* Crawl checkpoint system for large site recovery
* Redundant data storage for critical extraction results
* Alternative extraction method activation protocols
* Manual override procedures for critical failures

---

* This SOP establishes mandatory standards for all technical SEO data extraction
* Compliance with all protocols is required for system certification
* Changes require approval from SEO Technical Lead and Quality Assurance Lead
* All extraction implementations must demonstrate SOP compliance before deployment
* Monthly performance reviews mandatory with quarterly compliance audits