SOP: Page Priority Crawling Standards

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

# SOP: Page Priority Crawling Standards

|  |  |
| --- | --- |
| **Document ID:** | **DWS-SOP-CRAWLING-001** |
| \*\*Version:\*\* | 1.0 |
| \*\*Status:\*\* | Final |
| \*\*Approved By:\*\* | Craig Cottle |
| \*\*Date of Issue:\*\* | 01-Sep-2025 |
| \*\*Next Review Date:\*\* | 01-Mar-2026 |

---

## 1.0 Purpose

This Standard Operating Procedure (SOP) establishes mandatory standards for website crawling prioritization within the Autonomous Agentic Marketing System. This SOP ensures all website analysis focuses on the most important pages first, providing comprehensive coverage of critical site elements while optimizing analysis time and ensuring all essential pages receive proper SEO evaluation and documentation.

## 2.0 Scope

This SOP applies to all website crawling and analysis activities, including:

* Technical SEO analysis and metadata extraction
* Performance testing and Core Web Vitals assessment
* Accessibility testing and WCAG compliance evaluation
* Content analysis and competitive intelligence gathering
* All agents utilizing Playwright MCP and Scrapy for website analysis

## 3.0 Definitions

* \*\*Priority Page\*\*: High-value website page essential for SEO and user experience analysis
* \*\*Core Pages\*\*: Fundamental website pages present on most business websites (home, about, services, contact)
* \*\*Service Pages\*\*: Pages describing specific products, services, or solutions offered by the business
* \*\*Conversion Pages\*\*: Pages designed to drive specific user actions (contact, purchase, signup)
* \*\*Crawl Depth\*\*: Number of link levels from homepage to analyse during automated crawling
* \*\*Page Discovery\*\*: Process of identifying all important pages on a website before detailed analysis

## 4.0 Procedures

### 4.1 Procedure: Priority Page Identification Framework

Establish systematic approach to identifying the most important pages for analysis.

### \*\*Step 1: Mandatory Core Pages (Priority Level 1)\*\*

Always analyze these pages if they exist:

### Core Pages Checklist (MANDATORY)  
- [ ] \*\*Homepage\*\* (/) - Primary entry point and brand representation  
- [ ] \*\*About/About Us\*\* (/about, /about-us, /who-we-are) - Company information and credibility  
- [ ] \*\*Services/Products/Solutions\*\* (/services, /products, /solutions) - Core business offerings  
- [ ] \*\*Contact/Get In Touch\*\* (/contact, /contact-us, /get-in-touch) - Contact information and conversion  
- [ ] \*\*Features\*\* (/features, /what-we-do) - Detailed service/product features  
  
### Page Discovery Methods  
1. \*\*Navigation Menu Analysis\*\*: Extract all main navigation links  
2. \*\*Sitemap Parsing\*\*: Check /sitemap.xml for comprehensive page list   
3. \*\*Footer Link Analysis\*\*: Important pages often linked in footer  
4. \*\*Internal Link Discovery\*\*: Follow high-authority internal links from homepage

### \*\*Step 2: Business-Specific Priority Pages (Priority Level 2)\*\*

Industry and business model specific pages:

### Service-Based Business Pages  
- [ ] \*\*Service Category Pages\*\* (e.g., /orthodontics, /invisalign, /braces)  
- [ ] \*\*Location Pages\*\* (e.g., /canberra-orthodontist, /belconnen-clinic)  
- [ ] \*\*Pricing/Cost Information\*\* (/pricing, /costs, /investment)  
- [ ] \*\*FAQ/Frequently Asked Questions\*\* (/faq, /common-questions)  
- [ ] \*\*Testimonials/Reviews\*\* (/testimonials, /reviews, /success-stories)  
- [ ] \*\*Team/Doctor Profiles\*\* (/team, /doctors, /specialists)  
  
### E-Commerce Business Pages   
- [ ] \*\*Product Category Pages\*\* (/category/\*, /shop/\*)  
- [ ] \*\*Popular Product Pages\*\* (best sellers, featured products)  
- [ ] \*\*Checkout Process\*\* (/cart, /checkout, /payment)  
- [ ] \*\*Account/Login Pages\*\* (/account, /login, /register)  
- [ ] \*\*Shipping/Returns\*\* (/shipping, /returns, /policy)  
  
### Professional Services Pages  
- [ ] \*\*Case Studies\*\* (/case-studies, /portfolio, /work)  
- [ ] \*\*Resources/Blog\*\* (/blog, /resources, /insights)  
- [ ] \*\*Careers\*\* (/careers, /jobs, /join-us)  
- [ ] \*\*Legal Pages\*\* (/privacy, /terms, /legal)

### \*\*Step 3: Content Discovery and Expansion (Priority Level 3)\*\*

Identify additional high-value content:

### Content Analysis Pages  
- [ ] \*\*Blog/News Archive\*\* (/blog, /news, /insights)  
- [ ] \*\*Resource Downloads\*\* (/downloads, /resources, /guides)  
- [ ] \*\*Event/Webinar Pages\*\* (/events, /webinars, /workshops)  
- [ ] \*\*Partner/Client Pages\*\* (/partners, /clients, /affiliates)  
- [ ] \*\*Industry-Specific Pages\*\* (specialized services, niche offerings)

### 4.2 Procedure: Systematic Page Discovery Protocol

Implement structured approach to finding all important pages.

### \*\*Phase 1: Initial Discovery (15 minutes)\*\*

# Implementation template for agents  
def discover\_priority\_pages(target\_url):  
 """Systematic page discovery for priority analysis"""  
   
 discovered\_pages = {  
 'core\_pages': [],  
 'service\_pages': [],  
 'content\_pages': [],  
 'conversion\_pages': []  
 }  
   
 # Step 1: Main navigation analysis  
 nav\_links = extract\_navigation\_links(target\_url)  
 discovered\_pages['core\_pages'].extend(nav\_links)  
   
 # Step 2: Sitemap parsing  
 sitemap\_pages = parse\_sitemap(f"{target\_url}/sitemap.xml")  
 categorize\_sitemap\_pages(sitemap\_pages, discovered\_pages)  
   
 # Step 3: Footer analysis  
 footer\_links = extract\_footer\_links(target\_url)  
 discovered\_pages['content\_pages'].extend(footer\_links)  
   
 return prioritize\_pages(discovered\_pages)

### \*\*Phase 2: Priority Ranking (10 minutes)\*\*

### Page Prioritization Scoring  
| Factor | Weight | High Score (3) | Medium Score (2) | Low Score (1) |  
|--------|--------|---------------|------------------|---------------|  
| \*\*Navigation Prominence\*\* | 30% | Main nav menu | Footer/secondary | Buried link |  
| \*\*Business Relevance\*\* | 25% | Core service | Supporting info | General content |  
| \*\*Conversion Impact\*\* | 25% | Direct conversion | Influence conversion | Informational |  
| \*\*SEO Value\*\* | 20% | High traffic potential | Medium traffic | Low traffic |  
  
### Final Priority Categories  
- \*\*Priority 1 (9-12 points)\*\*: Analyze immediately, full detail  
- \*\*Priority 2 (6-8 points)\*\*: Analyze after core pages, standard detail   
- \*\*Priority 3 (3-5 points)\*\*: Analyze if time permits, basic overview

### 4.3 Procedure: Crawling Execution Protocol

Define systematic approach to analyzing priority pages.

### \*\*Crawling Strategy Implementation\*\*

### Phase 1: Core Page Analysis (Must Complete)  
\*\*Time Allocation\*\*: 60-90 minutes  
\*\*Pages\*\*: Homepage, About, Services, Contact, Features  
\*\*Analysis Depth\*\*: Comprehensive  
  
\*\*Required Analysis per Page\*\*:  
- [ ] Page title extraction and optimization assessment  
- [ ] Meta description extraction and length validation  
- [ ] H1-H6 heading structure analysis  
- [ ] Internal linking assessment  
- [ ] Performance testing (Core Web Vitals)  
- [ ] Accessibility compliance check  
- [ ] Mobile responsiveness validation  
- [ ] Screenshot documentation  
  
### Phase 2: Service/Product Page Analysis (High Priority)  
\*\*Time Allocation\*\*: 30-60 minutes   
\*\*Pages\*\*: All service/product category pages  
\*\*Analysis Depth\*\*: Standard  
  
\*\*Required Analysis per Page\*\*:  
- [ ] SEO metadata optimization  
- [ ] Content quality assessment  
- [ ] Call-to-action effectiveness  
- [ ] Internal linking structure  
- [ ] Performance impact assessment  
- [ ] Screenshot of key elements  
  
### Phase 3: Supporting Content Analysis (If Time Permits)  
\*\*Time Allocation\*\*: 15-30 minutes  
\*\*Pages\*\*: Blog, FAQ, testimonials, resources  
\*\*Analysis Depth\*\*: Overview  
  
\*\*Required Analysis per Page\*\*:  
- [ ] Basic SEO metadata check  
- [ ] Content relevance assessment  
- [ ] Internal linking contribution  
- [ ] Overall site value contribution

### \*\*Smart Crawling Limits\*\*

def implement\_smart\_crawling(discovered\_pages, time\_limit\_minutes=120):  
 """Implement intelligent crawling based on time constraints"""  
   
 crawling\_plan = {  
 'priority\_1\_pages': discovered\_pages['priority\_1'][:10], # Max 10 core pages  
 'priority\_2\_pages': discovered\_pages['priority\_2'][:20], # Max 20 service pages   
 'priority\_3\_pages': discovered\_pages['priority\_3'][:30] # Max 30 content pages  
 }  
   
 # Adjust based on available time  
 if time\_limit\_minutes < 60:  
 # Limited time - focus on core pages only  
 crawling\_plan['priority\_2\_pages'] = crawling\_plan['priority\_2\_pages'][:5]  
 crawling\_plan['priority\_3\_pages'] = []  
 elif time\_limit\_minutes < 90:  
 # Medium time - include key service pages  
 crawling\_plan['priority\_2\_pages'] = crawling\_plan['priority\_2\_pages'][:10]  
 crawling\_plan['priority\_3\_pages'] = []  
   
 return crawling\_plan

### 4.4 Procedure: Analysis Depth Standards

Define different levels of analysis based on page priority.

### \*\*Comprehensive Analysis (Priority 1 Pages)\*\*

### Full Analysis Requirements  
\*\*Technical SEO Analysis\*\*:  
- [ ] Page title: Extract actual title, measure length, assess optimization  
- [ ] Meta description: Extract actual description, validate length and relevance  
- [ ] Meta keywords: Check for presence (flag if present - outdated practice)  
- [ ] Canonical URL: Verify proper canonical implementation  
- [ ] Open Graph tags: Assess social media optimization  
- [ ] Schema markup: Identify and validate structured data  
- [ ] Robots meta: Check for crawling/indexing directives  
  
\*\*Content Analysis\*\*:  
- [ ] H1 tag: Single H1 validation and optimization assessment  
- [ ] Heading hierarchy: H1-H6 structure and SEO effectiveness  
- [ ] Content length: Word count and depth assessment  
- [ ] Keyword optimization: Primary keyword integration analysis  
- [ ] Internal linking: Link structure and anchor text optimization  
- [ ] External linking: Outbound link quality and relevance  
  
\*\*Technical Performance\*\*:  
- [ ] Page load speed: GTMetrix integration for actual measurements  
- [ ] Core Web Vitals: LCP, FID, CLS measurement and optimization  
- [ ] Mobile responsiveness: Cross-device compatibility testing  
- [ ] Image optimization: Alt tags, file sizes, format optimization  
- [ ] JavaScript/CSS: Resource optimization opportunities  
  
\*\*User Experience\*\*:  
- [ ] Navigation clarity: Menu structure and user flow assessment  
- [ ] Call-to-action effectiveness: CTA placement and clarity  
- [ ] Form functionality: Contact forms and conversion elements  
- [ ] Accessibility compliance: WCAG guideline adherence  
- [ ] Visual design: Professional appearance and brand consistency

### \*\*Standard Analysis (Priority 2 Pages)\*\*

### Standard Analysis Requirements  
\*\*Essential SEO Elements\*\*:  
- [ ] Page title and meta description extraction and optimization  
- [ ] H1 tag validation and heading structure overview  
- [ ] Basic content quality assessment  
- [ ] Internal linking evaluation  
- [ ] Core performance metrics (basic GTMetrix check)  
  
\*\*Key Performance Indicators\*\*:  
- [ ] Page load speed (basic measurement)  
- [ ] Mobile compatibility check  
- [ ] Basic accessibility scan  
- [ ] Content relevance to services/products  
- [ ] Call-to-action presence and clarity

### \*\*Overview Analysis (Priority 3 Pages)\*\*

### Basic Analysis Requirements  
\*\*Quick Assessment\*\*:  
- [ ] Page title and meta description presence  
- [ ] H1 tag identification  
- [ ] Content type and relevance  
- [ ] Basic functionality check  
- [ ] Integration with overall site structure

### 4.5 Procedure: Quality Assurance and Validation

Ensure comprehensive coverage of priority pages.

### \*\*Completion Validation Checklist\*\*

### Pre-Report Quality Check  
- [ ] All Priority 1 pages analyzed with full comprehensive analysis  
- [ ] Minimum 80% of Priority 2 pages analyzed with standard analysis  
- [ ] Page discovery methodology documented in report  
- [ ] Screenshots captured for all Priority 1 pages  
- [ ] Any skipped pages documented with reasoning  
- [ ] Total pages analyzed clearly stated in report summary  
  
### Coverage Documentation  
\*\*Required in All Reports\*\*:

## Page Analysis Coverage Summary

: [NUMBER]

: [NUMBER] / [TOTAL] (100% target)

: [NUMBER] / [TOTAL] (80% target)

: [NUMBER] / [TOTAL] (time permitting)

### Pages Not Analyzed

## 5.0 Agent Implementation Requirements  
  
### 5.1 Technical SEO Analyst Integration

|  |  |  |
| --- | --- | --- |
| **Page URL** | **Reason** | **Impact** |
| /old-blog-post | Low priority | Minimal SEO impact |
| /archive/2019 | Outdated content | No current relevance |

# Implementation for priority-based crawling

class PriorityCrawler:

def \_\_init\_\_(self, target\_url, max\_pages=50):

self.target\_url = target\_url

self.max\_pages = max\_pages

self.discovered\_pages = {}

def execute\_priority\_crawl(self):

# Phase 1: Page discovery

self.discover\_priority\_pages()

# Phase 2: Priority ranking

self.rank\_pages\_by\_priority()

# Phase 3: Systematic analysis

return self.analyze\_by\_priority()

def discover\_priority\_pages(self):

"""Implement systematic page discovery"""

# Implementation details...

pass

### 5.2 Reporting Integration  
  
#### \*\*Priority Page Analysis Report Template\*\*

# Priority Page Analysis Report

## Executive Summary

This analysis focused on the most critical pages for SEO and user experience optimization, ensuring comprehensive coverage of core business pages while optimizing analysis time and resource allocation.

## Page Discovery Methodology

:

* [✓] Main navigation menu analysis
* [✓] XML sitemap parsing (/sitemap.xml)
* [✓] Footer link extraction
* [✓] Internal link following (2-level depth)
* [✓] Manual business-specific page identification

## Priority Page Analysis Results

### Priority 1 Pages (Comprehensive Analysis)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Page** | **Title Length** | **Meta Desc** | **H1 Issues** | **Performance** | **Priority Actions** |
| Homepage | 45 chars ✓ | Missing ❌ | Single H1 ✓ | 82/100 | Add meta description |
| About | 73 chars ❌ | 165 chars ❌ | Multiple H1 ❌ | 78/100 | Shorten title, fix H1 |
| Services | 52 chars ✓ | 145 chars ✓ | Single H1 ✓ | 85/100 | Minor optimizations |

### Priority 2 Pages (Standard Analysis)

[Similar table format with key findings]

### Analysis Coverage Summary

: 23 of 45 discovered pages

:

* Priority 1: 8/8 pages (100% - Complete)
* Priority 2: 12/15 pages (80% - Target met)
* Priority 3: 3/22 pages (14% - Time limited)