

Redirect Hygiene Audit SOP v1.0 (Agent-Ready)

Purpose: Encode all business logic so an automated agent can evaluate a site's redirects, push structured results into a database, and produce prioritized fixes that another agent (or VA/dev) can execute.

Outcomes:

- Zero internal 3xx in the link graph
- Zero internal 4xx
- No loops; no chains (>1 hop)
- External link equity preserved (final 200 in ≤ 1 hop)
- Canonicals align with final 200

1) Inputs & Data Contracts

1.1 Required inputs

- **Crawl exports** (from Screaming Frog CLI, Sitebulb, JetOctopus, or in-house crawler):
 - Pages with status codes and final URLs
 - Internal links (source $\rightarrow$ href)
 - Redirect chains (A $\rightarrow$... $\rightarrow$ Z with codes)
 - Canonical targets
- **Config** (JSON):

```
{
  "host_preferred": "www.example.com",
  "http_to_https": true,
  "enforce_trailing_slash": true,
  "enforce_lowercase": true,
  "max_redirect_hops": 5,
  "allowed_internal_redirects": 0,
  "max_chain_hops": 1,
  "backlink_data_available": false
}
```

- **Optional feeds:**
 - Backlinks (domain $\rightarrow$ target URL, RD metrics)
 - GA4/Search Console (page clicks/impressions)
 - Revenue per lead or per session (for business impact scoring)

1.2 Database schema (normalized)

pages

```

url TEXT PRIMARY KEY,
status INT,
final_url TEXT,
hop_count INT DEFAULT 0,
content_type TEXT,
canonical TEXT,
indexable BOOLEAN,
last_crawled TIMESTAMP

```

links_internal

```

source_url TEXT,
href TEXT,
href_final_url TEXT,
href_status INT,
anchor_text TEXT,
PRIMARY KEY (source_url, href)

```

redirects

```

chain_id TEXT,
step INT,
from_url TEXT,
status_code INT,
to_url TEXT,
PRIMARY KEY (chain_id, step)

```

policy_violations

```

violation_id TEXT PRIMARY KEY,
rule_code TEXT,
url TEXT,
context JSONB,
severity INT,
impact_score NUMERIC,
created_at TIMESTAMP DEFAULT now()

```

tasks

```

task_id TEXT PRIMARY KEY,
owner_role TEXT,          -- "VA", "DEV", "SEO"
action TEXT,              -- see Task Types below

```

```

payload JSONB,          -- action-specific fields
priority INT,
status TEXT,            -- "open","in_progress","done","blocked"
related_violation TEXT  -- FK to policy_violations.violation_id

```

backlinks_checked (*optional*)

```

source_domain TEXT,
target_url TEXT,
final_url TEXT,
hop_count INT,
result_status INT,
rd_score NUMERIC

```

runs

```

run_id TEXT PRIMARY KEY,
started_at TIMESTAMP,
finished_at TIMESTAMP,
config JSONB,
summary JSONB

```

run_artifacts

```

run_id TEXT,
artifact_type TEXT, --
"summary_pdf","fix_sheet","server_rules_apache","server_rules_nginx"
uri TEXT,
PRIMARY KEY (run_id, artifact_type)

```

2) Business Rules (authoritative)

Prefix each rule with an ID so the agent can log precise violations.

R01 – No internal 4xx. Any internal URL returning 4xx is a violation.

R02 – No internal 3xx in link graph. Internal links must resolve directly to a 200 (no hops). 0 allowed.

R03 – Max chain hops = 1 for external requests. Legacy/external hits may go through **one** 301 to final 200.

R04 – No redirect loops. Any cycle detected in redirect graph is critical.

R05 – Canonical must target a 200 final URL. Canonical cannot point to 3xx/4xx or a different final URL than rendered.

R06 – Preferred host enforced. Only `host_preferred` allowed; others 301→preferred.

R07 – HTTP→HTTPS enforced when `http_to_https = true`.

R08 – Trailing slash policy enforced. All internal links must match `enforce_trailing_slash`.

R09 – Lowercase policy enforced. All internal links lowercase when `enforce_lowercase = true`.

R10 – Backlink preservation. Any known backlink target must return 200 in ≤ 1 hop (or mapped to best match).

R11 – Internal link update required. When a target moved, update all internal sources to the new destination (don't rely on 301 internally).

R12 – Homepage is not a catch-all. Only map to homepage when no materially relevant equivalent exists.

R13 – Orphan pages flagged. Indexable pages with no inlinks from crawl or sitemaps are violations.

R14 – Redirect to self disallowed. Same URL redirecting to itself is a loop variant.

R15 – Canonical consistency across variants. `http/www/trailing-slash/case` variants must consolidate to one canonical.

R16 – Max redirect hops (global). Any hop chain $> \text{max_redirect_hops}$ is a hard fail.

R17 – Mixed canonicalization signals. Conflicts between canonical, hreflang, and redirects are violations.

R18 – 302 use is temporary only. 302s older than 14 days are violations unless exempted.

R19 – 301 longevity. Keep 301s for legacy URLs with backlinks indefinitely.

R20 – Sitemaps must list canonical 200s only. Non-200 or redirected URLs in sitemaps are violations.

3) Decision Logic (agent algorithm)

3.1 Normalize

- Normalize URLs to preferred host/lowercase/trailing-slash for comparison keys.
- Build maps: `final(url)`, `status(url)`, `inlinks(url)`, `canonical(url)`.

3.2 Build redirect graph

- For each URL, follow redirects (cap `max_redirect_hops`).
- Record `chain_id`, steps, and detect cycles using `visited` set.

3.3 Detect & classify

- **4xx internal** ⇒ R01
- **Internal link to 3xx** ⇒ R02 + create **T02: UpdateInternalLink**
- **Chain length >1 (external context)** ⇒ R03
- **Loop/cycle** ⇒ R04 (critical)
- **Canonical → non-200 or to redirect** ⇒ R05
- **Variant violations** (host/http/slash/case) ⇒ R06–R09
- **Backlink target 4xx or >1 hop** ⇒ R10
- **Orphan** (indexable=true & inlinks=0 & not in sitemaps) ⇒ R13
- **Redirect to self** ⇒ R14
- **Hop chain >max** ⇒ R16
- **Mixed signals** (canonical ≠ final) ⇒ R17
- **302 older than 14d** ⇒ R18
- **Legacy 301 removed** where backlink exists ⇒ R19
- **Sitemaps non-200/3xx** ⇒ R20

3.4 Best-match mapping (for retired URLs)

Order of heuristics (stop at first hit):

1. **Exact slug rename** (path diff minimal; same parent folder)
2. **Template + taxonomy match** (same practice area/state/city)
3. **Anchor/context similarity** (from inlinks/backlinks anchor text)
4. **Title/content similarity** (if HTML/extracts provided; cosine ≥ 0.6)
5. **Nearest parent category** page
6. **Homepage** (*last resort only; R12 warns*)

Emit decision rationale into `context` JSON (e.g., `{"heuristic":"taxonomy_match","confidence":0.82}`).

4) Task Types & Payloads

T01: AddRedirect (owner: DEV)

```
{
  "from_url": "https://example.com/old/",
  "to_url": "https://www.example.com/new/",
  "status_code": 301,
  "rationale": "best_match: taxonomy_match",
}
```

```
"permanent": true
}
```

T02: UpdateInternalLink (owner: VA/DEV)

```
{
  "source_url": "https://www.example.com/page/",
  "old_href": "/old/",
  "new_href": "/new/"
}
```

T03: FixCanonical (owner: DEV)

```
{
  "url": "https://www.example.com/page/",
  "current_canonical": "https://www.example.com/page-old/",
  "new_canonical": "https://www.example.com/page/"
}
```

T04: RemoveLoop/CollapseChain (owner: DEV)

```
{
  "chain_id": "abc123",
  "chain": [
    {"from": "/a", "code": 301, "to": "/b"},
    {"from": "/b", "code": 301, "to": "/c"}
  ],
  "fix": {"a": "/c"}
}
```

T05: PublishOr410 (owner: SEO/DEV)

```
{
  "url": "/retired-page/",
  "decision": "301_to_best_match | 410_gone",
  "best_match": "/category/",
  "rationale": "no equivalent content; parent category"
}
```

T06: SitemapFix (owner: VA/DEV)

```
{
  "sitemap": "https://www.example.com/sitemap.xml",
  "bad_entries": ["/redirected/", "/404/"]
}
```

5) Severity, Priority & Impact Scoring

Severity (S):

- S1 Critical: Loops (R04), massive 4xx on key templates, hop>max (R16)
- S2 High: Chains>1 (external), internal 3xx (R02), backlink 4xx (R10)
- S3 Medium: Canonical mismatches (R05/R17), policy variants (R06–R09)
- S4 Low: Orphans (R13) on low-value pages, sitemap nits (R20)

ImpactScore (0–1000)

```
= 40*(status_is_4xx)
+ 50*(is_loop)
+ 30*max(chain_hops-1,0)
+ 10*internal_inlinks
+ 5*organic_clicks
+ 100*has_backlinks
+ 0.5*backlink_count
+ 5*rd_score
+ 15*canonical_mismatch
+ 5*template_priority
```

Missing metrics default to 0. Agent sets **priority** by (Severity desc, ImpactScore desc).

6) Acceptance Criteria (run-level)

- A1: `internal_4xx_count == 0`
 - A2: `internal_3xx_in_graph == 0`
 - A3: `loops == 0` and `chains_collapsed_to<=1`
 - A4: `backlink_targets_ok == 100%` (or explicit exceptions)
 - A5: `canonicals_point_to_200 == 100%`
-

7) Pipeline & Agent Roles

1. **Crawler (Producer)** → writes raw CSV/JSON to storage.

2. **Ingestor** → loads into `pages`, `links_internal`, `redirects`.
3. **Policy Evaluator** → applies R01-R20; writes `policy_violations`.
4. **Mapper** → computes best matches; emits **T01/T05**.
5. **Task Generator** → emits **T02/T03/T04/T06** and sets priority.
6. **Reporter** → produces PDF summary, Fix Sheet, server rule bundles; writes to `run_artifacts`.
7. **Reviewer Agent** → validates deltas; toggles tasks → `done` when verified by re-crawl.

Event contracts (JSON)

CrawlCompleted

```
{"event": "CrawlCompleted", "run_id": "2025-08-14-001", "artifacts": ["s3://.../pages.csv", ".../redirect_chains.csv"]}
```

TasksReady

```
{"event": "TasksReady", "run_id": "2025-08-14-001", "open_tasks": 127}
```

Idempotency: use `(run_id, url, rule_code)` as composite key to avoid duplicate violations.

8) Reference Logic (pseudocode)

Redirect graph & loop detection

```
for each url in pages:
    visited = {}
    hops = []
    cur = url
    for i in 1..config.max_redirect_hops:
        if cur in visited: emit R04(loop); break
        visited.add(cur)
        nxt = redirect_to(cur)
        if not nxt: break
        hops.append({from:cur,to:nxt,code:status(cur)})
        cur = nxt
    record final_url=cur, hop_count=len(hops)
```

Internal 3xx hygiene


```

for each (source, href) in links_internal:
  if is_internal(href) and status(href) in [301,302]:
    emit R02; create T02 UpdateInternalLink(source, href, final(href))

```

Chain collapse suggestion

```

for each chain where hops > 1:
  suggest 301 from first.from_url to last.to_url

```

Best-match mapping (heuristic order)

```

candidates = find_same_template_and_taxonomy(old_url)
if empty: candidates = find_anchor_context_matches(old_url)
if empty: candidates = find_title_content_similarity(old_url)
if empty: candidates = [parent_category(old_url)]
if empty: candidates = [homepage]
choose best with score and emit T01

```

9) Server Rule Generation (templates)

Nginx

```

# Force HTTPS
if ($scheme = http) { return 301 https://$host$request_uri; }

# Force preferred host
if ($host != www.example.com) { return 301 https://
www.example.com$request_uri; }

# Lowercase (example map)
map $request_uri $lc_uri { default ""; ~[A-Z] tolower($request_uri); }
if ($lc_uri) { return 301 https://$host$lc_uri; }

# Trailing slash
location ~* ^/([^.?]+[^\/]$) { return 301 https://$host/$1/; }

# Custom redirects (generated)
rewrite ^/old-path/$ https://www.example.com/new-path/ permanent;

```

Apache (.htaccess)

```
RewriteEngine On
RewriteCond %{HTTPS} !=on
RewriteRule ^ https://%{HTTP_HOST}%{REQUEST_URI} [R=301,L]

RewriteCond %{HTTP_HOST} !^www\.example\.com$ [NC]
RewriteRule ^ https://www.example.com%{REQUEST_URI} [R=301,L]

RewriteCond %{REQUEST_URI} [A-Z]
RewriteRule (.*?) ${lc:$1} [R=301,L]

RewriteCond %{REQUEST_URI} !\.$
RewriteCond %{REQUEST_URI} !/$
RewriteRule ^(.+)$ https://www.example.com/$1/ [R=301,L]

Redirect 301 /old-path/ https://www.example.com/new-path/
```

The agent fills custom redirects from **T01** tasks.

10) Reporting Artifacts

- **Executive Summary (HTML/PDF):** counts by rule (R01–R20), top 10 impacts.
- **Fix Sheet (CSV/Sheet):** concatenated tasks (T01–T06) with owners and priority.
- **Server Rules:** Nginx/Apache files built from T01 + global policy.
- **Delta Report:** Before vs After metrics.

11) SLAs & SLOs

- **SLA:** High-severity (S1/S2) tasks actioned within 3 business days.
- **Re-crawl cadence:** Weekly for active migrations; monthly otherwise.
- **SLO:** Maintain A1–A5 $\geq 99\%$ across runs.

12) Edge Cases

- JS-only routes: sample render; don't blow crawl budget.
 - Locale/hreflang: ensure canonicals and redirects align per locale.
 - Query parameters: whitelist meaningful params; strip others.
 - Pagination: rel="next/prev" (if used) must not redirect to non-self.
 - Files (PDF, images): validate 200 and correct Content-Type.
-

13) Rollback Plan

- All server rule changes shipped behind a flag; keep previous rules for 7 days.
 - If 5xx spikes after deploy, auto-rollback to last stable artifact.
-

14) Example End-to-End (fictional)

- `CrawlCompleted(run_id=2025-08-14-001)` ingested.
 - Evaluator flags:
 - R02: 178 internal 301s → T02 updates
 - R03: 12 chains (>1 hop) → T04 collapse
 - R04: 1 loop → T04 remove
 - R05: 9 canonicals → T03 fixes
 - R10: 6 backlink 404s → T01 mappings
 - Total tasks: 206; Priorities assigned by ImpactScore.
 - Reporter emits artifacts; Reviewer schedules re-crawl T+72h.
-

This SOP is the single source of truth for the redirect audit agent. Update version and changelog on material policy changes.