MRD – Self-Hosted Website Ingestion & Citation Service

Version	Date	Author
1.0 (Hardened Spec)	2025-07-20	Warlack

1. Purpose

To provide a self-hosted tool that crawls a website, indexes its content, and exposes a retrieval API that returns answers with precise, verifiable in-text citations.

2. Problem / Opportunity

Large Language Models (LLMs) lack authoritative context for private or niche websites and often hallucinate or provide broken links, eroding user trust. Existing SaaS solutions can compromise data privacy and lack the deep customisation required by technical users. Enthusiasts and small teams running home servers need a local, open-source, and trustworthy alternative.

3. Target Users

- **Technical Hobbyists:** Users with home labs (running Docker, Kubernetes, Proxmox).
- **Researchers:** Academics and professionals requiring offline citation accuracy.
- **Small Technical Teams:** Groups of up to 10 engineers with sensitive documentation.

4. Key Differentiators

- **100% Self-Hosted:** No external API calls after initial setup.
- **Reduced Detection Footprint:** Crawls from a residential/business IP with configurable rates and proxy support, making it less visible than cloud scrapers.
- **Trustworthy Citations:** Core architecture is built around preventing stale links and providing verifiable sources.
- **Plug-and-Play:** Distributed as a secure, containerized stack.

5. Functional Requirements (v1.0 Scope)

ID	Requirement	Tier
F1	Crawl entire domain, respect robots.txt & sitemaps, with a	Community

	basic scheduler.	
F2	**Manual Deletion Detection:** CLI command to find and prune stale links (404 checks).	Community
F3	**Pro: Automated & Scheduled** deletion detection.	Pro
F4	Clean HTML → Markdown, recursive splitter ≤ 512 tokens, handle code & tables.	Community
F5	Selectable embedding model (OpenAI, local GGUF, BGE).	Community
F6	Store vectors + source URL in a pluggable DB (ChromaDB officially supported; others are community-supported).	Community
F7	/query endpoint for retrieval.	Community
F8	CLI + simple React dashboard for status monitoring.	Community
F9	**Pro:** Access control (API key) and a Data Purge API to remove a URL's data.	Pro

6. Non-Functional Requirements

- **Performance:**
  - **Static Content:** Query latency ≤ 700 ms P95.
  - **JS-Rendered Content:** Query latency ≤ 2 s P95 (excluding initial render).
- **Security:** Non-root containers. Default deployment includes a Caddy/Traefik container for automatic HTTPS and rate limiting.
- **Reliability:** 95% queries within SLA; < 0.1% unnoticed crawl failures (alerting).
- **Backup & Disaster Recovery:** A daily, automated snapshot of the vector index and metadata. A documented make restore script for recovery.

7. Success Metrics

Metric	Target
Citation Accuracy*	≥ 95%
Setup Time (with HTTPS)	≤ 30 min
Crawl Coverage	≥ 99%

**Golden set of 50–100 Q/A/source triplets. CI fails if accuracy < 95%.*

8. Out of Scope (v1.0)

- Advanced bot evasion (e.g., JS fingerprinting, CAPTCHA solving).
- Mobile UI, PDF OCR, on-the-fly translation.
- Enterprise features (SSO, RBAC, advanced compliance).

9. Milestones

Date	Milestone
2025-08-30	Prototype CLI crawler + Chroma + Flask API
2025-10-31	Beta Docker stack, MCP support, basic UI
2026-01-31	**v1.0 Release:** Stable, documented release with Helm chart.

10. Risks & Mitigations

- **JS-Heavy Sites / Advanced Bot Blocks:** Use Playwright fallback. Acknowledge that complex client-side challenges (e.g., Turnstile) are out of scope for v1.0. Provide clear documentation on using external proxy services as the primary mitigation.
- **Upstream Dependencies:** Contribute features back to upstream projects. Avoid private forks. Pin tested versions and use integration tests to catch breaking changes.
- **Component Licensing (Playwright):** Document the licensing of bundled browser binaries and provide instructions for users to leverage system-installed browsers to avoid redistribution concerns.
- **Timeline:** The v1.0 release date has been extended to Jan 2026 to realistically account for the complexity of a polished and reliable release.

11. Market & Pricing Strategy (v1.0)

Tier	Price	Key Features
Community	Free	2 websites, 50k tokens, manual deletion detection, basic scheduler.
Pro	US$249 one-time	Unlimited websites/pages, **automated deletion detection**, API key access, Data Purge API, priority email support.
Updates (Optional)	US$59/yr	Provides ongoing patches and model updates for the Pro tier.

12. Future Scope (Post-v1.0)

- **Enterprise Edition:** A separate, subscription-based offering for larger teams with features like SSO, RBAC, and advanced compliance tools.
- **Advanced Features:** Integrated chat UI, multilingual support, advanced analytics.

13. Change Log

Version	Date	Change
1.0	2025-07-20	**Hardened Spec.** Addressed fault lines: Moved deletion detection to Community, refined performance SLAs, added DR/backup spec, extended timeline, clarified security/licensing, and simplified the GTM plan to focus on a robust core product.