**Business Requirements Specification: Nexus Scraper**

Version: 1.3

Date: April 5, 2025

**1. Introduction**

* **1.1 Purpose:** This document defines the requirements for Nexus Scraper, an advanced, scalable, and robust web scraping platform designed to extract data from complex, dynamic websites efficiently, reliably, and ethically. Its primary value lies in reliably handling dynamic content and bypassing sophisticated anti-scraping measures while offering seamless integration for AI agents. It aims to serve data scientists, market researchers, developers, and business analysts who require high-quality data extraction capabilities beyond standard scraping tools. It also includes requirements for integration with external AI tools and agents, emphasizing AI-interpretable inputs and outputs.
* **1.2 Scope:** ... [unchanged from v1.2] ...
* **1.3 Definitions:** ... [unchanged from v1.2] ...

**2. Overall Description**

* **2.1 Product Perspective:** Nexus Scraper will be a standalone platform... [rest of section mostly unchanged] ... It will expose functionality through a GUI, CLI, a comprehensive REST API for general use, and specific interfaces designed for programmatic consumption by AI agents, featuring optimized data structures for AI interaction.
* **2.2 User Characteristics:** ... [unchanged] ...
* **2.3 General Constraints:**
  + Must be compatible with major operating systems (Linux, macOS, Windows).
  + Leverage modern, maintainable technology stacks (e.g., Python/Node.js backend, React/Vue frontend, Chromium-based headless browser).
  + Must adhere to best practices for security and data privacy.
  + The system must have a clear strategy for managing external dependencies, including the headless browser engine and security patching. *(Added)*
* **2.4 Assumptions and Dependencies:** ... [unchanged] ...

**3. Functional Requirements**

* **FR-CORE-001 to FR-CORE-005:** ... [unchanged from v1.2] ...
* ... [FR-DYN unchanged] ...
* ... [FR-ANTIBLOCK unchanged] ...
* ... [FR-DATA unchanged, including FR-DATA-AI-001] ...
* **FR-MGMT-001 to FR-MGMT-004:** ... [unchanged] ...
* **FR-MGMT-005: Distributed Scraping:** The system shall be designed to run in a distributed mode across multiple nodes/workers for high-throughput scraping, coordinating task distribution, *maintaining distributed state*, and result aggregation. *(Enhanced description)*
* **FR-MGMT-ERR-001: Granular Error Reporting:** The system *shall* provide detailed and categorized error reporting, distinguishing between network issues, HTTP status code errors (e.g., 4xx, 5xx), content parsing errors, anti-blocking events, and internal system errors. *(New requirement)*
* ... [FR-AI unchanged] ...
* ... [FR-ETHICS unchanged] ...
* **FR-IFACE-001 to FR-IFACE-003:** ... [unchanged] ...
* **FR-IFACE-AI-001 to FR-IFACE-AI-006:** ... [unchanged from v1.2] ...
* **FR-IFACE-AI-007: API Versioning:** The AI Agent API *shall* implement a versioning strategy (e.g., URL path versioning) to allow for future enhancements without breaking compatibility for existing AI agent integrations. *(New requirement)*

**4. Non-Functional Requirements**

* **NFR-PERF-001: Throughput:** The system must be capable of achieving high scraping throughput, scalable with the number of distributed workers. The system *shall* be benchmarked to demonstrate scalable throughput (e.g., target X pages/minute/worker under specified conditions TBD). *(Made more concrete)*
* **NFR-PERF-002: Latency:** Headless browser rendering time should be optimized. Direct HTML fetch latency should be minimal, primarily dependent on network conditions. The system *shall* provide metrics on different stages of request latency. *(Made more concrete)*
* **NFR-PERF-003: Resource Utilization:** The system should be efficient in CPU and RAM usage... [rest unchanged] ...
* **NFR-SCAL-001: Horizontal Scalability:** ... [unchanged] ...
* **NFR-SCAL-002: Vertical Scalability:** ... [unchanged] ...
* **NFR-REL-001: Reliability:** ... [unchanged] ...
* **NFR-REL-002: Fault Tolerance:** ... [unchanged] ...
* **NFR-REL-003: Data Accuracy:** ... [unchanged] ...
* **NFR-SEC-001: Credential Security:** ... [unchanged] ...
* **NFR-SEC-002: API Security (General):** ... [unchanged] ...
* **NFR-SEC-003: Input Validation:** ... [unchanged] ...
* **NFR-SEC-004: Dependency Security:** The system *shall* incorporate mechanisms or support processes to scan and manage security vulnerabilities in its third-party dependencies regularly. *(New requirement)*
* **NFR-SEC-AI-001: Secure AI Agent Authentication:** ... [unchanged] ...
* **NFR-USAB-001: GUI Ease of Use:** ... [unchanged] ...
* **NFR-USAB-002: Configuration Clarity:** ... [unchanged] ...
* **NFR-USAB-003: Documentation:** ... [unchanged from v1.2] ...
* **NFR-USAB-AI-001: AI Discoverability:** ... [unchanged] ...
* **NFR-MAINT-001: Modularity:** ... [unchanged] ...
* **NFR-MAINT-002: Code Quality:** ... [unchanged] ...
* **NFR-MAINT-003: Logging:** ... [unchanged] ...
* **NFR-MAINT-004: Testability:** The system architecture *shall* facilitate unit, integration, and end-to-end testing. *(New requirement)*
* **NFR-PORT-001: OS Compatibility:** ... [unchanged] ...
* **NFR-PORT-002: Containerization:** ... [unchanged] ...

**5. Interface Requirements**

* **5.1 User Interfaces:** ... [unchanged] ...
* **5.2 API Interfaces:**
  + General RESTful API (Details in FR-IFACE-003, NFR-SEC-002).
  + AI Agent RESTful API (Details in FR-IFACE-AI-001 to FR-IFACE-AI-007, FR-DATA-AI-001, NFR-SEC-AI-001, NFR-USAB-AI-001). Must support JSON request/response bodies optimized for AI interpretation, provide machine-readable capability/schema descriptions, and implement API versioning. *(Revised description to include versioning)*
* **5.3 Data Interfaces:** ... [unchanged] ...
* **5.4 External Service Interfaces:** ... [unchanged] ...

**6. Data Requirements**

* ... [unchanged] ...

**7. Deployment and Operations**

* **7.1 Installation:** ... [unchanged] ...
* **7.2 Configuration:** ... [unchanged] ...
* **7.3 Monitoring:** ... [unchanged] ...
* **7.4 Updates:** ... [unchanged] ...
* **7.5 AI Policy Management:** ... [unchanged] ...
* **7.6 Backup and Recovery:** Procedures or capabilities for backing up critical system configuration and job state *shall* be defined and documented. *(New requirement)*