

Web Data Product Specification Document

Project name: e.g. “competitive pricing monitor for Acme product line”

1. Problem statement

- **Business use case:** *What decision or operational process will this dataset support?*
- **Goal state:** *The workflow or system outcome the data should enable (e.g. “daily alerts when SKU landed price deviates by $\geq 3\%$ ”).*
- **Success criteria:** *Measurable before-and-after impacts or ROI indicators (e.g. “reduce manual price checks by 80%”).*

2. Stakeholders

Role	Name	Responsibilities
Project owner	<i>Name</i>	Defines business need, goal state, and success criteria
Data user	<i>Name</i>	Consumes or acts on the delivered data
Legal / compliance	<i>Name</i>	Ensures alignment with policy and regulation
Technical lead	<i>Name</i>	Oversees crawler build, maintenance, and delivery

Note: One person may fill multiple roles, but each functional responsibility should be explicitly listed.

3. Technical requirements

3.1 Risk and compliance notes

Document the operational guardrails and compliance requirements that shape how the project is built and run.

Include as applicable:

- **Robots.txt policy** — whether to follow site robots.txt directives.
- **Rate limiting** — max requests per host/IP, concurrency settings.
- **Data retention policy** — how long raw and processed data will be stored.

- **Privacy regulations** — GDPR, CCPA, or other jurisdictional requirements.
- **Content restrictions** — no personal-data capture, sensitive-category data handling.
- **Security measures** — encryption in transit/at rest, secure credential storage.

3.2 Dependencies

List all systems, services, and integrations the crawlers will rely on:

- Upstream sources (files, APIs)
- Storage solutions (databases, cloud buckets)
- Downstream systems (business intelligence (BI) tools, analytics pipelines)
- Authentication services or credentials

3.3 Timeline and milestones

Milestone	Target date	Owner	Notes
Discovery and sample crawl	2025-08-15	Dev lead	Sample data approved
MVP scraper for all sites	2025-09-01	Dev team	
Production roll-out	2025-10-01	PM	

4. Implementation details

4.1 Data sources

#	Source URL or name	Crawl type	Delivery frequency	Geographic considerations	Navigation actions needed	Notes
1	https://example.com	Full crawl	One-time / daily / weekly	None	Pagination, filters	Main product catalog
2	https://example2.com	Category URL(s)	Daily	UK	Pagination, filters	Main product catalog

4.2 Field schema

Field name	Description	Sample value	Required (Y / N)	Transformation / cleaning
product_name	Title text shown on listing	"Widget 3000"	Y	Strip HTML, trim whitespace
price	Numeric price	24.99	Y	Remove currency symbol, convert to number
image_url	Link to product image	https://...jpg	N	Download file

Create separate tables for each entity type as required (e.g. article, seller).

4.3 Scale and cadence

- **Number of sources / sites:** e.g. two sites (listed above)
- **Estimated volume per run:** Approx. 2,500 product records total
- **Run-time SLA:** e.g. complete crawl within two hours
- **Project duration:** e.g. full scrape in month one, incremental weekly thereafter

4.4 Delivery details

- **Output format:** CSV, JSON, Parquet, SQL table, etc.
- **Delivery mechanism:** S3 bucket upload, FTP, REST API, email
- **Delivery frequency:** After each crawl / real-time stream / weekly batch
- **Post-processing / transformation:** Currency normalisation, deduplication

4.5 Out of scope

- Scraping of mobile apps
- User reviews or comments
- Historical data prior to project start date

4.6. Acceptance criteria

- **Data quality target:** ≥ 95 % of expected records; ≤ 1 % field error rate
- **Delivery SLA:** Dataset delivered by 06:00 UTC daily
- **User acceptance:** Analytics team validates a random 5 % sample meets needs
- **Monitoring and alerts:** Slack integration; #scrape-alerts channel active. Trigger job failure on partial/incomplete records.
- **Pilot exit criteria (if applicable):** One-month pilot on Site A meets all above targets

Version: 0.1

Last updated: YYYY-MM-DD