

Web Scraping Services for Responsible, Structured Data Collection

Public web data such as prices, listings and market information, collected within site rules and rate limits and delivered clean, structured and ready to use.

OVERVIEW

Web Scraping & Data Collection

We build web scraping and data collection pipelines that turn public web pages into structured data you can use: product prices and availability, property and job listings, public directories, news, tenders and market information. Data is cleaned, de-duplicated and delivered as CSV, JSON or XML, or loaded straight into your database or CRM.

We scrape responsibly. Before building anything, we check each site's robots.txt and terms of use, whether an official API or data licence exists, and whether the data includes personal information. We collect publicly available data only, never get past logins, paywalls or CAPTCHAs, and keep request rates low so we do not burden the sites we visit.

Where pages are messy, we combine scrapers with AI and machine learning to extract fields from free text, and add human review where accuracy matters. Scrapers are monitored and maintained, because websites change and a silent failure is worse than no data.

BUILT FOR

- E-commerce and retail pricing teams
- Market research and strategy teams
- Property, travel and automotive marketplaces
- Data science teams building training data sets
- Analysts tracking tenders, news and public filings

FEATURES

What you get with Web Scraping & Data Collection

01

Source and compliance review

For each site we check robots.txt, the terms of use and whether an API or licensed feed exists, and flag any personal data before collection starts.

02

Structured data extraction

Prices, product details, listings, directories, reviews and news turned into consistent fields that match the schema you need.

03

JavaScript-heavy sites

Pages that load content with JavaScript are read with headless browsers, within the same rules on access, rate and public data.

04

Scheduled collection and change tracking

Daily, weekly or hourly runs that record what changed since the last visit, such as a new price, a sold listing or an updated tender.

05

Cleaning and normalisation

Duplicates removed, units, currencies and dates standardised, and records matched across sources, so the data is ready for analysis.

06

AI extraction from unstructured pages

Language models and machine learning pull fields out of free text and tag records into schema-compliant output, with checks and sampled human review.

- 07 Web research**
Targeted research on markets, competitors and customer segments from public online sources, compiled into a structured data set.
- 08 Data mining and analysis**
Clustering, classification and trend analysis on the collected data, so you receive patterns and forecasts as well as raw records.
- 09 Delivery and integration**
Files by SFTP or cloud storage, a simple API, or direct loads into your relational database, CRM or CMS on each run.

WHY IT MATTERS

The difference it makes to your business

- 01 Current data, not last quarter's**
Scheduled runs keep prices, listings and market data up to date without anyone copying pages by hand.
- 02 Data in the shape you need**
Clean, consistent fields arrive ready for your spreadsheets, dashboards or models, with no reformatting.
- 03 Scales from a few sites to many**
The same pipeline grows from a handful of sources to thousands, with monitoring that shows which source failed and why.
- 04 Lower legal and reputational risk**
Compliance checks, polite crawling and minimal personal data keep collection within the rules and away from complaints.

HOW WE DELIVER

How we deliver Web Scraping & Data Collection

- 01 Discovery**
We agree your business objectives, the exact fields you need, how often you need them and the list of candidate sources.
- 02 Compliance check**
Each source is reviewed for robots.txt rules, terms of use, API alternatives and personal data. Sources that do not pass are dropped or replaced.
- 03 Build and pilot**
Scrapers are built with rate limits, retries and validation rules, then run on a sample so you can check the output against the schema.
- 04 Collect and clean**
Runs happen on schedule, data is cleaned and checked, and alerts tell us when a site changes or a run returns unusual results.

Data is delivered in your format and destination, and scrapers are updated as sites change, for as long as you need the feed.

FAQ

Questions about Web Scraping & Data Collection

Is web scraping legal?

It depends on what is collected, how and where. Collecting public, non-personal data at a reasonable rate while respecting a site's terms is generally lower risk. Getting past logins, ignoring terms, copying copyrighted content or collecting personal data raises legal questions under laws such as the CFAA in the USA, the GDPR in the EU and India's DPDP Act. This is not legal advice, so for sensitive projects we recommend checking with your own lawyer.

Do you respect robots.txt and website terms?

Yes. We check robots.txt and the terms of use for every source before building a scraper. If a site disallows the pages you need or prohibits automated collection, we look for an official API, a data licence or another source instead.

Can you scrape sites behind a login or a CAPTCHA?

No. We do not bypass logins, paywalls, CAPTCHAs or other access controls. If the data sits behind an account, the right route is an API or a data agreement with the site owner, which we can help you integrate.

What about personal data?

We avoid it by default and design scrapers to skip fields such as names, emails and phone numbers. If personal data is genuinely needed, the project must have a lawful basis under the GDPR, India's DPDP Act or the laws that apply to you, and we minimise, pseudonymise and secure what is kept.

In what format will we receive the data?

CSV, Excel, JSON or tagged XML files, a simple API, or direct loads into your relational database, CRM or CMS. We agree the schema with you first, so every delivery has the same fields.

What happens when a website changes its layout?

Scrapers can break when a site changes. Every run is validated, and alerts tell us when a source returns missing or unusual data, so we can fix the scraper rather than deliver bad data quietly.

Should we use an API instead?

If the site offers an official API or data feed that covers what you need, usually yes. It is more stable and clearly permitted. We check for one during discovery and only scrape where no suitable alternative exists.

How quickly can we get data?

Discovery and the source check take one to two weeks, and a pilot sample usually follows within one to three weeks, depending on the number of sources and how complex the pages are. Scheduled deliveries start once you approve the sample.

How much does web scraping cost?

It depends on the number of sources, how complex the pages are, how often data is collected and how much cleaning it needs. Building the scrapers is a fixed price after the source check, and keeping them running is a monthly plan.

NEXT STEP

Tell us what data you need

Share the fields, sources and frequency you have in mind. We will check what can be collected responsibly and propose a pipeline with a fixed quote.

Book a free consultation: <https://appsin.ai/book-a-consultation/>

WhatsApp or call: [+91-9025555584](https://wa.me/919025555584)

About AppsInAi

AppsInAi Private Limited is an AI development company in Coimbatore, India. We build generative AI assistants, AI agents, chatbots, machine learning models, computer vision and data platforms on our clients' own data, from a first proof of concept to production, with monitoring and support after launch.

Contact

sales@appsin.ai

+91-9025555584

Housing Unit, Sf No. 393, CMS School Backside, 3rd Street, Nehru Nagar, Ganapathy, Coimbatore 641006, India

<https://appsin.ai>