

AN OVERVIEW OF THE 50 MOST COMMON WEB SCRAPING TOOLS

WEB SCRAPING IS THE PROCESS OF USING BOTS TO EXTRACT CONTENT AND DATA FROM A WEBSITE. UNLIKE SCREEN SCRAPING, WHICH ONLY COPIES PIXELS DISPLAYED ON SCREEN, WEB SCRAPING EXTRACTS UNDERLYING CODE – AND WITH IT, STORED DATA – AND OUTPUTS THAT INFORMATION INTO A DESIGNATED FILE FORMAT.

While legitimate uses cases exist for data harvesting, illegal purposes exist as well, including undercutting prices and theft of copyrighted content.

Understanding web scraping bots starts with understanding the diverse and assorted array of web scraping tools and existing platforms. Following is a high-level overview of the 50 most common web scraping tools and platforms currently available.

50 OF THE MOST COMMON WEB SCRAPING TOOLS

	NAME	DESCRIPTION
1	Apache Nutch	Apache Nutch is an extensible and scalable open-source web crawler software project.
2	A-Parser	A-Parser is a multithreaded parser of search engines, site assessment services, keywords and content.
3	Apify	Apify is a Node.js library similar to Scrapy and can be used for scraping libraries in JavaScript.
4	Artoo.js	Artoo.js provides script that can be run from your browser's bookmark bar to scrape a website and return the data in JSON format.
5	Blockspring	Blockspring lets users build visualizations from the most innovative blocks developed by engineers within your organization.
6	BotScraper	BotScraper is a tool for advanced web scraping and data extraction services that helps organizations from small and medium-sized businesses.
7	Cheerio	Cheerio is a library that parses HTML and XML documents and allows use of jQuery syntax while working with the downloaded data.
8	Colly	Colly is a scraping framework to extract structured data from websites.
9	Connotate	Connotate is a cloud-based web data extraction platform.
10	Content Grabber	Content Grabber is an automated web scraping tool that allows for pagination, scrolling pages and pop-ups and can process AJAX and JavaScript.
11	Crawljax	Crawljax is a tool for crawling JavaScript-based AJAX web applications.
12	Crawly	A web crawler that automatically extracts site content.
13	DataMiner	DataMiner is a web scraping tool that is an extension for Google Chrome and provides basic web scraping capabilities.
14	Data Scraper	Data Scraper is a simple web scraping tool for extracting data from a single page.
15	Dexi.io	Dexi.io supports data collection from any website and provides different types of robots to scrape data — crawlers, extractors, autobots and pipes.
16	Diffbot	Diffbot provides machine learning and computer vision algorithms and public APIs for web scraping.
17	Diggernaut	Diggernaut is a cloud-based service for web scraping, data extraction and other extract, load, transform (ELT) tasks.

50 OF THE MOST COMMON WEB SCRAPING TOOLS

NAME		DESCRIPTION
18	FMiner	FMiner is a visual web data extraction tool for web scraping and web screen scraping. In addition to the basic web scraping features, it also has AJAX/JavaScript processing and CAPTCHA solving.
19	GrabzIt	GrabzIt offers tools to capture screens and scrape websites.
20	Grepsr	Grepsr is a web scraping platform to source data from the web.
21	Helium Scraper	Helium Scraper is a multilevel page extractor tool.
22	iMacros	iMacros is a web automation, data extraction and web testing solution.
23	Import.io	Import.io provides easy-to-use web data extraction for price monitoring, lead generation, market research, big data analysis and more.
24	Jaunt	Jaunt is a Java library for web scraping and web automation.
25	Kimono	Kimono is a web crawler that turns websites into structured APIs.
26	Mozenda	Mozenda is an enterprise cloud-based web scraping platform that comprises two parts: an application to build the data extraction project and a web console to run agents, organize results and export data.
27	MyDataProvider	MyDataProvider is drop shipping and web scraping software for e-commerce.
28	Octoparse	Octoparse is a cloud-based web crawler and automated web scraping software.
29	Outwit Hub	Outwit Hub is a free data extractor built in a web browser and is available as an extension or stand-alone application.
30	ParseHub	ParseHub is a more advanced web-based scraping tool, which is built to crawl single and multiple websites that are using JavaScript, AJAX, cookies, sessions and redirects.
31	Playment	Playment is a fully managed solution offering training data for AI, transcription, data collection and enrichment services at scale.
32	PowerMapper	PowerMapper is a web crawler that automatically creates a site map of a website using thumbnails of each webpage.
33	PromptCloud	PromptCloud is a large-scale custom big data crawl and extraction solution.
34	PySpider	A Python web crawler that supports JavaScript pages and leverages a distributed architecture.

50 OF THE MOST COMMON WEB SCRAPING TOOLS

	NAME	DESCRIPTION
35	Scrape It	A Node.js web scraper.
36	ScrapeBox	ScrapeBox enables users to harvest URLs, verify backlinks and post mass comments and increase online presence.
37	Scraper	Scraper is a Chrome extension for scraping simple webpages. It is simple to use and will help users scrape a website's content and upload the results to Google Docs.
38	Scrape Hero	An advanced service web scraper/service that provides capabilities to bypass anti-bot technologies to collect structured and unstructured data.
39	Scrapy	Scrapy is an open-source web scraping framework in Python used to build web scrapers.
40	Scrapy Cloud (Scrapinghub)	A cloud-based web crawling platform that allows users to deploy custom-developed scrapers and scale them on demand.
41	sCrawler	sCrawler is web crawler software that enables users to collect data from directory websites.
42	Sitebulb	Website crawler for more comprehensive audits of webpages and out-of-the-box reporting capabilities.
43	Sphider	Sphider is a lightweight web spider and search engine written in PHP that uses MySQL as its back-end database.
44	SunTec	A more advanced tool to mine data from sites using JavaScript, jQuery, MooTools framework, etc., and bypass CAPTCHA systems.
45	Supplejack	Supplejack is a Ruby on Rails (RoR) development platform for managing the harvesting and manipulation of data.
46	Visual Scraper	Visual Scraper offers web scraping software and services to collect data from webpages.
47	Visual Web Ripper	Visual Web Ripper is a webpage scraper used to easily extract website data.
48	Web Scraper	Web Scraper is a Chrome extension free tool to scrape dynamic webpages.
49	WebHarvy	WebHarvy is a visual web scraper that can automatically scrape text, images, URLs and emails from websites.
50	Worth Web Scraping	Worth Web Scraping Services provides easy-to-integrate, high-quality data and metadata from hundreds of thousands of global online sources.

LEARN MORE ABOUT

THE BOT LANDSCAPE AND HOW TO TELL THE GOOD ONES FROM THE BAD