



DATASHEET · CAPTURE · WEBSITE SOURCES

It's all on your website. Who retypes it? Your website becomes your agent's knowledge.

Website sources read the public HTTPS pages of your website on a schedule. Each page becomes an internal article in the knowledge base you choose, and your agents answer from it on the phone and in website chat.

AT A GLANCE

Up to 2000 pages per source, link depth up to 6, a schedule from every 6 hours to monthly. Changed pages are rewritten, vanished pages are retired after two runs.

01 · WHAT IT IS ABOUT

Two versions of the same answer. Now your website sets the facts.

Opening hours, product pages and help texts have been on your website for ages. Retyping them for the agent takes time, and every change on the site makes the copy stale. Website sources read the pages themselves and bring changes along with the next run.

INPUT

Start URL with https, optional sitemap and more start URLs on the same host

RESULT

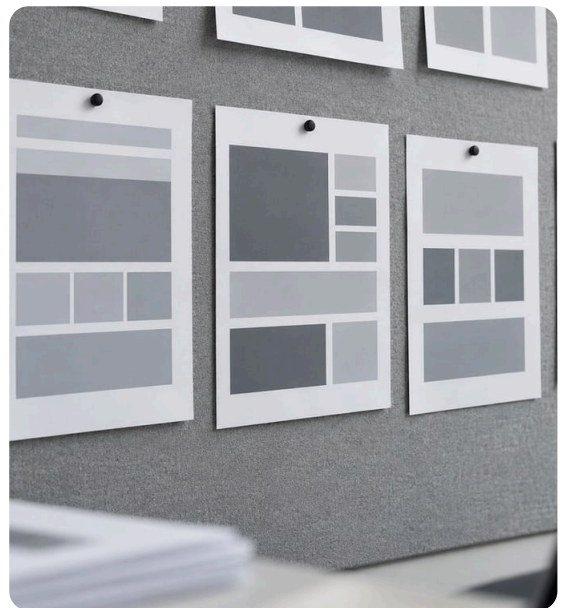
One internal article per page, with the page URL as source and the detected language

SCHEDULE

Every 6 or 12 hours, daily, weekly, every 14 days or monthly

READERS

Conversation agents on the phone and in website chat, chat agents in Agent AI

**One host, many pages**

From the start URL the crawler follows links on the same host, with or without www. Sitemaps from robots.txt and a sitemap of your own extend the list.

Set the scope

Maximum pages and link depth per source, plus path prefixes, exclude globs, URLs with a query string and list paths such as /tag.

JavaScript included

Fast reads the HTML, Full renders in a browser, Auto switches by itself when a page looks empty. A click selector opens FAQ accordions, for example.

Clean articles

CSS exclude keeps areas such as navigation or the footer out of the article. If you like, the article takes a short summary from the page.

The website's rules

robots.txt and noindex are respected. If an indexed page is blocked later, the crawler retires its article.

Changes, not full rewrites

The crawler spots changed pages by their content and leaves unchanged ones alone. Pages that vanish are retired after two runs without a hit.

02 · HOW IT WORKS

Add a source, check it, assign it to the agent.

A website source takes just a few fields. The preview shows what the crawler finds before the first run, and the catalog then shows what happened to each URL.

1

Add a source

Under **Knowledge Base AI › Website sources**, use **Add source**: enter name, start URL, target knowledge base and refresh interval.

2

Check the preview

The **Preview** shows robots.txt, sitemaps and candidates without indexing anything. Then set scope and crawler.

3

Run and monitor

The first run starts a few minutes after saving, and **Crawl now** schedules another. **Indexed pages** shows status and HTTP code per URL.

4

Assign it to the agent

Under **Conversation AI › Conversation Agents**, choose the knowledge base on the **Knowledge** tab. From then on the agent answers from the website articles.

TECHNICAL DATA

FEATURE	DETAILS
Input	Public HTTPS pages of one host, HTML text
Scope per source	Up to 2000 pages (default 200), link depth up to 6 (default 3)
Sources per account	Up to 25
Crawler	Fast, Full or Auto, with click and wait selectors
Output	Internal article per page with page URL, detected language and the tags website and host
Overview	Indexed pages with status, render, HTTP and last fetch; Sync history per source
Access and billing	Web app under Knowledge Base AI › Website sources ; billed by usage

03 · IN PRACTICE

Where website knowledge reaches phone and chat.

How-tos for the website chat

A software company keeps its how-tos under /docs. A website source with the path prefix /docs reads them daily, and the website chat names the matching page in its answer.

Product pages for the hotline

A retailer has its product pages read weekly. The conversation agent on the phone answers questions about sizes and materials from the same texts customers see online.

Website plus your own articles

Anything that is not on the website, such as a holiday closure, you add as your own article in the same knowledge base. The agent reads both together.

FREQUENTLY ASKED QUESTIONS

Which pages does a website source read?

Public HTTPS pages of one host, from the start URL and, if you like, via sitemaps and further start URLs. It reads the text of the HTML pages, with no login, cookies or credentials.

What happens to my own edits to a website article?

When the page changes, the next run rewrites the article. Additions therefore belong in an article of their own in the same knowledge base.

How current is the agent's knowledge?

As current as the last run. You choose the schedule from every 6 hours to monthly, and **Crawl now** schedules an extra run. Two manual runs are 15 minutes apart.

Do the pages appear in the customer portal?

Website articles are internal: your team and the agents of the knowledge base read them, and they stay out of the customer portal.

GET STARTED

Make your website your first source.

In the web app, go to **Knowledge Base AI › Website sources**, use **Add source** to enter the start URL, check the preview and save. The first run starts a few minutes later.

paperoffice.ai/en/ai-web-sources-website-crawler-knowledge



MORE ABOUT THE MODULE

paperoffice.ai/en/ai-web-sources-website-crawler-knowledge

PRICING

paperoffice.ai/en/pricing

CONTACT

paperoffice.ai/en/contact