

NOSIBLE Scrape URL - Data Dictionary V2.1.0

As of: 2026-07-20

This document describes every input and output field of the NOSIBLE Scrape URL endpoint so that a technical reader can understand the API at a glance. Scrape URL ([POST https://www.nosible.ai/search/v2/scrape-url](https://www.nosible.ai/search/v2/scrape-url)) turns any webpage into beautiful JSON: provide a target URL and NOSIBLE returns a clean, structured document containing the article text, metadata, authors, publish dates, snippets, and more. Cached copies are served when available unless you force a [recrawl](#). If the URL fails NOSIBLE's preflight checks or cannot be fetched, the request fails with HTTP 400.

Interactive API documentation: this endpoint is documented live in the NOSIBLE Swagger UI at <https://www.nosible.ai/search/v2/docs/> (OpenAPI spec: <https://www.nosible.ai/search/v2/openapi.json>). The version of this data dictionary tracks the API version published there: this document describes NOSIBLE V2.1.0.

Contents

Request Headers.....	2
Input Fields.....	2
Output Fields.....	3
Example Request and Response.....	7

Request Headers

Every request to this endpoint is an HTTP POST with a JSON body. Authentication is handled by a single header, documented below. You should also send **Content-Type: application/json**, and **Accept-Encoding: gzip** is recommended to cut transfer sizes.

Header	Details
api-key	Definition: Your NOSIBLE Search API key. To get an API key please visit app.nosible.com , where you can manage organizations, users, subscriptions, and API keys. Type: string Length: 42–50 characters Required: Yes - requests without a valid key fail with HTTP 422. Example: "nos_sk_XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX"

Input Fields

All input fields are passed in the JSON body of the POST request. Only **url** matters in most cases; the rest control how the page is fetched.

Input Field	Details
url	Definition: The URL that you would like NOSIBLE to scrape, process, and turn into JSON. The URL must pass NOSIBLE's preflight checks; otherwise the request fails with HTTP 400. Type: string Default: "https://nosible.ghost.io/ensemble-and-distil/" Example: "https://www.rainforest-times.com/birdwing-sanctuary"
html	Definition: The HTML associated with this URL, if you already have it. When provided, NOSIBLE parses your HTML instead of fetching the page. Type: string Default: "" Example: "<html><head>...</head><body>...</body></html>"
render	Definition: Force the crawler to render the JavaScript on the page

Input Field	Details
	before extracting content. Type: Boolean Default: <code>false</code> Example: <code>true</code>
<code>recrawl</code>	Definition: Whether to re-crawl the URL again instead of serving NOSIBLE's cached copy of the page. Type: Boolean Default: <code>false</code> Example: <code>true</code>

Output Fields

A successful scrape returns HTTP 200 with a small envelope around the structured document. The document itself has nine blocks; the dotted rows below describe each one.

Output Field	Details
<code>message</code>	Definition: A human-readable acknowledgement. Type: string Example: <code>"Document retrieved"</code>
<code>added_to_batch</code>	Definition: Whether the page was queued for indexing into the search engine. Type: Boolean Example: <code>false</code>
<code>response</code>	Definition: The structured document for the page. Its nine blocks are described in the rows below. Type: object Example: <pre>{ "request": {...}, "page": {...}, "statistics": {...}, "languages": {...}, "snippets": {...}, "full_text": "...", "metadata": {...}, "structured": [...],</pre>

Output Field	Details
	<pre>"url_tree": {...} }</pre>
response.request	Definition: The URL preflight information for the response URL: how NOSIBLE parsed, normalised, and geolocated the URL you submitted. Type: object Fields: <code>raw_url</code> (str) - the URL exactly as submitted <code>url</code> (str) - the cleaned, normalised URL <code>hash</code> (str) - the 24-character url_hash of the page <code>geo</code> (str) - the inferred country of the site, ISO 3166-1 alpha-2 <code>proxy</code> (str) - the proxy region used to fetch the page <code>scheme</code> / <code>netloc</code> / <code>prefix</code> / <code>domain</code> / <code>suffix</code> / <code>path</code> / <code>query</code> / <code>fragment</code> (str) - the parsed URL components <code>query_allowed</code> / <code>query_blocked</code> (dict) - query parameters kept/stripped during normalisation Example: <pre>{ "raw_url": "https://www.rainforest-times.com/birdwing-sanctuary", "url": "https://www.rainforest-times.com/birdwing-sanctuary", "hash": "aB3dE_fG7hIj-KlMnOpQrStU", "geo": "PG", "netloc": "www.rainforest-times.com", ... }</pre>
response.page	Definition: Cleaned page-level metadata. Type: object Fields: <code>title</code> (str) - the page title/headline <code>description</code> (str) - the page description <code>author</code> (str) - the author, or empty string <code>published</code> (str) - publish date, YYYY-MM-DD (3000-01-01 if unknown) <code>modified</code> (str) - modified date, YYYY-MM-DD, or empty string <code>visited</code> (str) - the date NOSIBLE crawled the page, YYYY-MM-DD <code>certain</code> (bool) - whether the publish date is trusted <code>published_timestamp</code> (str) - full ISO publish timestamp, or empty <code>modified_timestamp</code> (str) - full ISO modified timestamp, or empty <code>visited_timestamp</code> (str) - full ISO UTC crawl timestamp

Output Field	Details
	Example: <pre>{ "title": "New sanctuary offers hope for the world's largest butterfly", "author": "Jane Doe", "published": "2025-08-12", "visited": "2026-06-30", "certain": true, ... }</pre>
response.statistics	Definition: Document-level aggregate counts rolled up from all snippets. Type: object Fields: snippets, sentences, words, characters, images, videos, audio, tables, lists, blocks, links, files - all integers. Example: <pre>{ "snippets": 14, "sentences": 96, "words": 1843, "characters": 11208, "images": 4, "links": 37, ... }</pre>
response.languages	Definition: The languages detected on the page: a mapping from ISO 639-1 code to the proportion of characters in that language (sums to roughly 1.0). Type: object Example: <pre>{ "en": 0.978, "fr": 0.022 }</pre>
response.snippets	Definition: The embeddable text chunks of the page: a mapping from snippet_hash to snippet object. Snippets form a linked list via their prev/next hashes. Type: object Fields per snippet:

Output Field	Details
	<code>url_hash</code> (str) - the parent page's hash <code>snippet_hash</code> (str) - this snippet's hash (also the key) <code>prev_snippet_hash</code> / <code>next_snippet_hash</code> (str/null) - neighbouring snippets <code>content</code> (str) - the cleaned chunk text <code>words</code> (str) - space-joined normalized words <code>language</code> (str) - the snippet's language <code>statistics</code> (dict) - sentences, words, characters (+ links when present) <code>links</code> / <code>images</code> / <code>videos</code> / <code>audio</code> / <code>files</code> / <code>tables</code> / <code>lists</code> / <code>blocks</code> (list, optional) - media and elements associated with this snippet Example: <pre>{ "kY8pQ2nWjR4tXz-Lm1CvDbEa": { "content": "The Queen Alexandra's Birdwing ...", "language": "en", "statistics": { "sentences": 4, "words": 87, "characters": 512 }, ... }, ... }</pre>
<code>response.full_text</code>	Definition: The whole page's extracted text, newline-joined. Type: string Example: "New sanctuary offers hope... \nConservationists in Papua New Guinea..."
<code>response.metadata</code>	Definition: The raw HTML meta tags of the page. Single-token keys map directly to string values; multi-token keys are grouped into sub-dicts by their first token (e.g. <code>og</code>, <code>twitter</code>). Also carries a cleaned <code>nosible</code> sub-dict with: <code>html_lang</code>, <code>title</code>, <code>description</code>, <code>published</code>, <code>modified</code>, <code>author</code>, <code>published_timestamp</code>, <code>modified_timestamp</code>. Type: object Example:

Output Field	Details
	<pre> { "viewport": "width=device-width", "og": { "title": "New sanctuary offers hope ...", "type": "article" }, "nosible": { "title": "New sanctuary offers hope ...", "published": "2025-08-12", ... } } </pre>
response.structured	Definition: The raw JSON-LD (application/ld+json) objects extracted from the page's script tags. Structure varies per site. Type: array of objects Example: <pre> [{ "@type": "NewsArticle", "headline": "New sanctuary offers hope ...", ... }] </pre>
response.url_tree	Definition: A nested URL prefix tree of up to roughly 30 similar on-page URLs, space-compressed so that many links can be stored compactly. Type: object Example: <pre> { "https://www.rainforest-times.com": { "/birdwing-sanctuary": {}, "/about": {}, ... } } </pre>

Example Request and Response

The example below scrapes a page end to end. The response is shown truncated. All field values are illustrative.

Example Request:

```
import requests

response = requests.post(
    url="https://www.nosible.ai/search/v2/scrape-url",
    json={
        "url": "https://www.rainforest-times.com/birdwing-sanctuary"
    },
    headers={
        "Accept-Encoding": "gzip",
        "Content-Type": "application/json",
        "api-key": "YOUR_NOSIBLE_API_KEY"
    }
)

print(response.json())
```

Example Response:

```
{
  "message": "Document retrieved",
  "added_to_batch": false,
  "response": {
    "request": {
      "raw_url": "https://www.rainforest-times.com/birdwing-sanctuary",
      "url": "https://www.rainforest-times.com/birdwing-sanctuary",
      "hash": "aB3dE_fG7hIj-KlMnOpQrStU",
      "geo": "PG",
      "netloc": "www.rainforest-times.com",
      ...
    },
    "page": {
      "title": "New sanctuary offers hope for the world's largest butterfly",
      "description": "Conservationists in Papua New Guinea have secured
                     a 10,000-hectare corridor.",
      "author": "Jane Doe",
      "published": "2025-08-12",
      "modified": "2025-08-13",
      "visited": "2026-06-30",
      "certain": true,
      "published_timestamp": "2025-08-12T09:30:00+10:00",
      "modified_timestamp": "2025-08-13T08:00:00+10:00",
      "visited_timestamp": "2026-06-30T14:12:05.421Z"
    },
    "statistics": {
      "snippets": 14,
      "sentences": 96,
      "words": 1843,
      "characters": 11208,
      "images": 4,
      "videos": 0,
      "audio": 0,
    }
  }
}
```

```

    "tables": 0,
    "lists": 2,
    "blocks": 1,
    "links": 37,
    "files": 0
  },
  "languages": {
    "en": 0.978,
    "fr": 0.022
  },
  "snippets": {
    "kY8pQ2nWjR4tXz-Lm1CvDbEa": {
      "url_hash": "aB3dE_fG7hIj-KlMnOpQrStU",
      "snippet_hash": "kY8pQ2nWjR4tXz-Lm1CvDbEa",
      "prev_snippet_hash": null,
      "next_snippet_hash": "zW3xV9bN2mK5jH8gF1dS4aPo",
      "content": "The Queen Alexandra's Birdwing, the world's
                  largest butterfly, now has a dedicated
                  sanctuary in Oro Province.",
      "words": "the queen alexandra s birdwing the world s ...",
      "language": "en",
      "statistics": {
        "sentences": 4,
        "words": 87,
        "characters": 512
      }
    },
    ...
  },
  "full_text": "New sanctuary offers hope ...",
  "metadata": {
    "og": {
      "title": "New sanctuary offers hope ...",
      "type": "article"
    },
    "nosible": {
      "title": "New sanctuary offers hope ...",
      "published": "2025-08-12",
      ...
    },
    ...
  },
  "structured": [
    {
      "@type": "NewsArticle",
      "headline": "New sanctuary offers hope ...",
      ...
    }
  ],
  "url_tree": {
    "https://www.rainforest-times.com": {
      "/birdwing-sanctuary": {},
      "/about": {},
      ...
    }
  }
}

```

}
}
}
}