NOSIBLE Bulk Search - Data Dictionary V2.1.0

As of: 2026-07-20

This document describes every input and output field of the NOSIBLE Bulk Search endpoint so that a technical reader can understand the API at a glance. Bulk Search (`POST https://www.nosible.ai/search/v2/bulk-search`) gives access to the long tail: up to 10,000 results per query instead of the top 100. It runs asynchronously. The endpoint immediately returns a pointer to an encrypted file on S3 and the results materialise there 15–45 seconds later. Requests are authenticated with the `api-key` header. Only one bulk search may run per account at a time; submitting a second while one is running returns HTTP 409.

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

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. Every field is optional: any field you omit falls back to the default shown in its row. In practice, `question` is the only field you must think about; everything else tunes recall, filtering, and snippet length. Note that n_results starts at 1,000: for fewer results, use Fast Search or Rich Search instead.

Input Field	Details
`question`	**Definition:** The question you would like your search results to answer. This is the primary query that drives retrieval. **Type:** string **Length:** 1–500 characters **Default:** `"Why did Argus Filch confiscate the Marauder's Map?"` **Example:** `"What conservation efforts exist for the Queen Alexandra's Birdwing butterfly?"`
`expansions`	**Definition:** Semantically equivalent but lexically diverse rephrasings of the question. Optional and recommended: including ten expansions materially improves recall. **Type:** array of strings **Items:** 0–10 **Default:** `[]` **Example:**

Input Field	Details

	`[`
	`  "How is the Queen Alexandra's Birdwing protected in PNG?",`
	`  "What threats face the largest butterfly species?"`
	`]`

`instruction`	**Definition:** The instruction passed along to the embedding model. USE WITH CAUTION. **Type:** string **Length:** 1–500 characters **Default:** `"Retrieve semantically similar text."` **Example:** `"Retrieve semantically similar text."`
`sql_filter`	**Definition:** A SQL filter to narrow down your search results. **Type:** string **Default:** `"SELECT loc FROM engine"` **Columns:** `netloc` - include or exclude specific websites from your search results `published` - a YYYY-MM-DD string for filtering on publish dates `modified` - a YYYY-MM-DD string for filtering on modified dates `visited` - a YYYY-MM-DD string for filtering on visited/crawled dates `language` - e.g. en-en (english document, english snippet) `stot` - include or exclude documents based on total number of snippets `snum` - include or exclude snippets based on their index on the page `words` - include or exclude snippets based on their word count `chars` - include or exclude snippets based on their character count **Example:** `"SELECT loc FROM engine WHERE words>=30"`
`algorithm`	**Definition:** The algorithm to use when scoring search results. **Type:** string **Default:** `"hybrid-3"` - the recommended choice **Options:** `string`, `lexical`, `baseline`, `hamming`, `hybrid-1`, `hybrid-2`, `hybrid-3`, `company` **Note:** `company` is remapped to `hybrid-3` by the engine. **Example:** `"hybrid-3"`
`min_similarity`	**Definition:** The minimum similarity that a result must have to be returned. **Type:** float **Range:** 0.0–1.0 **Default:** `0.0`

Input Field	Details
	Example: `0.65`

`must_include`	**Definition:** Any of the following strings must be included in each result. Be careful when using this: aggressive lists can filter out everything.
	Type: array of strings
	Items: 0–100
	Default: `[]`
	Example:
	`[`
	`  "Ornithoptera alexandrae"`
	`]`

`must_exclude`	**Definition:** All of the following strings must be excluded from each result. Be careful when using this: aggressive lists can filter out everything.
	Type: array of strings
	Items: 0–100
	Default: `[]`
	Example:
	`[`
	`  "casino",`
	`  "betting"`
	`]`

`brand_safety`	**Definition:** The brand safety level to apply to the source websites. **Type:** string or null **Options:** `Safe`, `Sensitive`, `Unsafe` **Default:** `null` **Example:** `"Safe"`
`language`	**Definition:** The language you would like your search results to be refined to. **Type:** string or null **Options:** one of 94 ISO 639-1 two-letter codes, e.g. `en`, `fr`, `de`, `es`, `pt`, `zh`, `ja`, `ar` **Default:** `null` **Example:** `"en"`
`continent`	**Definition:** The continent you would like your search results to be refined to. **Type:** string or null

Input Field	Details
	Options: `Africa`, `Asia`, `Europe`, `North America`, `Oceania`, `South America`, `Worldwide` **Default:** `null` **Example:** `"Oceania"`
`region`	**Definition:** The geographic region you would like your search results to be refined to. **Type:** string or null **Options:** one of 31 named world regions, e.g. `Western Europe`, `Middle East`, `Southeast Asia`, `Sub-Saharan Africa`, `The Balkans`, `The Caucasus` **Default:** `null` **Example:** `"Southeast Asia"`
`country`	**Definition:** The country you would like your search results to be refined to. **Type:** string or null **Options:** one of 203 full English country names (including `Worldwide`), e.g. `United States`, `Germany`, `South Africa` **Default:** `null` **Example:** `"Papua New Guinea"`
`sector`	**Definition:** The GICS sector you would like your search results to be refined to. **Type:** string or null **Options:** `Other`, `Communication Services`, `Consumer Discretionary`, `Consumer Staples`, `Energy`, `Financials`, `Health Care`, `Industrials`, `Information Technology`, `Materials`, `Real Estate`, `Utilities` **Default:** `null` **Example:** `"Materials"`
`industry_group`	**Definition:** The GICS industry group you would like your search results to be refined to. **Type:** string or null **Options:** one of 26 GICS industry groups, e.g. `Banks`, `Capital Goods`, `Media & Entertainment`, `Software & Services` **Default:** `null` **Example:** `"Food, Beverage & Tobacco"`

Input Field	Details
`industry`	**Definition:** The GICS industry you would like your search results to be refined to. **Type:** string or null **Options:** one of 75 GICS industries, e.g. `Biotechnology`, `Aerospace & Defense`, `Automobiles` **Default:** `null` **Example:** `"Biotechnology"`
`sub_industry`	**Definition:** The GICS sub-industry you would like your search results to be refined to. **Type:** string or null **Options:** one of 164 GICS sub-industries, e.g. `Advertising`, `Apparel Retail`, `Aluminum` **Default:** `null` **Example:** `"Environmental & Facilities Services"`
`iab_tier_1`	**Definition:** The tier 1 IAB content category you would like your search results to be refined to. **Type:** string or null **Options:** one of 39 tier 1 IAB categories, e.g. `Automotive`, `Business and Finance`, `Technology & Computing` **Default:** `null` **Example:** `"Science"`
`iab_tier_2`	**Definition:** The tier 2 IAB content category you would like your search results to be refined to. **Type:** string or null **Options:** one of 388 tier 2 IAB categories **Default:** `null` **Example:** `"Environment"`
`iab_tier_3`	**Definition:** The tier 3 IAB content category you would like your search results to be refined to. **Type:** string or null **Options:** one of 644 tier 3 IAB categories **Default:** `null` **Example:** `"Environment"`
`iab_tier_4`	**Definition:** The tier 4 IAB content category you would like your search results to be refined to.

Input Field	Details
	Type: string or null **Options:** one of 704 tier 4 IAB categories **Default:** `null` **Example:** `"Nature"`

`companies`	**Definition:** A list of company names. This will be used to refine the search results.
	Type: array of strings
	Items: 0–3
	Default: `[]`
	Example:
	`[`
	`  "Tesla"`
	`]`

`collection`	**Definition:** The collection to search over. Use `this-week` to search only recently added data. **Type:** string **Options:** `everything`, `this-week` **Default:** `"everything"` **Example:** `"this-week"`
`deduplicate`	**Definition:** Whether to deduplicate search results based on duplicate headlines. **Type:** Boolean **Default:** `false` **Example:** `true`
`n_results`	**Definition:** The number of search results that you would like to be returned. **Type:** integer **Range:** 1,000–10,000 **Default:** `1000` **Example:** `5000`
`n_probes`	**Definition:** The number of shards (collections) your search is evaluated over. More probes give deeper coverage at higher latency. **Type:** integer **Range:** 5–300 **Default:** `10` **Example:** `100`

Input Field	Details
`n_contextify`	**Definition:** The number of context words you would like back per search result. Controls snippet lengths. **Type:** integer **Range:** 128–1024 **Default:** `128` **Example:** `512`
`internal_use`	**Definition:** Internal-only. Used to enable hidden features and new algorithms from time to time. Hidden from the public API docs. **Type:** object **Default:** `{}` **Example:** `{}`

Immediate Response Fields

Bulk Search does not return results directly. A successful submission returns HTTP 200 with a small JSON object that tells you where your encrypted results file will appear and how to decrypt it.

Output Field	Details
`message`	**Definition:** A human-readable acknowledgement that the search was accepted. **Type:** string **Example:** `"Slow search task accepted."`
`decrypt_using`	**Definition:** The Fernet encryption key (derived from a hash of your API key) that decrypts the downloaded results file. **Type:** string **Example:** `"3vGx4kJ9pQ7rT2wY8zB1cD5fH6mN0aSuVeXgZiLqKo8="`
`download_from`	**Definition:** The S3 (Wasabi) URL where your encrypted results file will appear. The file name ends in `.json.zstd.fernet` and materialises 15–45 seconds after submission. S3 returns HTTP 404 until the file exists, so poll it in a loop. **Type:** string **Note:** A gzip-encrypted twin of the same payload is uploaded alongside it with the extension `.json.gzip.fernet` if you prefer gzip over

Output Field	Details
	Zstandard. **Example:** `"https://s3.eu-west-1.wasabisys.com/finweb-results/Xy12Ab...Cd34.json.zstd.fernet"`

Downloaded Result Fields

To retrieve your results: poll `download_from` until it returns HTTP 200 (it returns 404 while the search is still running), decrypt the payload with the Fernet key in `decrypt_using`, zstd-decompress it, and parse the JSON. The decrypted payload stored on Wasabi is a JSON object with three top-level keys: `message`, `query`, and `response` - the same envelope the synchronous search endpoints use. The rows below describe all three, plus every field of one search result.

Output Field	Details
`message`	**Definition:** A human-readable summary of how much of the index was evaluated and how long the search took. **Type:** string **Example:** `"Evaluated 48,120,334 out of 3,201,554,096 records in 32.512 seconds."`
`query`	**Definition:** An echo of how NOSIBLE interpreted your search: your inputs plus the retrieval strategy chosen by the engine. **Type:** object **Fields:** `instruction / question / expansions` - your query inputs, as interpreted `brand_safety / language / continent / region / country` (str/null) - the applied filters `sector / industry_group / industry / sub_industry` (str/null) - the applied GICS filters `iab_tier_1 / iab_tier_2 / iab_tier_3 / iab_tier_4` (str/null) - the applied IAB filters `algorithm / n_results / n_contextify / n_probes / min_similarity / pass_similarity / must_exclude / must_include / companies` - the applied hyperparameters `shard_selector / shard_reranker` (str) - internal routing strategies

Output Field	Details

	`search_lang` (str) - the language the query was searched in
	`search_words` (list) - the exact normalized words matched in the index
	`search_intents` (list) - the predicted intents of the query
	Example:
	`{`
	`  "question": "Conservation efforts for the Queen`
	`  Alexandra's Birdwing?",`
	`  "algorithm": "hybrid-3",`
	`  "n_results": 1000,`
	`  "search_lang": "en",`
	`  "search_words": ["conservation", "efforts", ...],`
	`  ...`
	`}`

`response`	**Definition:** The search results themselves: an array of flat search result objects (the same format as Fast Search), with at most n_results entries, sorted by descending similarity. All fields below describe one search result.
	Type: array of objects
	Example:
	`{`
	`  "response": [`
	`  {...},`
	`  {...},`
	`  ...`
	`  ]`
	`}`

`url_hash`	**Definition:** A hash that uniquely identifies the page: an 18-byte SHAKE-128 digest of the normalised URL, URL-safe base64 encoded (24 characters). Used as the document key across NOSIBLE systems. **Type:** string **Example:** `"aB3dE_fG7hIj-KlMnOpQrStU"`
`url`	**Definition:** The full URL of the source page. **Type:** string **Example:** `"https://www.rainforest-times.com/2025/08/12/birdwing-sanctuary"`
`netloc`	**Definition:** The network location (domain) of the source. **Type:** string

Output Field	Details
	Example: `"www.rainforest-times.com"`
`published`	**Definition:** The date the page was published, as a YYYY-MM-DD string. **Type:** string **Example:** `"2025-08-12"`
`visited`	**Definition:** The date NOSIBLE visited/crawled the page, as a YYYY-MM-DD string. **Type:** string **Example:** `"2026-06-30"`
`language`	**Definition:** The language of the document as an ISO 639-1 two-letter code. **Type:** string **Example:** `"en"`
`author`	**Definition:** The author of the page, when one could be detected. **Type:** string **Example:** `"Jane Doe"`
`title`	**Definition:** The title/headline of the page. **Type:** string **Example:** `"New sanctuary offers hope for the world's largest butterfly"`
`description`	**Definition:** The description of the page. **Type:** string **Example:** `"Conservationists in Papua New Guinea have secured a 10,000-hectare corridor for the Queen Alexandra's Birdwing."`
`best_chunk`	**Definition:** The single snippet of the page with the highest similarity to your question. **Type:** string **Example:** `"The Queen Alexandra's Birdwing, the world's largest butterfly, now has a dedicated sanctuary in Oro Province."`
`content`	**Definition:** The contextified content: the best snippet expanded with surrounding words from the page, up to `n_contextify` words.

Output Field	Details
	Type: string **Example:** `"... The Queen Alexandra's Birdwing, the world's largest butterfly, now has a dedicated sanctuary in Oro Province. The corridor links two forest reserves ..."`
`semantics`	**Definition:** Retrieval diagnostics for the result. **Type:** object **Fields:** `origin_shard` (int) - the shard the document was retrieved from `chunks_total` (int) - total chunks in the document `chunks_matched` (int) - chunks that matched the query `chunks_kept` (int) - chunks kept after filtering `similarity` (float) - the similarity score, between 0 and 1 **Example:** `{ "origin_shard": 143, "chunks_total": 42, "chunks_matched": 18, "chunks_kept": 6, "similarity": 0.874213 }`

Example Request and Response

The example below runs a bulk search end to end: submitting the search, receiving the S3 pointer, then downloading, decrypting, and decompressing the results. All field values are illustrative.

Example Request:

```
import requests

response = requests.post(
    url="https://www.nosible.ai/search/v2/bulk-search",
    json={
        "question": "Conservation efforts for the Queen Alexandra's Birdwing?",
        "algorithm": "hybrid-3",
        "n_results": 1000,
        "n_probes": 100,
        "n_contextify": 128,
        "sql_filter": "SELECT loc FROM engine WHERE words>=30"
```

```
    },
    headers={
        "Accept-Encoding": "gzip",
        "Content-Type": "application/json",
        "api-key": "YOUR_NOSIBLE_API_KEY"
    }
)

print(response.json())
```

Example Immediate Response:

```
{
    "message": "Slow search task accepted.",
    "decrypt_using": "3vGx4kJ9pQ7rT2wY8zB1cD5fH6mN0aSuVeXgZiLqKo8=",
    "download_from": "https://s3.eu-west-1.wasabisys.com/finweb-results/
                      Xy12Ab...Cd34.json.zstd.fernet"
}
```

Example Retrieval (polling S3):

```
import json
import time

import requests
import zstandard as zstd
from cryptography.fernet import Fernet

download_details = response.json()
download_from = download_details["download_from"]
decrypt_using = download_details["decrypt_using"]

# Poll S3 until the file materialises (15-45 seconds).
# S3 returns 404 until the worker has uploaded the file.
for attempt in range(30):
    time.sleep(15)
    s3_response = requests.get(download_from, timeout=60)
    if s3_response.status_code == 200:
        break
else:
    raise TimeoutError("Bulk search did not materialise in time.")

fernet = Fernet(decrypt_using.encode("utf-8"))
decrypted_data = fernet.decrypt(s3_response.content)

decompressor = zstd.ZstdDecompressor()
decompressed_data = decompressor.decompress(decrypted_data)

payload = json.loads(decompressed_data)
print(json.dumps(payload, indent=4))
```

Example Decrypted Payload (the final JSON stored on Wasabi, truncated to one result):


```
{
    "message": "Evaluated 48,120,334 out of 3,201,554,096 records
                in 32.512 seconds.",
    "query": {
        "instruction": "Retrieve semantically similar text.",
        "question": "Conservation efforts for the Queen Alexandra's Birdwing?",
        "expansions": [],
        "brand_safety": null,
        "language": null,
        "continent": null,
        "region": null,
        "country": null,
        "sector": null,
        "industry_group": null,
        "industry": null,
        "sub_industry": null,
        "iab_tier_1": null,
        "iab_tier_2": null,
        "iab_tier_3": null,
        "iab_tier_4": null,
        "algorithm": "hybrid-3",
        "n_results": 1000,
        "n_contextify": 128,
        "n_probes": 100,
        "min_similarity": 0.0,
        "pass_similarity": 1.0,
        "must_exclude": [],
        "must_include": [],
        "shard_selector": "...",
        "shard_reranker": "...",
        "search_lang": "en",
        "search_words": ["conservation", "efforts", "queen", ...],
        "search_intents": [...],
        "companies": []
    },
    "response": [
        {
            "url_hash": "aB3dE_fG7hIj-KlMnOpQrStU",
            "url": "https://www.rainforest-times.com/birdwing-sanctuary",
            "netloc": "www.rainforest-times.com",
            "published": "2025-08-12",
            "visited": "2026-06-30",
            "language": "en",
            "author": "Jane Doe",
            "title": "New sanctuary offers hope for the world's largest butterfly",
            "description": "Conservationists in Papua New Guinea have secured a
                            10,000-hectare corridor.",
            "best_chunk": "The Queen Alexandra's Birdwing ...",
            "content": "... The corridor links two forest reserves ...",
            "semantics": {
                "origin_shard": 143,
                "chunks_total": 42,
```

```
                "chunks_matched": 18,
                "chunks_kept": 6,
                "similarity": 0.874213
            }
        },
        ... up to n_results results ...
    ]
}
```