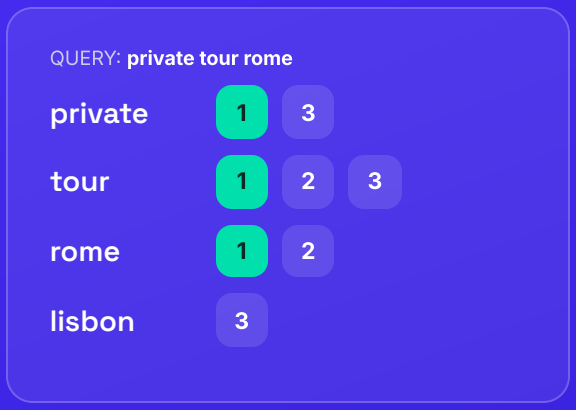
seogirl

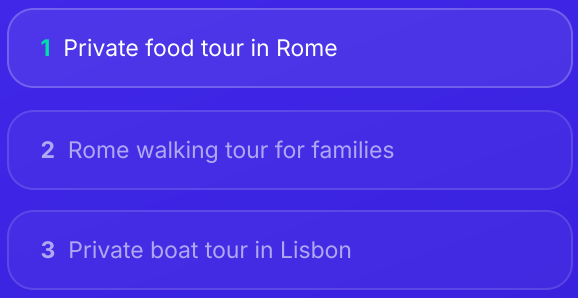
How Search Indexing Works

"Indexed" isn't a yes or no. What a search index actually is, how words get into it, and the practical SEO checks that follow from it.

Explainer · v0.1 · from posting lists to passages

AN INVERTED INDEX: WORDS FIRST, THEN THE PAGES THEY APPEAR ON





Only page 1 is on all three lists.

3	4	4
separate stages	layers: words, positions, passages, vectors	checks I run because of it

Ameneh Saeednia
Senior SEO and Generative Search Strategist

The basics

Crawling, indexing and serving are three jobs

Search Console says “Indexed” or “Not indexed,” and it's easy to treat that as the whole story. But an index isn't a list of approved URLs. It's a **data structure built to find the right documents for a query, fast**.

1

Crawling

Fetching the page. A URL can be **discovered and fetched** and still go no further.

2

Indexing

Analyzing what's on the page and **storing it in a searchable form**. Similar pages are grouped and one is chosen as canonical.

3

Serving

Finding and **ranking stored pages** when someone searches. Indexed pages can still almost never be shown.

The book index picture

Google explains its index with the **index at the back of a book**, and says it covers hundreds of billions of pages, well over 100,000,000 gigabytes. [1]

The important part

A book index lists **words, then the pages they appear on**. Not pages, then words. Search engines work the same way, and that reversal is the whole trick.

Why I care

This is the same reason I keep saying a **sitemap is not the website**: getting a URL discovered is only the first of several separate steps.

Layer 1: words

Forward index vs. inverted index

Take three tiny pages on a travel site: **1** “Private food tour in Rome,” **2** “Rome walking tour for families,” **3** “Private boat tour in Lisbon.”

Forward index	Words on the page
Page 1	private, food, tour, rome
Page 2	rome, walking, tour, families
Page 3	private, boat, tour, lisbon

Inverted index	Posting list
private	1, 3
tour	1, 2, 3
rome	1, 2
lisbon	3

Posting lists

Each row of the inverted index is a **posting list**: the pages where a term appears. To answer “private tour rome,” the engine jumps to those three lists and keeps pages on all of them. Here, page 1.

Why it's fast

This step is called **intersecting posting lists**. The engine never reads pages that don't contain the words. A forward index would have to scan every page, every time.

A good old source

One of the clearest explanations is a **2005 Google article for librarians** by Matt Cutts. The tools have changed. Matching first and ranking second hasn't. [2]

Before words are stored

Normalizing text, and the stop-word myth

Text is broken into tokens and normalized before it goes into the index, so that variations can still match.

Step	What it does	Example
Stemming	Cuts words to a rough root	running → run
Lemmatization	Maps words to their dictionary form	ran → run
Stop-word removal	Drops very common words	Mostly **not used** by web search engines today

Why small words stay

The standard information retrieval textbook notes that **web search engines generally don't use stop lists**. [3] “Flights to Paris” and “flights from Paris” are different searches. So are band names, book titles and product names.

What I do with that

I don't strip “small” words from titles or headings to make them look more keyword-dense. **They're doing work.** Exact names stay exact.

Layers 2 and 3

Positions, then passages

A simple posting list says **which** pages contain a word. Real indexes also store **where** it appears, and that changes what can be matched.

Positions

Stored positions make **phrase and proximity matching** possible. “Civil war” as neighbouring words means more than *civil* in one paragraph and *war* in another. The 2005 article also noted that words in a page title hint at its topic. [2]

Repetition

Scoring functions like **BM25** reward a term's presence, then quickly stop rewarding repetition. That's the math behind why keyword stuffing never worked as well as people thought.

Passages

In 2020 Google said it could better understand **individual passages** within a page, affecting around **7% of queries**. [4] Pages are still indexed as pages, but one section can be what makes a page relevant.

What that means for long guides

A great answer in paragraph 14, under a vague heading, is harder to match than the same answer in **a section built around it**. My answer-first retrieval test looked at exactly this.

Layer 4: meaning

Then vectors joined the index

The inverted index matches **words**. Vector search matches **meaning**: text becomes embeddings, and the engine finds content close to the query even when the words don't overlap.

Method	Matches on	Good at	Misses
Inverted index (BM25)	Exact terms	Names, codes, specific phrases	Different wording, same meaning
Vector search	Meaning	Paraphrases, related concepts	Exact specifics, rare terms
Hybrid retrieval	Both, then reranked	Most real queries	Still only what's on the page

The inverted index didn't go away. It got a partner.

Methods like **MUVERA** make the vector side fast enough at scale, and **RAG** systems behind AI answers still start with retrieval: find the right pieces first, then generate.

In practice

What I check because of this

Understanding the index changed four habits for me.

1

Ask “indexed for what?”

A page can only match searches whose words or meaning are **actually on it**. Facts inside images, skipped PDFs or late-loading tabs may not be in the index the way you assume.

2

One question per section

Clear headings with **the answer near the top** give passage-level matching something to work with.

3

Keep exact names exact

Product, place and brand names should appear **the way people search for them**, small words included.

4

Look for near-duplicates

Google groups similar pages and **chooses one canonical**. Two pages answering one question rarely double your chances; review overlaps before writing a third.

One claim I'd drop

Old lessons often say “Google indexes less than 5% of the internet.” I've found **no reliable, current source** for it, and it mixes private, logged-in, duplicate and low-value pages into one number.

Sources

Where these explanations come from

01 **How Google Search organizes information. Google**
The book-index analogy and the size of the Google index.
https://www.google.com/search/howsearchworks/how-search-works/organizing-information/

02 **Google librarian article on crawling, indexing and ranking, 2005 (archived)**
Posting lists, intersecting lists, word proximity and title hints. Historical context.
https://web.archive.org/web/20080701150733/https://www.google.com/librariancenter/articles/0512_01.html

03 **Introduction to Information Retrieval: dropping common terms. Manning, Raghavan and Schütze**
Stop lists and why web search engines generally don't use them.
https://nlp.stanford.edu/IR-book/html/htmledition/dropping-common-terms-stop-words-1.html

04 **How AI is powering a more helpful Google. Google, October 2020**
Passage understanding and the 7% of queries figure.
https://blog.google/products/search/search-on/

05 **In-depth guide to how Google Search works. Google Search Central**
Crawling, indexing and serving as separate stages.
https://developers.google.com/search/docs/fundamentals/how-search-works

Explainer by Ameneh Saeednia, prepared with AI assistance. The travel-page example is illustrative. Also available as a blog post on seogirl.ca.