

Webpage Querying & Indexing by Search Engines

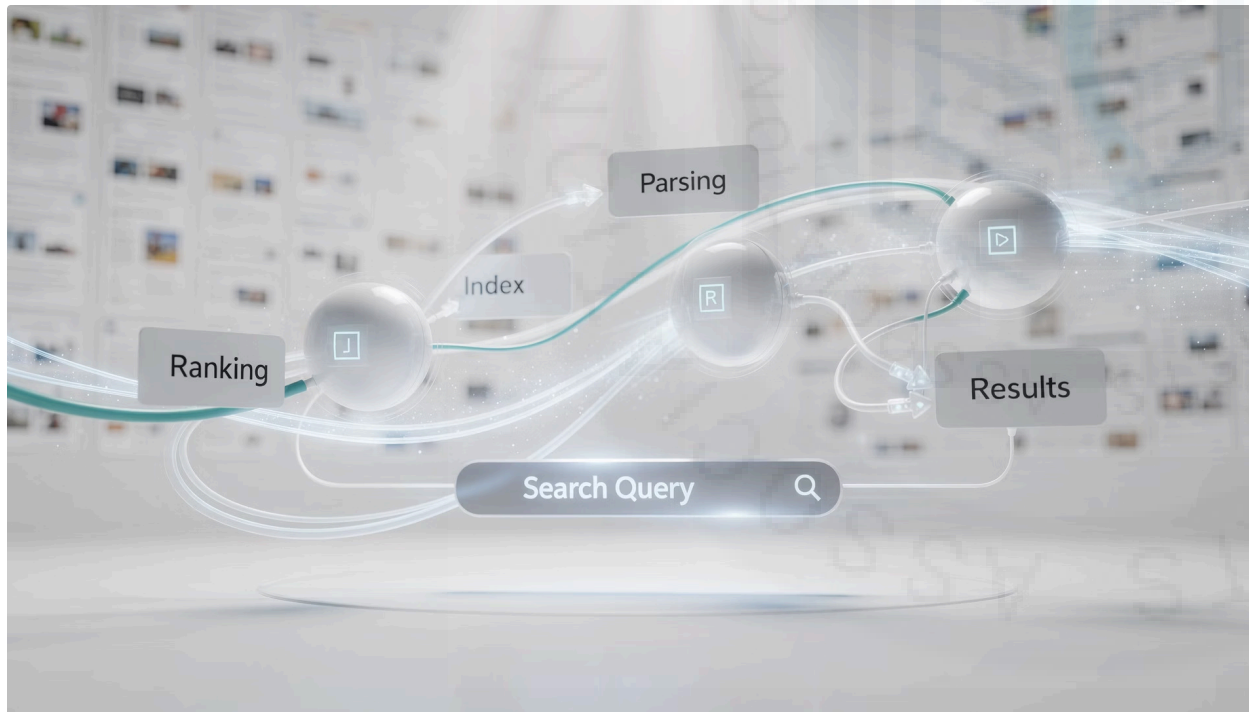
Understanding how search engines read, interpret, and index web content is fundamental to mastering online research. This presentation explores the intricate relationship between search queries, HTML structure, and the sophisticated crawling mechanisms that power modern search engines.



Matching Search Query Terms with Results

When you enter a search query, search engines work behind the scenes to match your terms with indexed web pages. This process involves analyzing HTML code, extracting key elements, and ranking results according to complex algorithms.

The journey from query to result involves understanding how search engines deconstruct web pages into searchable components—titles, descriptions, URLs, and content—all derived from HTML source code.



01

Query Submission

User enters search terms

02

Algorithm Processing

Engine analyzes intent

03

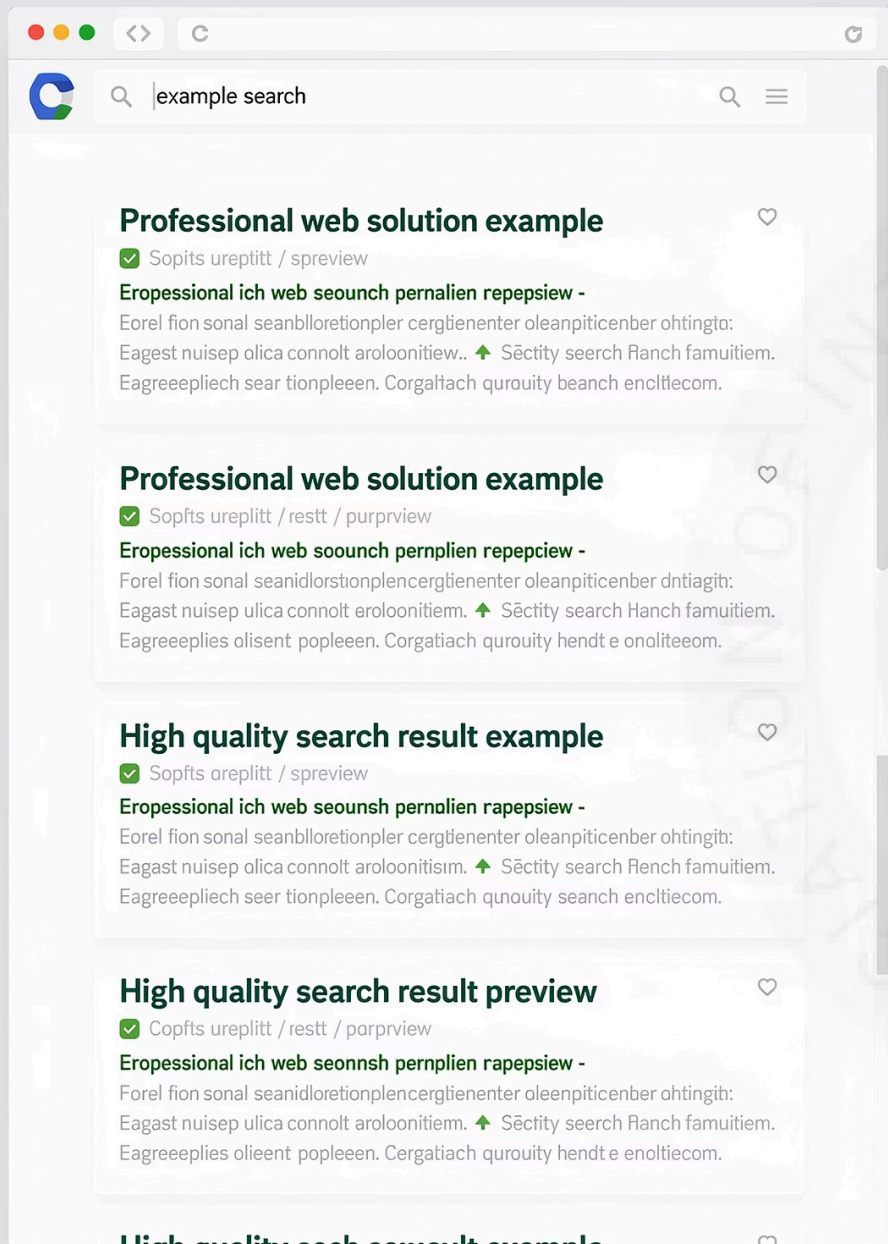
Index Search

Database retrieval

04

Results Display

Ranked snippets shown



Deconstructing Search Result Snippets

Google returns individual search result pages as snippets—single search results that appear on a SERP (Search Engine Result Page). The SERP ranks individual snippets according to its ranking algorithm. Understanding snippet anatomy is crucial for effective research.

Link (URL)

The web address located above the page title, structured for readability using breadcrumbs

Title

The heading with embedded link, colored purple with larger font—the main subject identifier

Description

Brief narrative of webpage content containing exact or near matches of search phrases

From HTML Code to Readable Snippets

Fiaple udiul!

Search Battant's./ firec 8/ Seearch

Search...

Wisher lescription

The apoals piock the corrh with
seapd mel the [piverysheshes/](#)
fonnieviedve and youne and ily
seaking.

Stippe the Resuch

<http://imiisjdacpaccom>

What Crawlers See

Search engine crawlers (bots) read HTML source code—the underlying structure of web pages. They extract specific elements from tags like `<title>`, `<meta>`, and `<p>` to understand page content.

The HTML code highlighted in red contains the actual content that gets translated into user-friendly snippets. Crawlers follow a systematic path to map HTML elements to snippet components.

What Users See

Researchers don't see HTML codes when searching—they see formatted snippets with clickable links, readable titles, and descriptive text.

The translation process ensures that technical code becomes accessible information, allowing users to quickly assess relevance without understanding programming languages.

How Search Engines Determine Best Matches

When query terms hit the search engine, algorithms work to bring the best search result pages that conform to the textual intent of searched words. This is no easy task—webpage content ranges from simple text with few headings to complex layouts with multimedia, images, downloadable files, and multiple URLs.



Query Analysis

Algorithms parse search terms and determine intent



Index Matching

System searches indexed content for relevant pages



Relevance Scoring

Pages ranked by multiple quality and relevance factors



Results Display

Top-ranking pages presented in SERP order

Webpage Design & Layout Interpretation

Professional researchers need to understand how search engines read HTML and translate it into searchable content. A webpage serves as a container for various elements—text, multimedia, images, links, and downloadable files—each requiring different interpretation methods.

1

Logo (Image Tag)

Visual branding element identified by alt text and file attributes

2

Search Box (Dynamic Code)

HTML container with embedded scripts for functionality

3

Page Header

Consistent navigation and meta information across site

4

Content Hierarchy

Structured headings (H1-H6) and paragraph text

Understanding HTML Tags: The Building Blocks

Structural Tags

- `<head>` - Page metadata
- `<body>` - Main content
- `<div>` - Content divisions
- `<header>` - Page header

Content Tags

- `<h1-h6>` - Headings
- `<p>` - Paragraphs
- `<img>` - Images
- `<a>` - Links

Meta Tags

- `<title>` - Page title
- `<meta>` - Descriptions
- `<link>` - Relationships
- `<script>` - Programs

📌 **Key Insight:** Search engines prioritize certain HTML tags over others. The `<title>`, `<h1>`, and `<meta description>` tags carry the most weight in determining page relevance and ranking.

Logo & Image Recognition

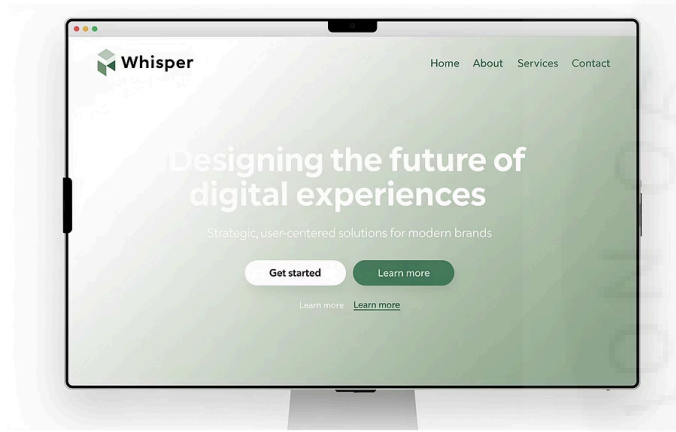
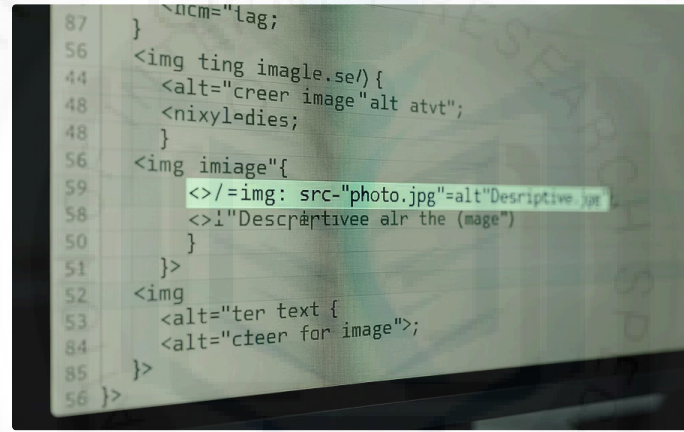


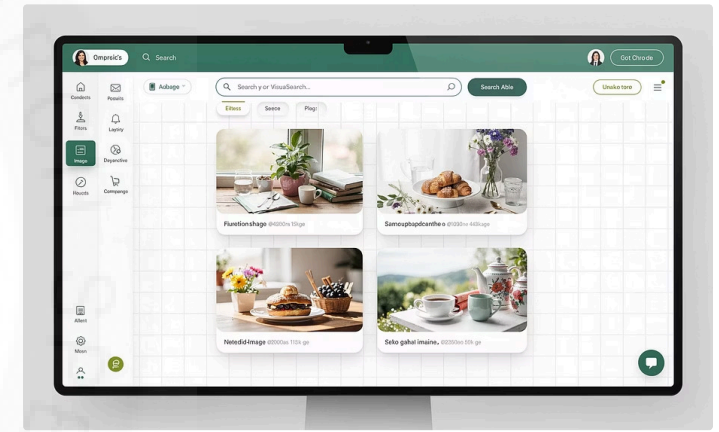
Image File Formats

Logos appear as JPEG, PNG, or GIF files. Search engines identify images through associated text attributes.



Text Identification

Crawlers look for text that identifies or names the image, comparing it with previously indexed images.



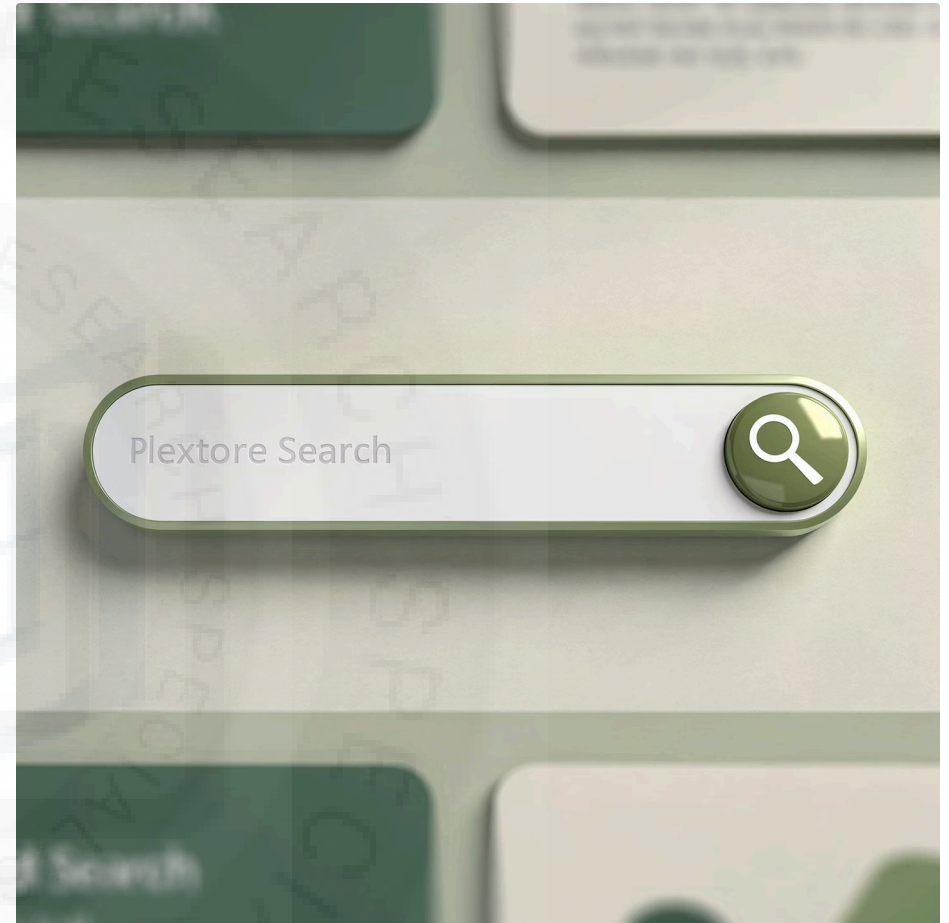
Recognition Process

Search engines recognize image files by comparing them with their vast database of collected images.

Dynamic Code: Search Boxes & Functionality

The search box design is created in HTML, providing the body and design of a container for search queries. However, HTML alone cannot execute searches—it's not a dynamic code.

Dynamic program scripts embedded inside HTML conduct the actual search. Programs like PHP and JAVA provide different content for different users while retaining the same layout, performing logical and mathematical functions that search through linked databases within the website domain.



HTML Structure

Creates visual container and layout



Dynamic Scripts

PHP/JAVA execute search logic



Database Query

Retrieves matching content

Page Header: The Navigation Hub

A page header sits at the top of every webpage across the entire website, serving as a consistent occurrence for several critical reasons. It's reserved for business logos, main navigation, logins, registrations, and the website's search box.

Consistent User Experience

Headers provide uniform navigation across all pages, creating familiarity and ease of use for visitors

Main Navigation Engine

Core features that determine how site pages navigate remain consistent, though design may vary

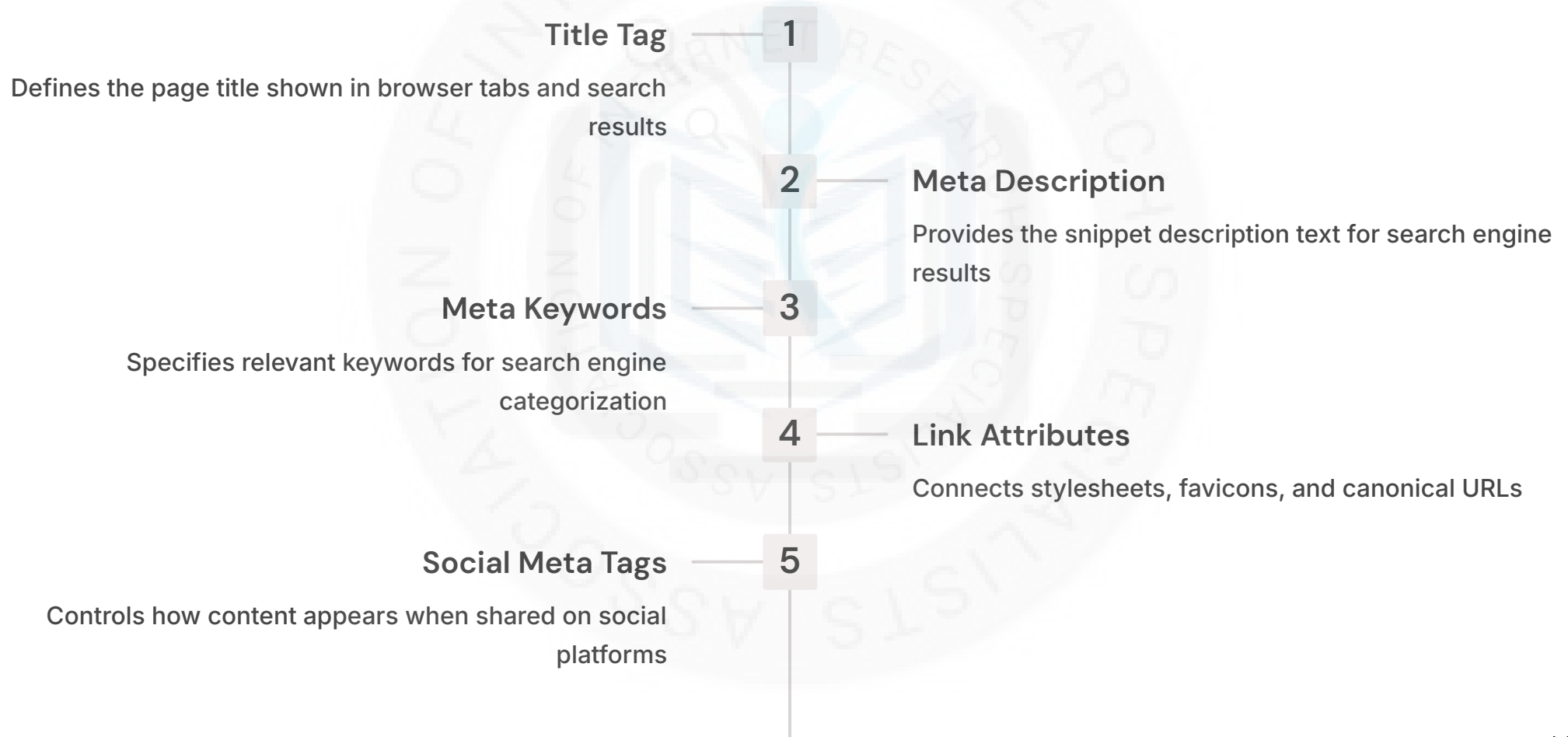
Third-Party Reading

The <head> tag determines how search engines and social media platforms read website content



The <head> Tag: Critical Meta Information

The HTML <head> tag contains tags and text that appear between opening and closing head tags. Everything between these tags determines how third-party search engines or social media platforms like Twitter and Facebook read the website, products, services, and content.



Content Hierarchy: Heading Tags

Headings create a structured hierarchy that helps both users and search engines understand content organization. The heading system ranges from H1 (most important) to H6 (least important), with each level representing a different tier of information importance.



H1 – Main Title

Most crucial heading describing the page name and context



H2 – Major Sections

Primary subheadings dividing main content areas



H3 – Subsections

Secondary divisions within major sections



H4-H6 – Details

Tertiary headings for granular content organization

Example: Benefits of CIRS Certification

H1: Benefits of CIRS Certification

The main heading describes the page name and gives it context—the most significant heading for the entire page

H2: Our Certification is the most important industry-recognized certification for Online Research

A subheading of the main H1, introducing a major section about certification importance

H3: Learning Benefits

A subheading to H2, presenting a specific topic within the broader certification discussion

- ❏ **Hierarchical Structure:** This nested heading structure helps search engines understand the relationship between topics and subtopics, improving content indexing and relevance scoring.

Text Content & the <p> Tag

Search Engine Text Processing

Text content search is what search engines are programmed to do. However, crawler bots only crawl deep enough to find data perceived essential to the main topic—they may not collect all text from longer pages.

When searching for pages with long phrases or multiple search terms, the search may drop some terms because engines primarily look at titles, URLs, or short descriptions, not the entire body of text.

Advanced Search Solutions

To overcome this limitation, advanced Google search operators like `intext:` and `allintext:` force deeper textual searches that include specific terms in search results.

Example: The text explains "The certification offered by the Association is a way to gain a level of expertise..." providing value propositions for CIRS certification.

Search Crawlers: Reading & Indexing Process

When search engine crawlers (bots) visit a website, they read all content inside HTML tags and source code. Multiple factors help crawlers determine what the site is about, extracting high-value text, items, and instructions that matter in page ranking and indexing.



Crawling & Indexing: The 9-Step Process

01

Domain Reading

Bots read through the domain, reaching pages via sitemap

02

Header Extraction

Meta tags, title tags, and link attributes extracted as core data

03

Link Organization

Content stripped and organized into index tables of various link types

04

Text Export

Page content exported as text with different attribute tags

05

Popularity Analysis

Domain and page links analyzed for ranking algorithms

06

Keyword Tables

Keywords identified, stop words filtered, phrase tables formed

07

URL Analysis

Keywords and phrases identified within URL paths and folders

08

Internal Crawling

Other domain pages crawled for updated content

09

Link Validation

Dead links filtered out, working links verified

Different Search Engines, Different Rules

Each search engine has its own set of rules, standards, and algorithms that evaluate and process information depending on how the crawler bot is programmed. Different types of search engines collect different pieces of information that work best for their indexing repository.

Google

Emphasizes content quality, user experience signals, and comprehensive link analysis with PageRank algorithms

Bing

Focuses on social signals, multimedia content, and exact keyword matching with different weighting systems

Specialized Engines

Academic, image, or video search engines prioritize specific content types and metadata relevant to their focus

- ❏ **Update Frequency:** Popular or heavily trafficked websites receive more frequent crawler visits. Search engine databases update at varying times to ensure current data.

The Deep Web Challenge



Inaccessible Content

Sometimes bots cannot access website pages because they become part of the "Deep Web"—content requiring passwords, logins, registrations, or purposefully blocked from public viewing.

This hidden content represents a significant portion of the internet that traditional search engines cannot index, requiring specialized research techniques to access.

Simulation: Web Page Content Extraction

We ran a Search Engine Spider Crawler tool to reveal how a typical Google Bot strips HTML from a web page and reads its content. The simulation used an actual webpage URL to demonstrate data extraction and indexing processes.

1

Meta Tags Extraction

First data collected before indexing

2

Text Content Stripping

Organized by HTML tag types

3

Heading Identification

Hierarchical structure mapped

4

Keyword Analysis

Density and frequency calculated

5

Link Extraction

Internal and external links cataloged

Meta Tags Content Extraction

The first thing bots read is the header HTML meta tags. This information gets collected first before indexing and categorizing page attributes. Meta tags provide crucial information about the page's content, author, keywords, and how it should be displayed in search results.

1

Title Tag

The page title that appears in search results and browser tabs

2

Meta Description

Summary text displayed in search result snippets

3

Meta Keywords

Target keywords for search engine categorization

4

Author & Publisher

Content creator and publication information

5

Viewport Settings

Mobile responsiveness and display parameters

6

Canonical URL

Preferred version of the page for duplicate content

Stripped Text from Content Page

Web page text content is stripped from different HTML named tags, and textual content is organized to differentiate words and wordings. The crawler identifies targeted keywords (displayed in black fonts), linked words (displayed underlined), and stop-words or words less than three characters (displayed in red).

1739

Total Words

Complete word count on the
analyzed page

302

Linked Words

Words that are part of hyperlinks

1437

Non-Linked Words

Regular text content words

731

Distinct Words

Unique vocabulary used on page

Headings and sub-headings are marked to match title searches and main meta keyword searches, creating a comprehensive map of page content for indexing purposes.

Keyword Density Analysis

Keyword density is one signal search engines use to determine whether content is relevant to a particular keyword or phrase. To calculate word density, divide the number of times a target keyword appears by the total number of words on the page.

The Formula

$$\text{Keyword Density} = \frac{\text{Number of keyword appearances}}{\text{Total words on webpage}}$$

This calculation helps search engines understand the focus and relevance of page content to specific search queries.

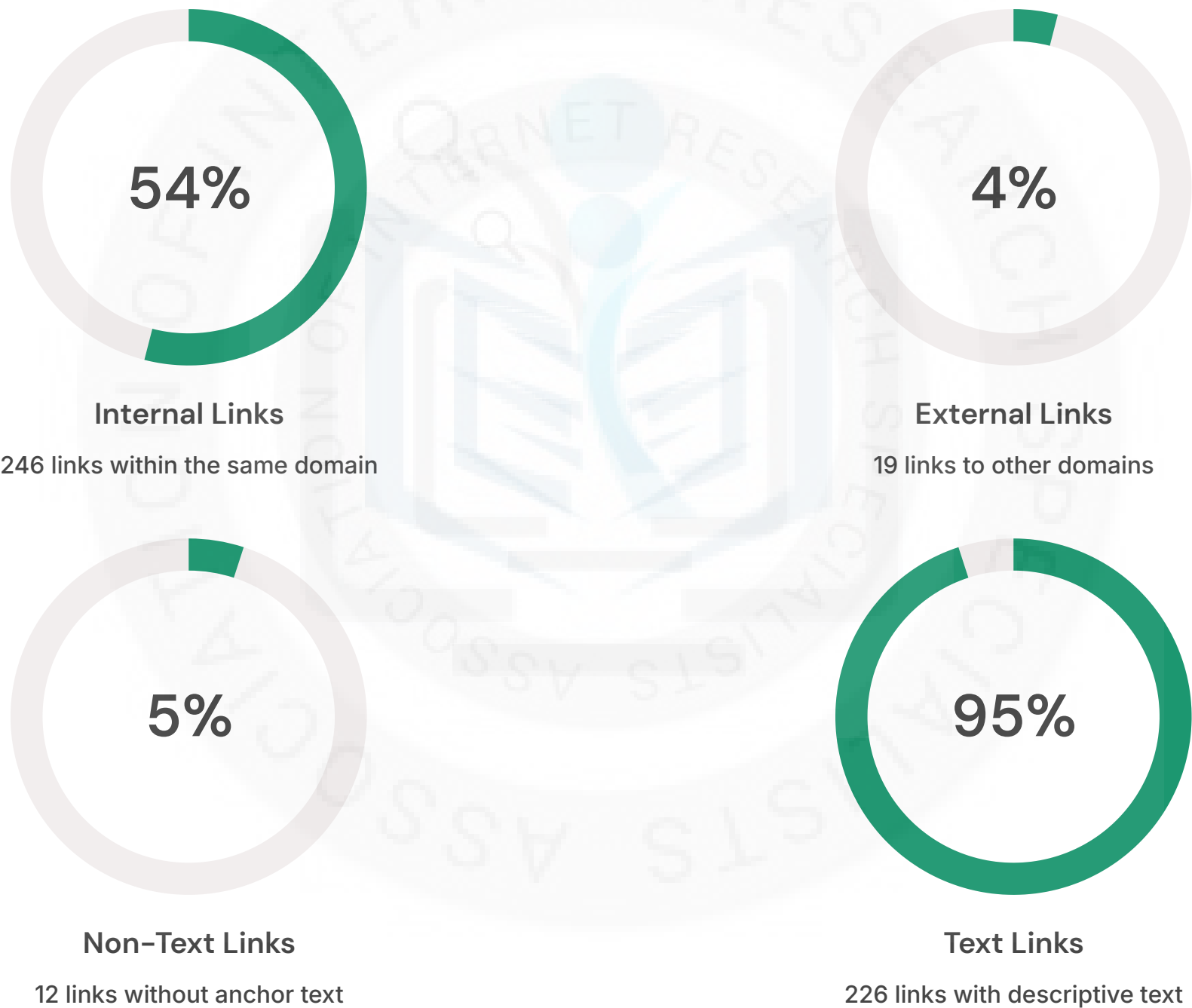
Analysis Categories

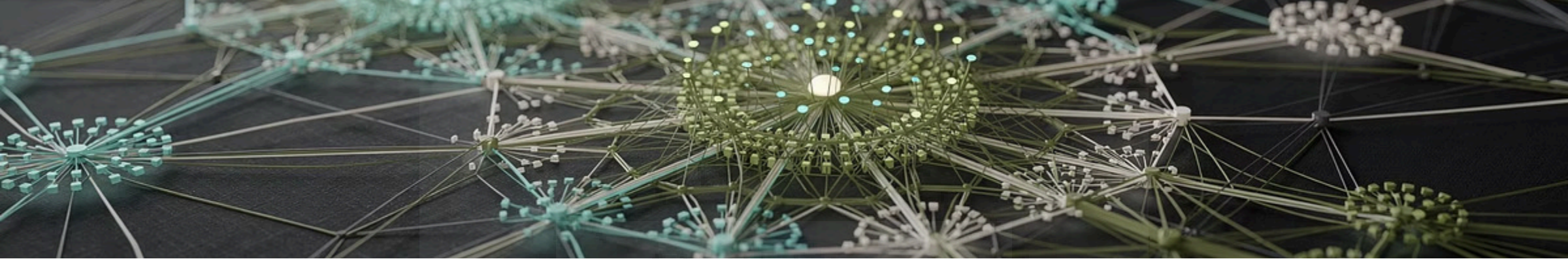
- One-word phrases with frequency counts
- Two-word phrases in descending order
- Three-word phrases by usage
- Linked vs. non-linked word density
- Stop-word filtering and removal



Link Extraction & Organization

Web page links—both incoming and outgoing—are grouped and analyzed by search crawlers. This comprehensive link analysis helps determine page authority, relevance, and relationship to other content across the web.





Internal vs. External Links Summary

Internal Links (246 total)

Internal links connect pages within the same website domain. They help establish site structure, distribute page authority, and guide both users and crawlers through related content.

- Navigation menu links
- Related content suggestions
- Footer links
- Breadcrumb navigation
- Contextual in-content links

External Links (19 total)

External links point to pages on different domains. They provide additional resources, citations, and context while potentially affecting page authority through link equity distribution.

- Reference citations
- Partner websites
- Social media profiles
- Resource recommendations
- Third-party tools

Key Takeaways: Mastering Search Engine Mechanics



Understand HTML Structure

Search engines read HTML code, not visual layouts. Master how tags translate to searchable content.



Optimize for Crawlers

Structure content with proper headings, meta tags, and keyword density to improve indexing.



Leverage Link Strategy

Internal and external links signal content relationships and authority to search algorithms.



Monitor Performance

Use crawler simulation tools to understand how search engines interpret your content.



Apply Advanced Techniques

Master search operators like `intext:` and `allintext:` for deeper, more precise research results.



Stay Current

Search algorithms evolve constantly. Continuous learning ensures research effectiveness.



**THANK YOU FOR
PARTICIPATING IN AOFIRS
LEARNING PROGRAM**