






-  _____

-  _____



 A look at search engines with their own indexes



  Rohan Kumar




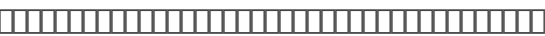
 2025-01-30 (commit hash: 15ddf9743664d7b1b6401ba6d49b1bbace9c06ed)










   Google  Bing  Yandex  GBY  GBY
 GBY 








 “”



 / RDFa
 JSON-LD 



 “”











  



Google



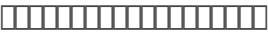
 Websub 













-  Startpage Google  Startpage  Bing[2]
- GMX Search
-  Runnaroo
- Mullvad Leta
- SAPO
- DSearch
- 13TABS
- Zarebin
-   _____ 

Bing

 IndexNow API 

Yandex  Seznam  IndexNow



- Yahoo OneSearch
- DuckDuckGo[3] Tor  JS  TUI  “  ”

- AOL
- Qwant [4]
- Ecosia
- Ekoru
- Privado
- Findx
- Disconnect Search[5]
- PrivacyWall
- Lilo
- SearchScene
- Peekier Peekr 
- Oscobo
- Million Short

- Yippy search[6]
- Lycos
- Givero
- Swisscows
- Fireball
- Netzzappen
- You.com[7]
- Vuhuv
- Metager []
- ChatGPT Search[8]
- [] Bing []

Yandex

[]

IndexNow API [] Bing [] Seznam [] IndexNow []

- Epic Search[] 2021 [] 6 []
- [] DuckDuckGo [] Bing[] [] *DuckDuckgo* [] Yandex “[] ”
- [] [] “[] ” [] []
- Petal[]

Mojeek

[]

GBY[]

Mojeek [] [eTools.ch](#) [] []

Mojeek [] GBY [] []

Google[] Bing [] Yandex [] microformats1[] microdata[] RDFa[] Open Graph markup [] JSON-LD[] Yandex [] microformats1 []

H-Card

[] Open Graph [] Schema.org [] Mojeek
[] Open Graph [] Schema.org []

[]

[] “[] ”
[]
[]

Stract

[] Stract [] “optics”[] Brave [] “
[] ”[] Stract [] [] AGPL-3.0
[] Stract [] Common Crawl

Secret Search Engine Labs

SEO
“”
”
CashRank

Gabanza

Jambo

2006

search.dxhub.de

Gigablast
Gigablast

Fynd

URL
/

Yessle

Bloopish

YaCy

/

Scopia

10 MetaGer Bing 2024 9

Artado Search

Category	Count
Plumb	10
JS	5
Google	5
Bing	5
Yahoo	5
Petal	5
MetaGer	5
"twitter"	5
"wikipedia"	5
"reddit"	5

Active Search Results

Crawlson

Diagram illustrating memory layout with two rows of blocks:

- Row 1: A block of 10 blocks, followed by a block labeled "URL", and the text "seirdy.one".
- Row 2: A block labeled "Crawson", followed by a block of 10 blocks.

Anoos

Yioop!

Dataset	Number of Records
WARC	200,000
Meorca	100,000
Yioop	50,000
Usenet	10,000
API	5,000

Spyda

James Mills So I'm a Knucklehead eh? Go
MIT Spyda

Slzii.com

seirdy.one

--	--	--	--	--	--	--

Qwant

Qwant

Bing

Bing

Neeva

Kagi Search

Neeva

TeclisTeclis

GoogleBing

Teclis

Kagi

Kagi.ai

TinyGem

Kagi.com

BraveAPI

PriEco

Google“”

GBY

Marginallia Search

/

GBY

SEO

SERPs

Teclis

Kagi2022-

05-28

Marginallia.nu

Ichido

Marginalia

CAPTCHA

SEO

Ichido

Teclis

```

graph TD
    Kagi[Kagi search] --> uBlock[uBlock Origin]
    Kagi --> Readability[Readability.js]
    Kagi --> Trafilatura[Trafilatura]
    uBlock --> Readability
    Readability --> Trafilatura
    Trafilatura --> HTML[HTML]
    Trafilatura --> Web[Web]
    Trafilatura --> Marginalia[Marginalia]
    Trafilatura --> API[API]
    Trafilatura --> RDFa[RDFa]
    HTML --> Web
    Web --> Marginalia
    Web --> API
    Web --> Kagi
    Marginalia --> API
    API --> Kagi
    RDFa --> Kagi
  
```

Clew


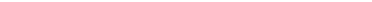
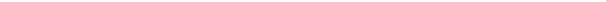



FOSS
seirdy.one

Lixia Labs Search

JavaScript

--	--	--	--	--

Kozmonavt

 800 
 Kozmonavt

 /  seirdy.one 

search.tl

The diagram shows a URL structure with the following components highlighted in boxes:

- TLD** (Top Level Domain)
- .com** (Domain extension)
- TLD** (Top Level Domain)
- [13]** (Index number)
- UI** (User Interface)
- TLD** (Top Level Domain)
- /** (Slash separator)
- tld** (Top Level Domain)
- URL** (Uniform Resource Locator)
- .org** (Domain extension)
- &tld=org** (Query parameter)
- URL** (Uniform Resource Locator)
- Amidalla** (Domain name)
- Amidalla** (Domain name)
- URL** (Uniform Resource Locator)
- search.tl** (Domain name)

Thunderstone

Diagram illustrating the structure of a URL:

- Protocol:** http
- Subdomain:** www
- Domain Name:** billybob
- Top-Level Domain:** com
- Second-Level Domain:** net
- Third-Level Domain:** org
- Path Segments:** web, yahoo, dogpile, aol

sengine.info

netEstate GmbH

Gnomit

2009
"IRC" IRC

--	--

High Browse

 SEO
 " " "

 "

Keybot

[illegible]

Quor

□□□□□□□□□□
2021 □ 6 □□□□□□□□□□
www dot quor dot com □□□

Semantic Scholar

Allen Institute for AI PDF

Bonzamate








 Boyter
 Bonzamate

searchcode

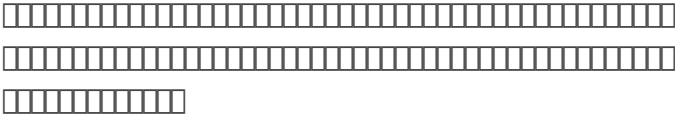
[illegible]

- Baidu GBY
- 360 Qihoo 360
- Toutiao
- Sogou
- Yisou
- Naver Searx
- Daum
- Seznam seirdy.one
- IndexNow Bing Yandex
- Cốc Cốc
- go.mail.ru
- LetSearch.ru URL

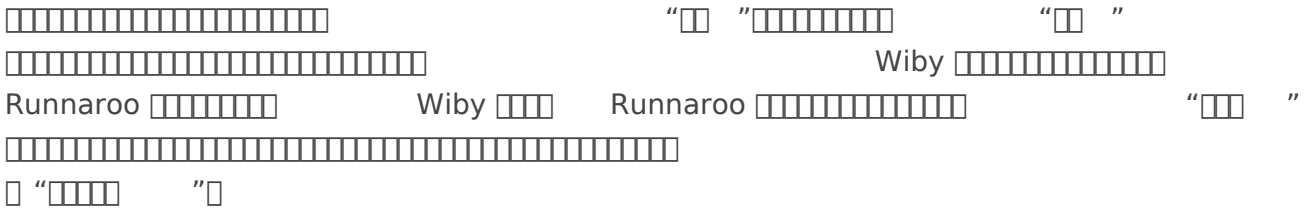
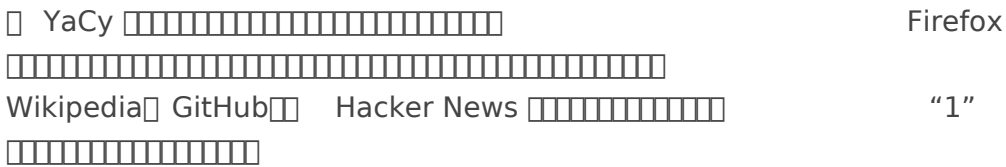


- ALibw.com 
- Vuhuv   
- search.ch 
- fastbot 
- SOLOFIELD 

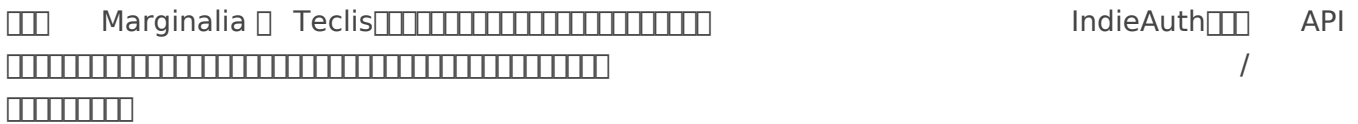
- kaz.kz [] [] [] [] [] [] [] [] [] [] " [] [] [] [] "



wiby.me □ wiby.org

Mwmbi

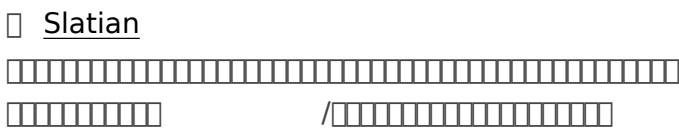
Search My Site



Kukei.eu



Unobtainium Search





Ask.com

Google Bing Yandex
"ask.com " directhit.com

Infinity Search

Infinity De

URL

Infinity Decentralized



-
-



Petal Search

Android

JavaScript

Yandex Qwant 2023 6

Neeva

“” Bing Bing Neeva
Bing Bing
OAuth
Snowflake 2023 5

Gigablast

Web
Private.sh Gigablast Right Dao 2023

wbsrch

Gowiki

seirdy.one
2022

Meorca

“” “”
seirdy.one

Ninfex

“” “”

Marlo

Marlo Haskell
marlo.sandymaguire.me

websearchengine.org tuxdex.com

inetdex.com
1000 Cookies

Entfer

[redacted]
[redacted]

Siik

[redacted] ToS [redacted] PHP
[redacted]
[redacted]

Blog Surf

[redacted] RSS/Atom
[redacted]
“MarketRank” [redacted] “Hacker News” [redacted]

[redacted]

- Parsijoo [redacted]
- Moose.at [redacted] Brave [redacted]



[redacted]

[redacted]



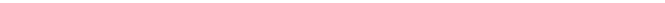

[redacted] (POC)
[redacted]



[redacted]



Google [redacted] Bing [redacted] Yandex [redacted] Google [redacted]
YouTube [redacted] LinkedIn
[redacted]
[redacted]
[redacted]


 Kagi  Ask.com



ToS

--	--	--	--	--	--

matza gebrent



matzo brei

Mojeek

20


30-40

--	--








--	--	--	--

Diagram illustrating memory allocation for three different programs (Google, Chromium, Firefox) across three rows of memory blocks.

- Row 1: 20 blocks. Labeled "Google" (15 blocks) and "Chromium" (5 blocks).
- Row 2: 15 blocks. Labeled "Google" (15 blocks).
- Row 3: 20 blocks. Labeled "Firefox" (20 blocks).

Diagram illustrating the distribution of bot traffic across different bots. The bots are listed on the right: Cloudflare, Googlebot, BingBot, and TwitterBot. The bars represent the relative volume of traffic from each bot, with segments indicating different categories of traffic (e.g., green, yellow, red, blue).

Google JavaScript " " GBY JavaScript

Google

Diagram illustrating the structure of a 100-bit message:

- Header: 10 bits
- Body: 70 bits
 - SEO: 10 bits
 - Data: 60 bits
- Footer: 20 bits



Gigablast Matt Wells GBY " Gigablast

_

☐ The New Leaf Journal ☐ Nicholas Ferrell ☐ “2021

Seznam☐ Naver☐ Baidu☐ Goo☐☐



1.

“indexes” “index” “indices”

2.

2023 3 8 Startpage Startpage Microsoft Bing

3.

DuckDuckGo DuckDuckBot
400
DuckDuckGo Bing
398
DuckDuckGo Bing Yandex
2022 3 DuckDuckGo

4.

Qwant Bing “”

13.

--	--	--	--	--	--	--	--

site:

TLD [][][][][][][][][][][]

site:.one

--	--	--	--	--	--	--

“.one” TLD ☐ ☐ ☐ ☐

14.

[illegible]

Google ☐ Microsoft☐☐☐

YouTube  LinkedIn



[TheCoffeMaker/shm: SelfHoster Manifesto, coz Arrr! Rulez - Codeberg.org](#)

"

"

"

19

"

[illegible][illegible]

15

14 