

Web Scale Search from the Ground Up

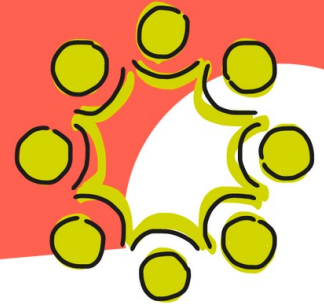
Establishing a Viable Alternative to Big Tech

[Colin Hayhurst](#), CEO

mojeek

4-6 October 2023

5th International
Open Search
Symposium
#ossym23



mojeek

[Web](#)

[Images](#)

[News](#)

[Substack^{demo}](#)

[Emotions^{demo}](#)

[Maps](#)

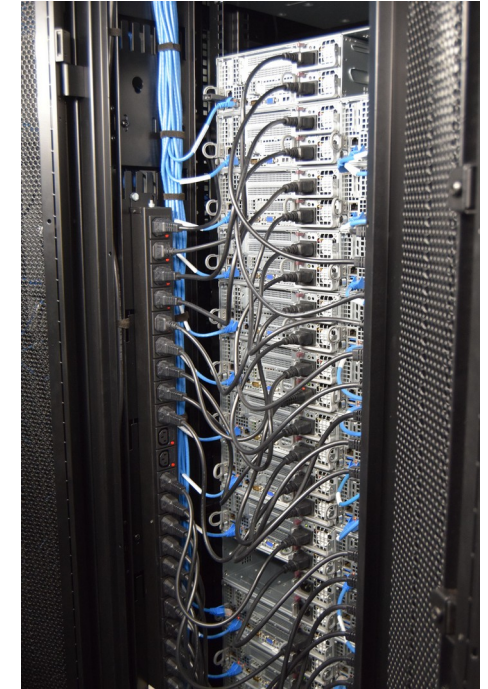


No Tracking. Just Search...



Mojeek Timeline

- 2004: Launched
- 2006: Privacy Declaration
- 2009: Incorporated
- 2011: Data centre
- 2015: 1 billion
- 2018: 2 billion
- 2020: 3 billion, Go to market
- 2022: 6 billion, Choices, Focus
- 2023: 7.4 billion, Maps



More details: <https://blog.mojeek.com/2021/03/to-track-or-not-to-track.html>

What Motivates Us?

Privacy not

Surveillance

Diversity not

Monopoly

Autonomy not

Control

Freedom to Seek not

Censorship



Privacy not Surveillance

Privacy-washing is not for us:

- No analytics services
- No captcha services
- Zero data passed on
- Do not sell or distribute any data
- Discard your IP address & log country code
- Browser usage data in a separate log
- Anonymous usage
- Truly cookie-less option (no local cookie)

<https://blog.mojeek.com/2021/12/personal-data-industry-the-new-tobacco.html>

Diversity not Monopoly

Social media monopolies have consequences














Search engine monopolies have more profound consequences few realise

Mojeek chose to take the hard road of total independence as our attempt to contribute to diversity

We advocate for, and deliver diversity with Search Choices; this allows one click referral to other search options

Search Selections

Select the Search Choices to show. [Learn More.](#)

	Bing	https://www.bing.com/search?q={searchTerms}	<input type="checkbox"/>
	Brave	https://search.brave.com/search?q={searchTerms}	<input type="checkbox"/>
	DuckDuckGo	https://duckduckgo.com/?q={searchTerms}	<input type="checkbox"/>
	Ecosia	https://www.ecosia.org/search?q={searchTerms}	<input checked="" type="checkbox"/>
	Google	https://www.google.com/search?q={searchTerms}	<input type="checkbox"/>
	Lilo	https://search.lilo.org/?q={searchTerms}	<input checked="" type="checkbox"/>
	Metager	https://metager.org/meta/meta.ger3?eingabe={searchTerms}	<input checked="" type="checkbox"/>
	Qwant	https://www.qwant.com/?q={searchTerms}	<input checked="" type="checkbox"/>
	Startpage	https://www.startpage.com/sp/search?q={searchTerms}	<input type="checkbox"/>
	Swisscows	https://swisscows.com/en/web?query={searchTerms}	<input checked="" type="checkbox"/>
	Yandex	https://yandex.com/search/?text={searchTerms}	<input type="checkbox"/>
	Yep	https://yep.com/web?q={searchTerms}	<input type="checkbox"/>
	You	https://you.com/search?q={searchTerms}	<input type="checkbox"/>

<https://blog.mojeek.com/2022/02/search-choices-enable-freedom-to-seek.html>

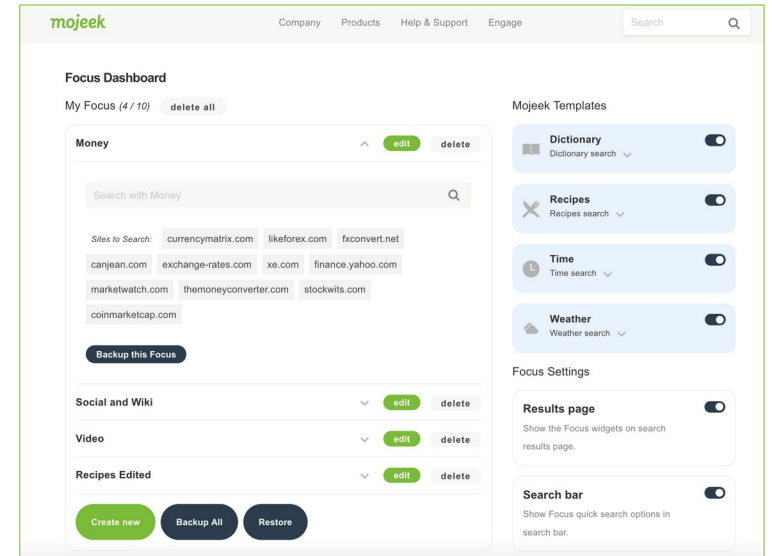
Autonomy not Control

We seek to provide more autonomy to users, for example:

- 1,000 results always, if available
- Specify snippet length (0 to 511 chars)
- Turn off infobox

Mojeek Focus enables easy ways to search part of the indexed web

<https://www.mojeek.com/focus/>



<https://blog.mojeek.com/2022/08/mojeek-focus-search-the-web-you-want.html>

Freedom to Seek not Censorship

“If Twitter has a freedom of speech problem,
then Google has a freedom to seek problem”

“What doesn’t show, you may may never know”



See our blog this topic
and relevant human rights declarations
(ICCPR Article 19, UDHR Articles 29(2) & 30),
that guide us

<https://blog.mojeek.com/2022/05/freedom-to-see-matters.html>

Technical Challenges

Data Centre
Servers
Crawling
Indexing
Ranking
Defence



Data Centre

Independent company
Dedicated, secure rooms
Environmental impact



mojeek.com



Mojeek Servers

Supermicro twin node bare bone servers

each node typically fitted with:

- Dual Intel Xeon Silver CPUs
- 128GB Micron ECC RAM
- 1TB Hard Disk Drives
- 1TB Solid State Drives



Video of some assembly and racking:

<https://youtu.be/rGIDyegXu8E>

Crawling

Much less of a computational challenge than indexing
Robots.txt

- MojeekBot has crawled respectfully since 2004
- Some sites only allow large search engines, or often just Google

Web Application Frameworks (WAF)

- Verified bot on Cloudflare and others

Ranking

Matching:

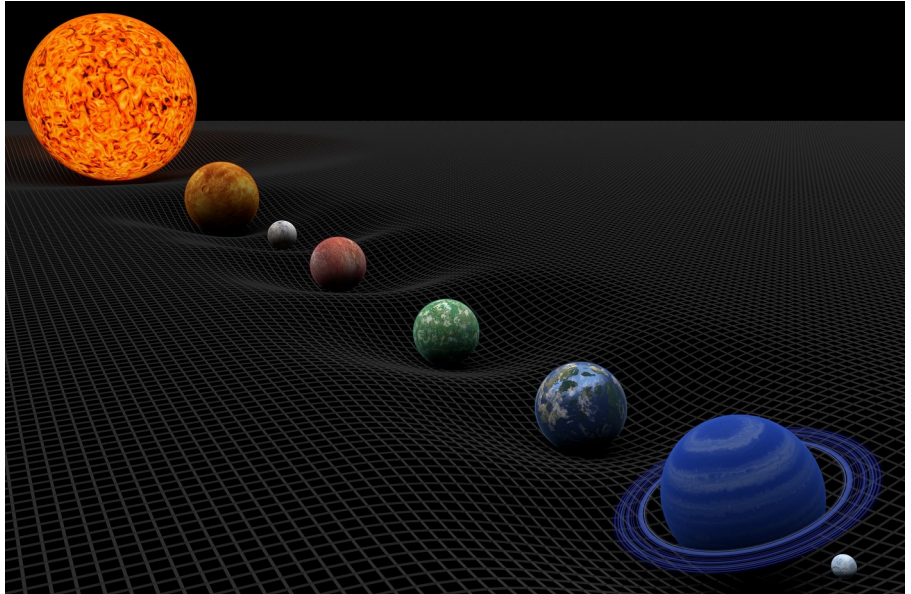
Keyword

Semantic

Numerous factors

Authority:

”Gravity” which is similar to PageRank



Retrieval from Index

Nodes ($n=s$ sets of k) rank the top 1,000 candidate results, for the search query

Slave (s) server receive these 1,000 results from their nodes (k), and then rank these into a top 1,000

Master server receives the top 1,000 from (s) slaves, and rank a final top 1,000 results, which are then available for usage.

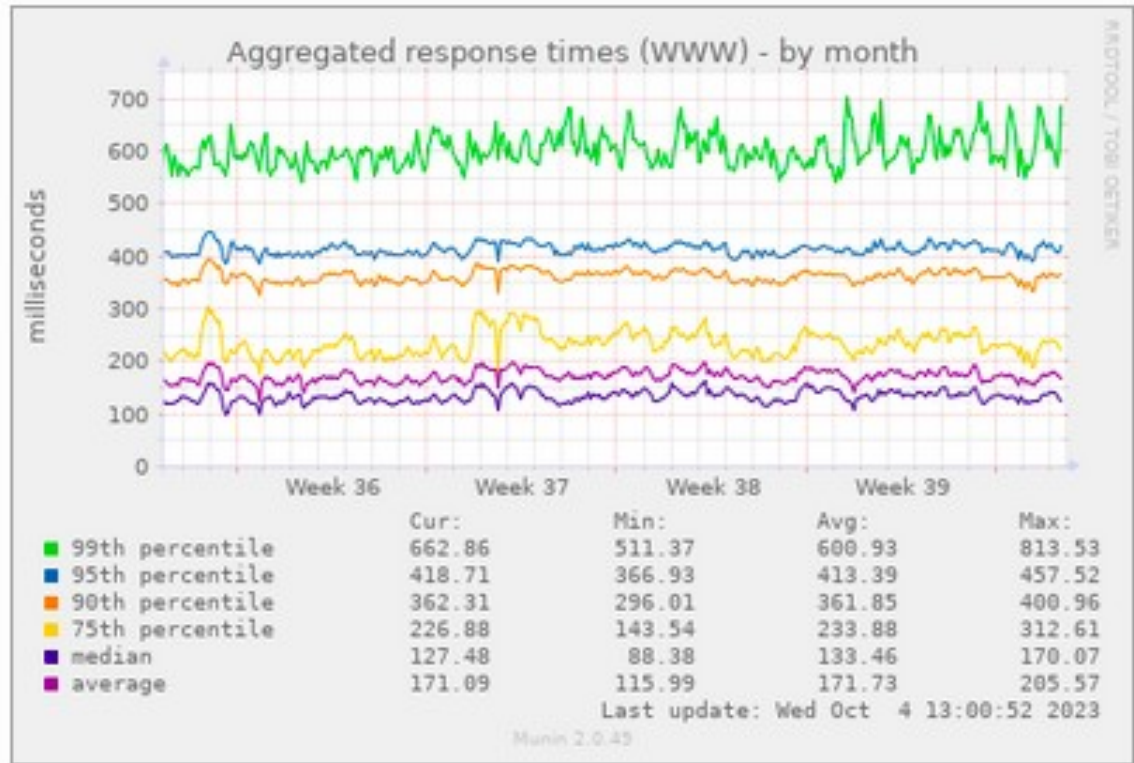
Scalable Service

Proxy servers prepare the client display, or API data:

- Request data for the results required (10 by default and up to 100) from text **database** server.
- The text database server requests the data from the relevant **nodes**.
- Snippets are computed from the full text for the pages, for the results required.
- Search results are then provided as **Title**, **URL** and **snippet**

Proxy servers can be scaled up as is needed to handle the volume of overlapping (in-time) searches.

Response Times



Median response time over the last month 133 ms

Bots

Legitimate human queries amount to less than 10%



















Non-human, non-API, queries

- General scrapers
- Scrapers targetting search services
 - to scrape results and/or
 - find content to scrape
- Botnets & viruses on personal computers
- Hackers looking for vulnerable sites to target

Commercial Challenges

Another longer off-the-record talk!

One huge example:

		Google	Microsoft	Apple					
									
		Browser							
		Chrome	Edge	Safari					
									
Search Engine	Search Service		Google	Google 		Default	Option		Default
	Bing	Bing 	Option	Default	Option				
	uses Bing	DuckDuckGo 	Option	Option	Option				
	uses Bing	Yahoo 	Option	Option	Option				
	uses Bing	Ecosia 	Option	-	Option				

<https://blog.mojeek.com/2022/05/gatekeepers-of-the-western-web.html>

Questions

No Tracking. Just Search...



<https://www.mojeek.com/>

colin@mojeek.com

[@colinhayhurst](#)