Architecture of Information Systems

Search Engine

John Samuel
CPE Lyon

Search

Year: 2017-2018
Email: john(dot)samuel(at)cpe(dot)fr

Outline: Search engine

- Frontend development
- Backend development
- Application programming interface

Search Engine

Search

Search Engine

Search

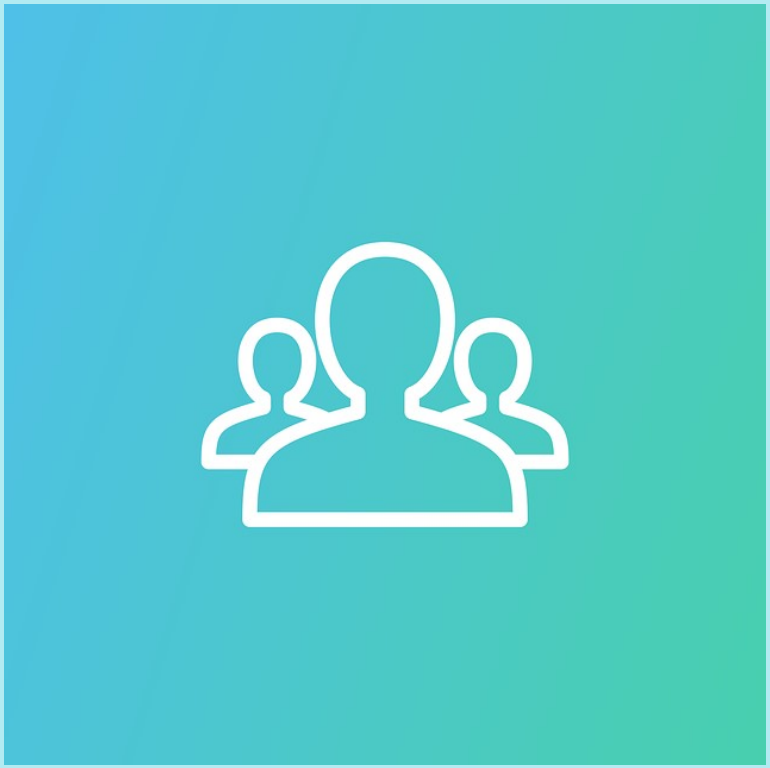
Time

Location

...

Target Audience

- Regular users
- Domain Experts



Front-end development

Search interface

- One-box search
- Advanced search (filters)

Personalized user experience

Interface: Simple search (One box)

Queries

- Keywords
- Natural language queries

Search Results

- Links
- Structured response
- Natural language response

Leonardo da Vinci (October 2017 Google results)

Advanced search (filters)

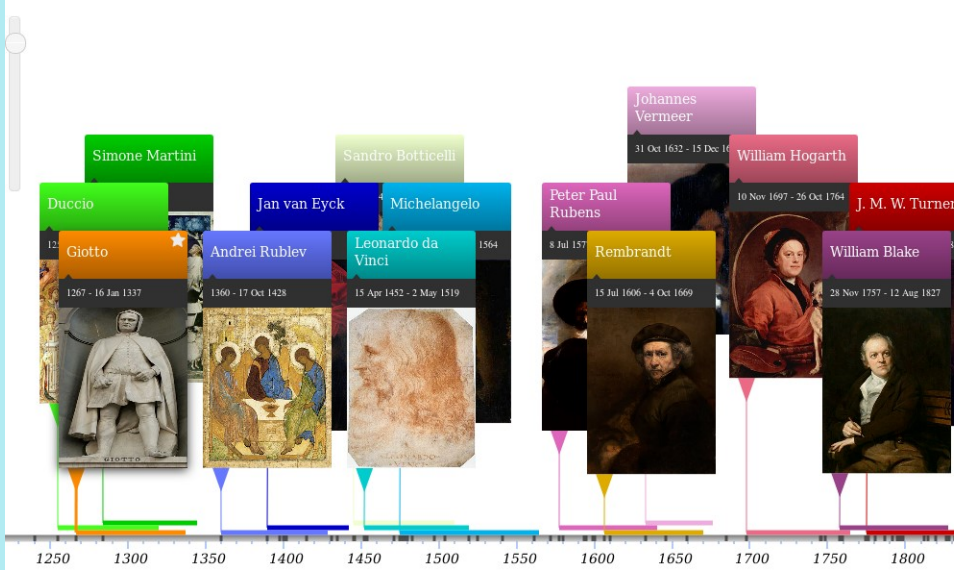
Filter search results (Multiple boxes)

Search

Time

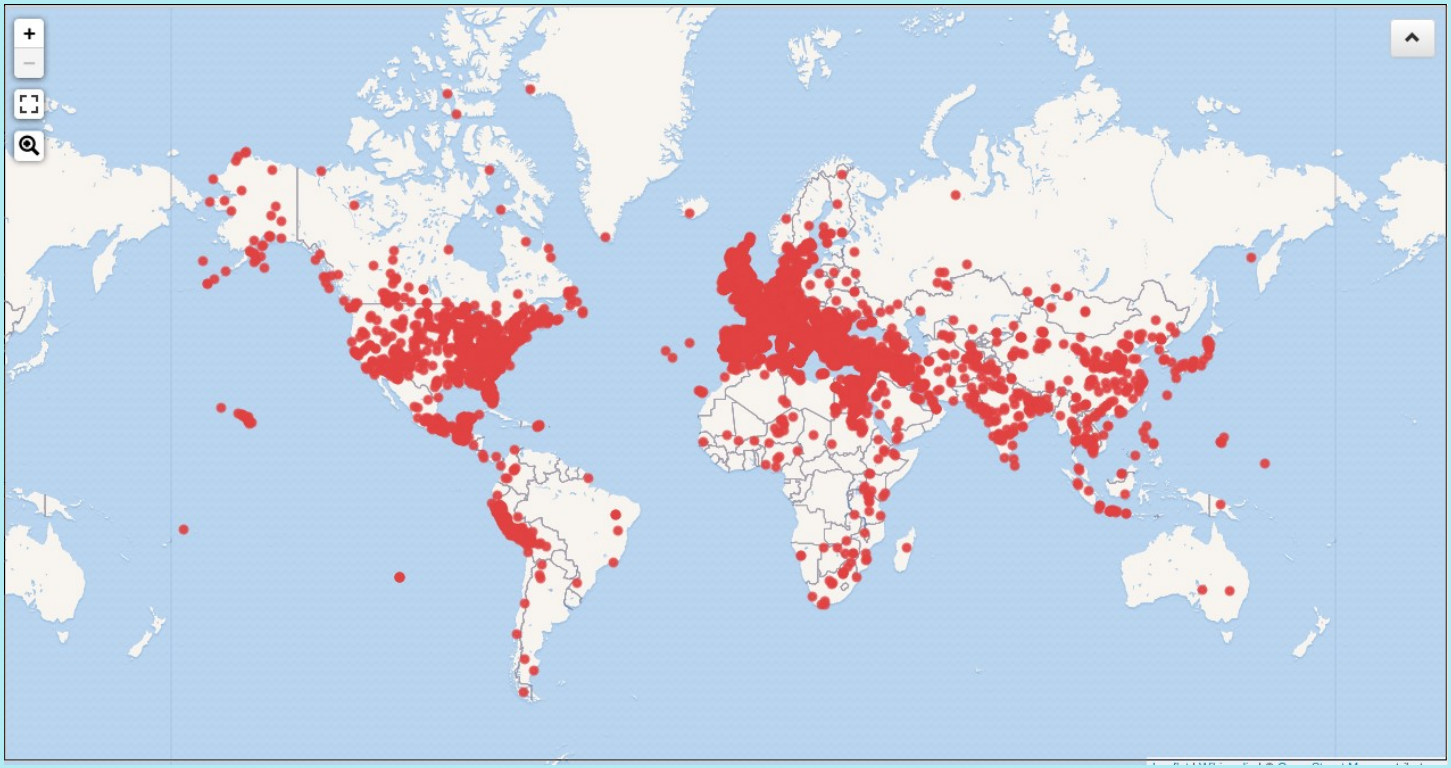
Location

...



Artists on Histropedia

Advanced search (filters)



Location of Archaelogical sites (Wikidata)

Advanced search (filters)

Why filters?

- Reduce information overload
- Precise queries
- Interactive search

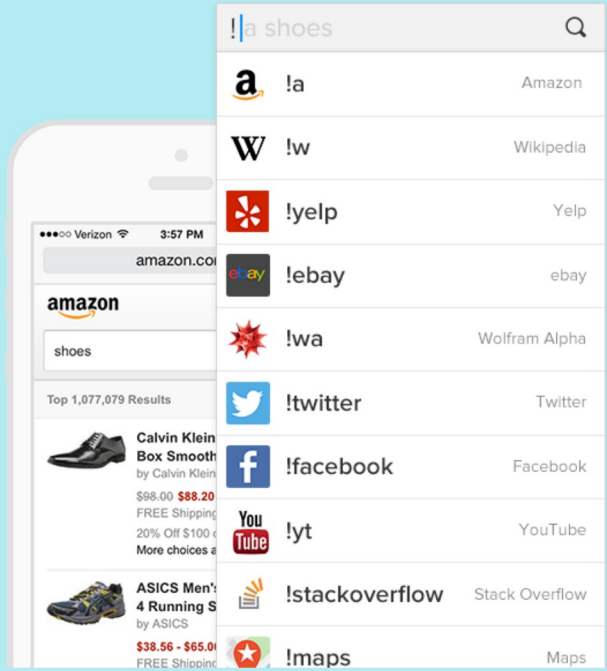
Advanced Search in one-box

Operators

- AND
- OR
- NOT

Advanced Search in one-box

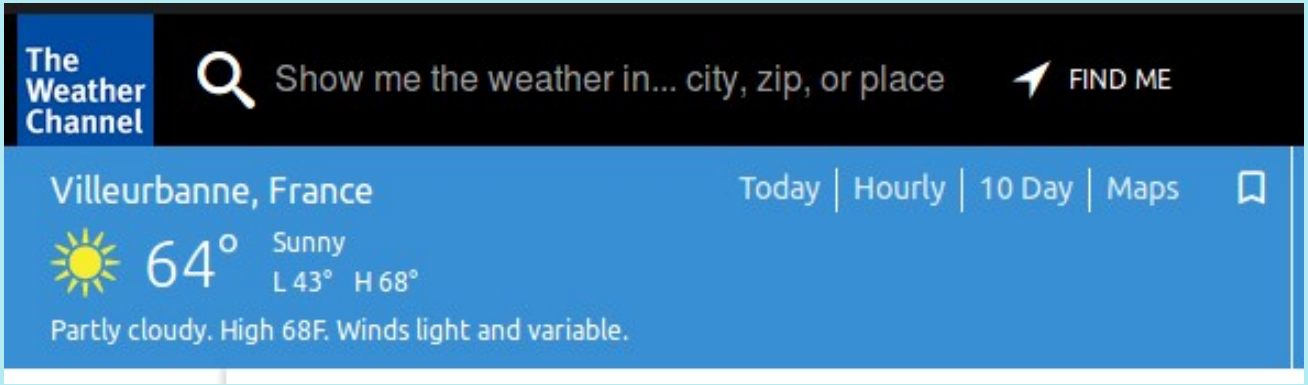
Mnemonics



Bangs (DuckDuckGo)

Personalized user experience

Time and location (Internationalization)



Weather (weather.com)

Personalized user experience

Past user search queries

User privacy

Backend development

- Data collection
- Data storage
- Configuration
- Logging
- Dashboard
- Security

Data collection

Data ownership

- Internal data (e.g., official website, internal wikis, databases etc.)
- External data (e.g., other websites, wikis, open data)

Data model (Data and Schema)

- Unstructured data (e.g., documents, texts, web pages etc.)
- Semi-structured data (e.g., JSON/XML files etc.)
- Structured data (e.g., relational databases, linked data)

Data collection

Data sources

- Web pages
- Documents, texts
- Sensors
- Databases
- ...

Data collection

Data acquisition

- Data dumps
- Crawlers
- Web scraping
- Application Programming Interface (API)

Data collection

Data cleaning and transformation

- Accuracy (e.g., verification with external sources)
- Validity (e.g., detect constraint violations)
- Uniformity (e.g., units)

Data storage

- Model
- Indexation
- Query optimization
- Caching
- Replication
- Backup

Data Model

- Database schema
- Schema-less

Data storage

- Relational Databases
- Object-oriented Databases
- NoSQL databases (e.g., graph databases)
- NewSQL databases (SQL + ACID guarantees)

Document indices and Query Optimization

Document indices

- Forward index
- Inverted index

Database Indexation

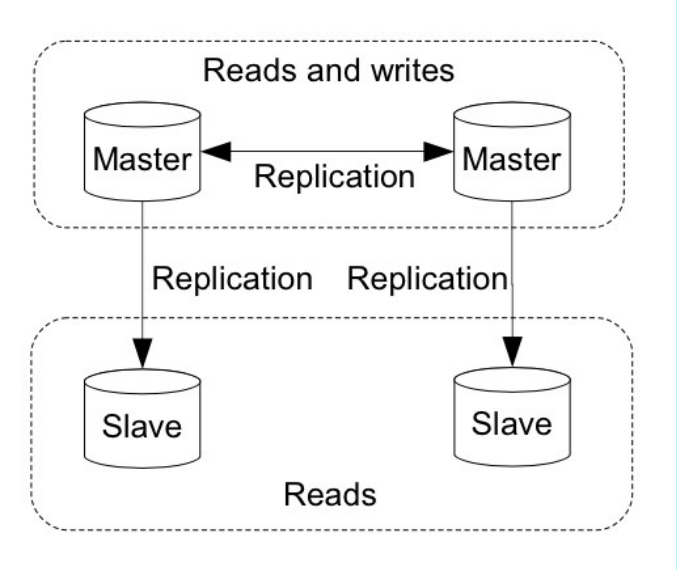
Query Optimization

- Join ordering
- Cost estimation

Caching

- Frequently asked questions and cached responses

Replication and Backup



Replication(Master-slave)

Resource management and configuration

Availability %	Downtime per year	Downtime per month	Downtime per week	Downtime per day
90% ("one nine")	**36.5 days**	**72 hours**	**16.8 hours**	**2.4 hours**
95% ("one and a half nines")	18.25 days	36 hours	8.4 hours	1.2 hours
97%	10.96 days	21.6 hours	5.04 hours	43.2 minutes
98%	7.30 days	14.4 hours	3.36 hours	28.8 minutes
99% ("two nines")	**3.65 days**	**7.20 hours**	**1.68 hours**	**14.4 minutes**
99.5% ("two and a half nines")	1.83 days	3.60 hours	50.4 minutes	7.2 minutes
99.8%	17.52 hours	86.23 minutes	20.16 minutes	2.88 minutes
99.9% ("three nines")	**8.76 hours**	**43.8 minutes**	**10.1 minutes**	**1.44 minutes**
99.95% ("three and a half nines")	4.38 hours	21.56 minutes	5.04 minutes	43.2 seconds
99.99% ("four nines")	**52.56 minutes**	**4.38 minutes**	**1.01 minutes**	**8.64 seconds**
99.995% ("four and a half nines")	26.28 minutes	2.16 minutes	30.24 seconds	4.32 seconds
99.999% ("five nines")	**5.26 minutes**	**25.9 seconds**	**6.05 seconds**	**864.3 milliseconds**
99.9999% ("six nines")	**31.5 seconds**	**2.59 seconds**	**604.8 milliseconds**	**86.4 milliseconds**
99.99999% ("seven nines")	**3.15 seconds**	**262.97 milliseconds**	**60.48 milliseconds**	**8.64 milliseconds**
99.999999% ("eight nines")	**315.569 milliseconds**	**26.297 milliseconds**	**6.048 milliseconds**	**0.864 milliseconds**
99.9999999% ("nine nines")	**31.5569 milliseconds**	**2.6297 milliseconds**	**0.6048 milliseconds**	**0.0864 milliseconds**

Availability (Wikipedia)

Resource management and configuration

- Machines (servers, disks etc.)
- Software packages and dependencies
- Energy consumption

Deployment

- Development setup
- Pre-production setup
- Production setup

Packaging

- Containers (e.g., Linux containers)

Load balancing

- Server-side
- Client-side

Selective Testing

A/B Testing

Logging

- Access logs
- Error logs
- Event logs
- Transaction logs

Logging

Why logs?

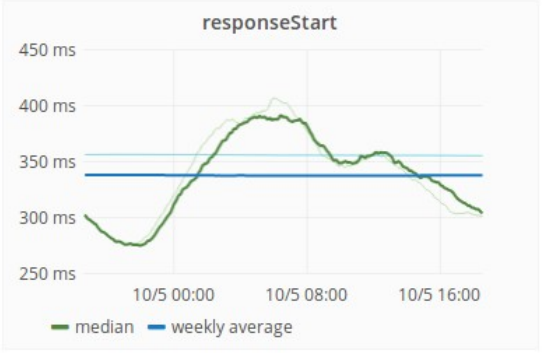
- Debugging
- Security (e.g., detect intrusion)
- Database rollbacks
- Audit
- Analysis (e.g., detecting patterns, resource planning)

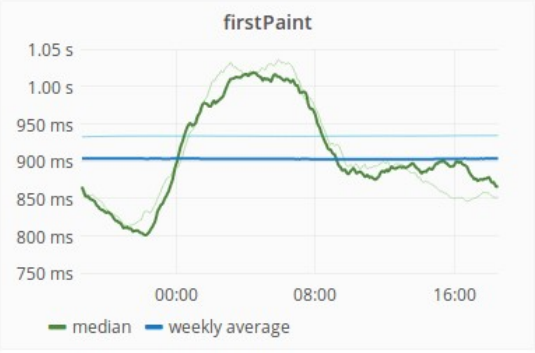
Logging

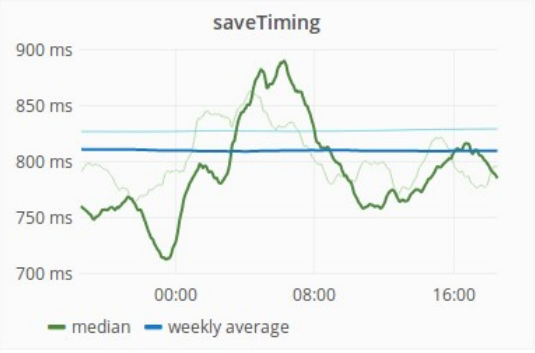
- IP address
- User
- Resource ID
- ...

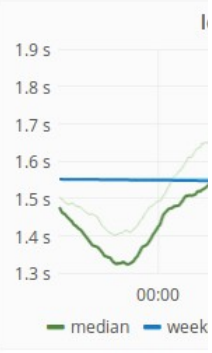
Dashboard

Current		Relative to last week		Relative to las
Metric	**Current**	**Metric**	**Current**	**Metric**
Median Time To First Byte	288.8 ms	Median Time To First Byte	4 ms	Median Time To First Byte
Median First Paint	825.5 ms	Median First Paint	25 ms	Median First Paint
Median Save Timing	744.5 ms	Median Save Timing	-26 ms	Median Save Timing
Median Page Load End	1.4 s	Median Page Load End	13 ms	Median Page Load End



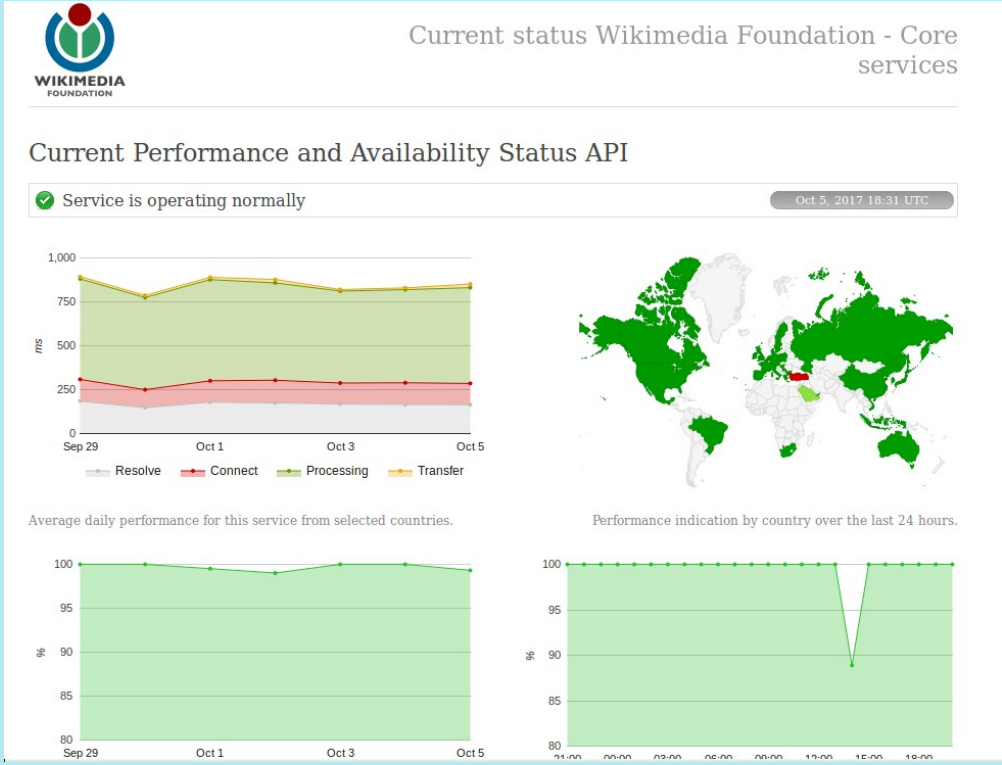






Wikimedia (Grafana: 5th October 2017)

Dashboard



Wikimedia (Availability: 5th October 2017)

Dashboard

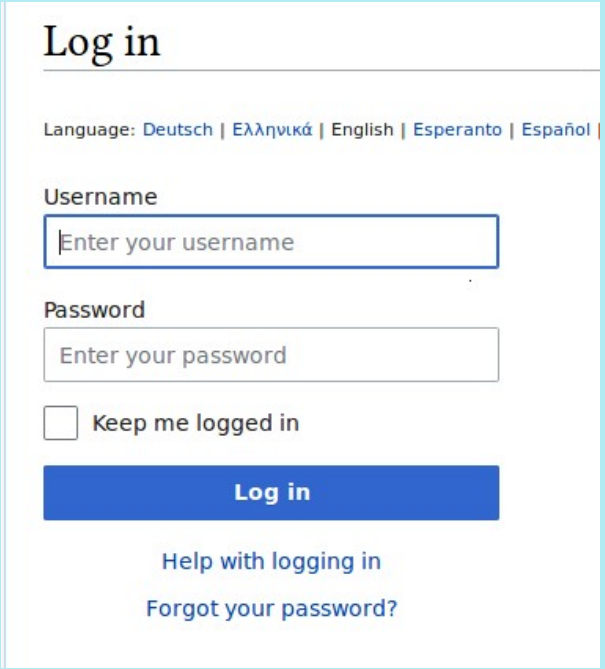
- Performance Metrics (active users, queries served etc.)
- Real-time metrics (e.g., downtime, latency, throughput)

Dashboard

- Email Alerts
- Visual indicators

Security

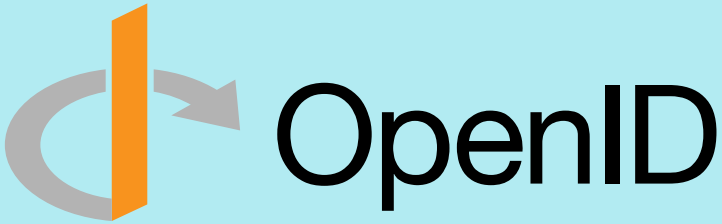
- Data protection
- Logged-in users or public access
- Third party access



Login (Wikipedia)

Security

- Authentication
- Authorization to third party access



OpenID

Mozilla Persona

Developer(s)	Mozilla Foundation
Initial release	July 2011
Written in	JavaScript
Operating system	Cross-platform
Available in	51 languages
Type	Authorization
License	MPL
Website	developer.mozilla.org/en-US/Persona

Mozilla Persona (2011-2016)

Detecting security vulnerabilities

- Intrusion
- SQL code injection
- Cross-site scripting
- Denial of service

Application programming interface

- Service-oriented (SOAP)
- Resource-oriented (REST)

API: Data formats

- XML
- JSON

API: (CRUDL) Operations

- Create
- Read
- Update
- Delete
- List

API: Data dumps

- Complete data dumps
- Selective data dumps

Application programming interface

- HTTP
- Software development kits (SDKs)

Interface definition

- Human-readable documentation
- Machine-readable documentation (WSDL, WADL etc.)
- Human and machine-readable documentation (microformats, semantic web languages)

Human readable Documentation

1. Read documentation
2. Develop application to integrate
3. Add business logic, if any

Machine-readable Documentation

1. Fully autonomous solution to integrate
2. Add business logic, if any

Quality of service

Resource usage limits

- Limits on API call count (per user, IP)
- Limits on data transfer
- Temporary blocks

Quality of service

- Analysis on frequently made API calls
- Resource planning and allocation

Security

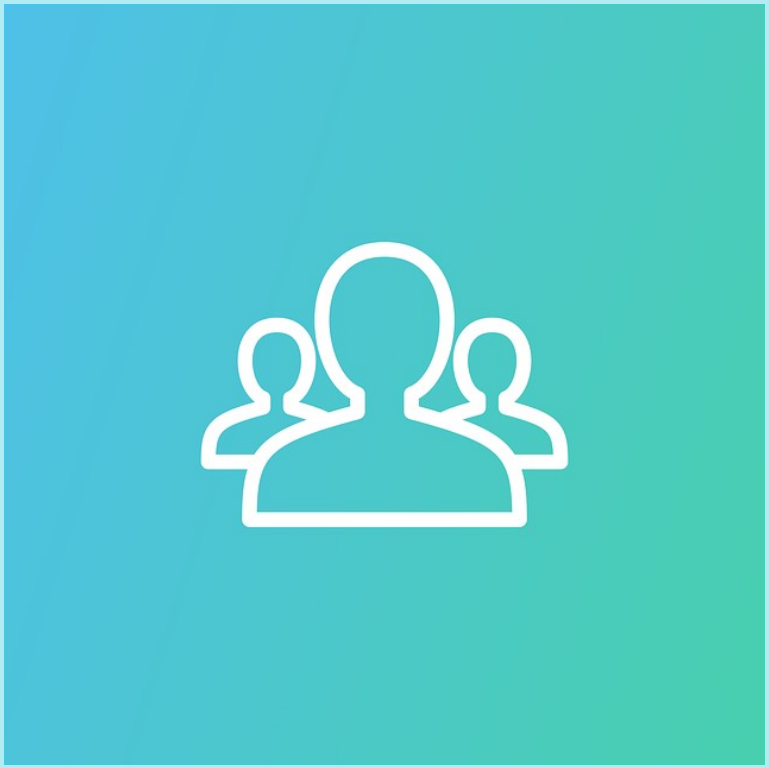
- No password
- Basic authentication (e.g., username, password)
- (Open) authentication protocols (e.g., OAuth)

OAuth

Project

Virtual Library

Project



Target audience

Project

Search

Search

Time

Location

...

References

References

- https://en.wikipedia.org/wiki/Information_system
- https://en.wikipedia.org/wiki/Search_engine_(computing)
- https://query.wikidata.org/
- https://weather.com/
- https://duckduckgo.com/bang
- http://commoncrawl.org/
- https://en.wikipedia.org/wiki/High_availability
- https://grafana.wikimedia.org
- https://status.wikimedia.org
- http://highscalability.com/

Image credits

- Wikimedia Commons
- http://histropedia.com/timeline/5b
- https://pixabay.com/