



# OpenWebSearch.EU

## Towards an Open Web Search Infrastructure



<https://openwebsearch.eu/>

Prof. Dr. Michael Granitzer

# Partners: 12+2, 8.5 MEur Funding



Research



ICT Solutions for Brilliant Minds

Infrastructure



NGOs



Businesses

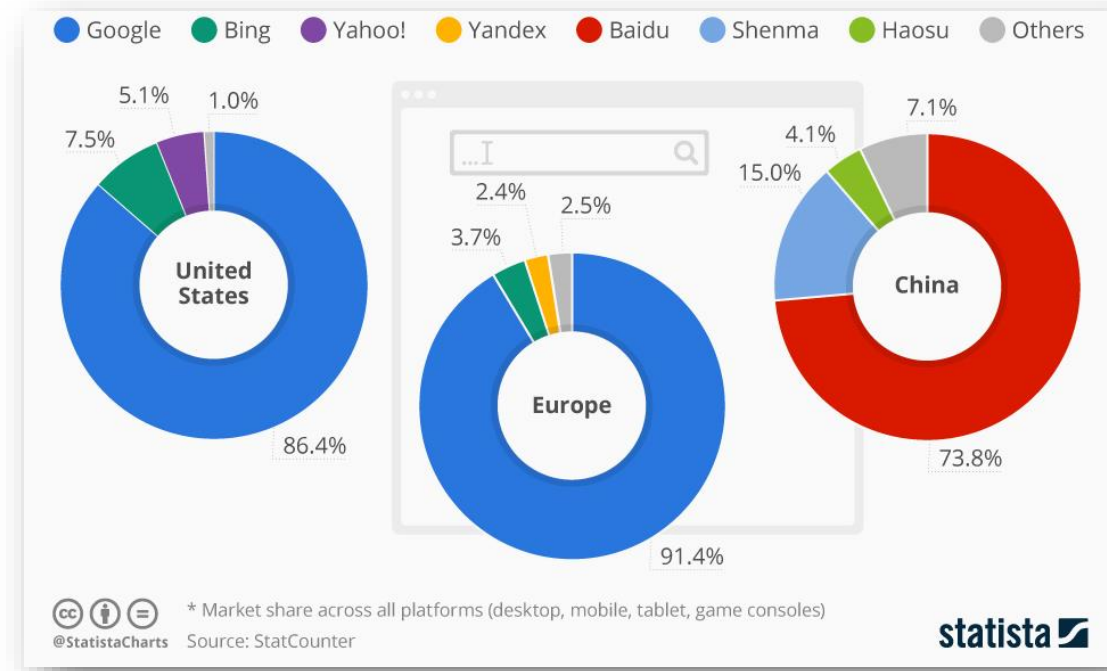
# Motivation

## Two properties of Web Search that don't fit

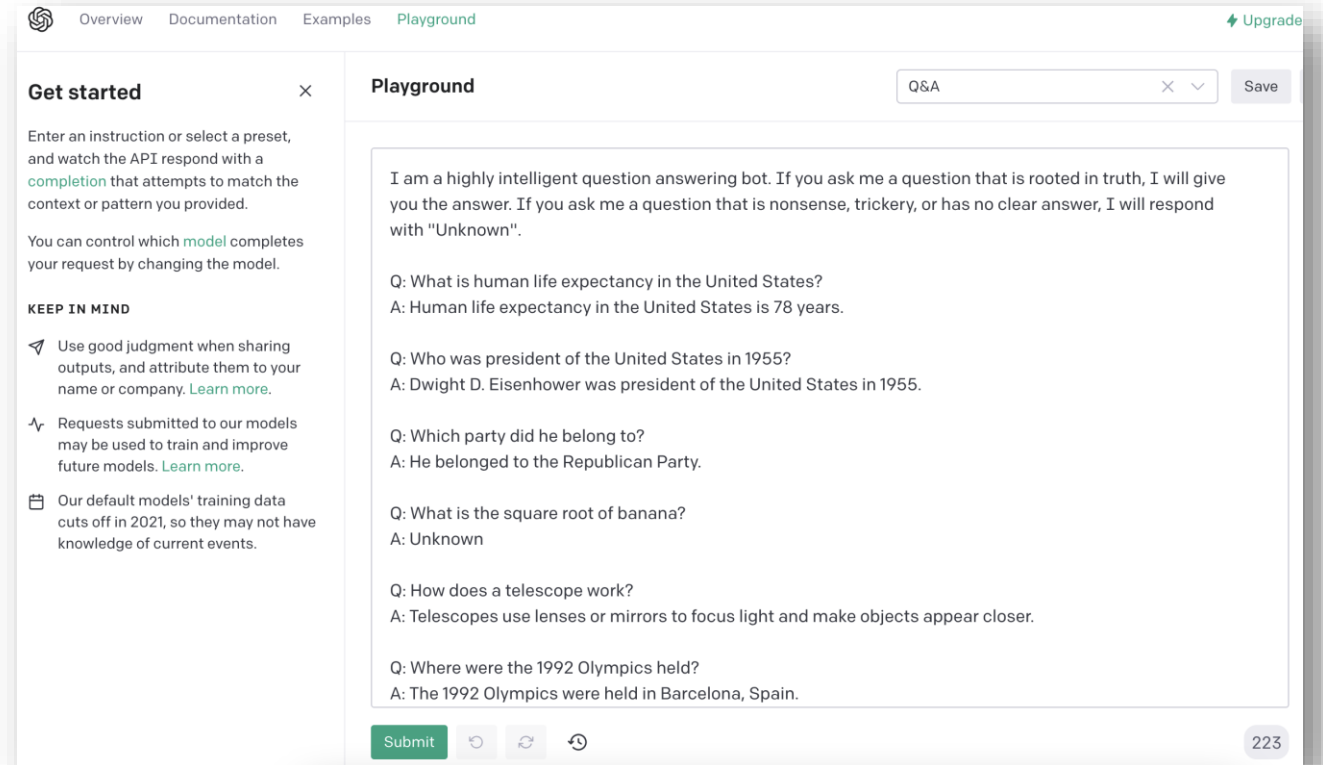
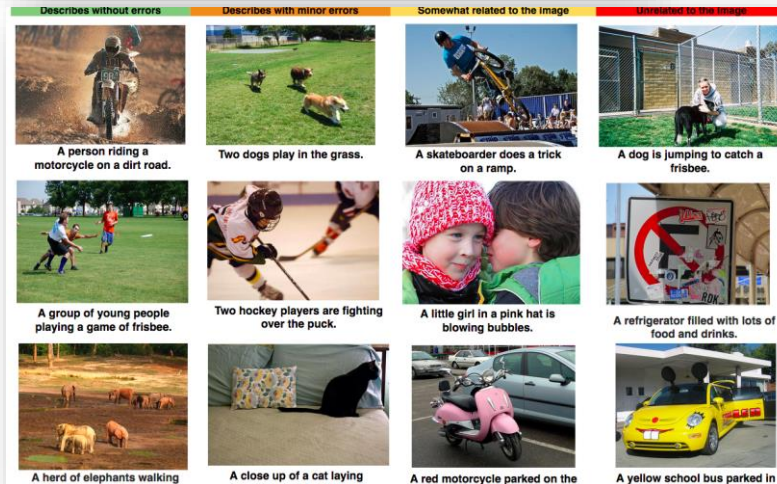
- A critical infrastructure for society, comparable to satellite navigation
- A market oligopoly: i.e. “a market structure in which a market or industry is dominated by a small number of large sellers or producers.” (Wikipedia)

## Effects

- Reduced User Choice
- User locked-in despite of “Open” technologies
- Rich-gets-richer effects through exclusive data
- Concerning market behaviour (e.g. Jedi Blue)
- SEO optimized ranking vs. best information delivery?
- Limited business models
- ....



## Web data drives innovation beyond search



Overview Documentation Examples Playground Upgrade

### Get started

Enter an instruction or select a preset, and watch the API respond with a completion that attempts to match the context or pattern you provided.

You can control which model completes your request by changing the model.

### KEEP IN MIND

- Use good judgment when sharing outputs, and attribute them to your name or company. [Learn more.](#)
- Requests submitted to our models may be used to train and improve future models. [Learn more.](#)
- Our default models' training data cuts off in 2021, so they may not have knowledge of current events.

### Playground

Q&A Save

I am a highly intelligent question answering bot. If you ask me a question that is rooted in truth, I will give you the answer. If you ask me a question that is nonsense, trickery, or has no clear answer, I will respond with "Unknown".

Q: What is human life expectancy in the United States?  
A: Human life expectancy in the United States is 78 years.

Q: Who was president of the United States in 1955?  
A: Dwight D. Eisenhower was president of the United States in 1955.

Q: Which party did he belong to?  
A: He belonged to the Republican Party.

Q: What is the square root of banana?  
A: Unknown

Q: How does a telescope work?  
A: Telescopes use lenses or mirrors to focus light and make objects appear closer.

Q: Where were the 1992 Olympics held?  
A: The 1992 Olympics were held in Barcelona, Spain.

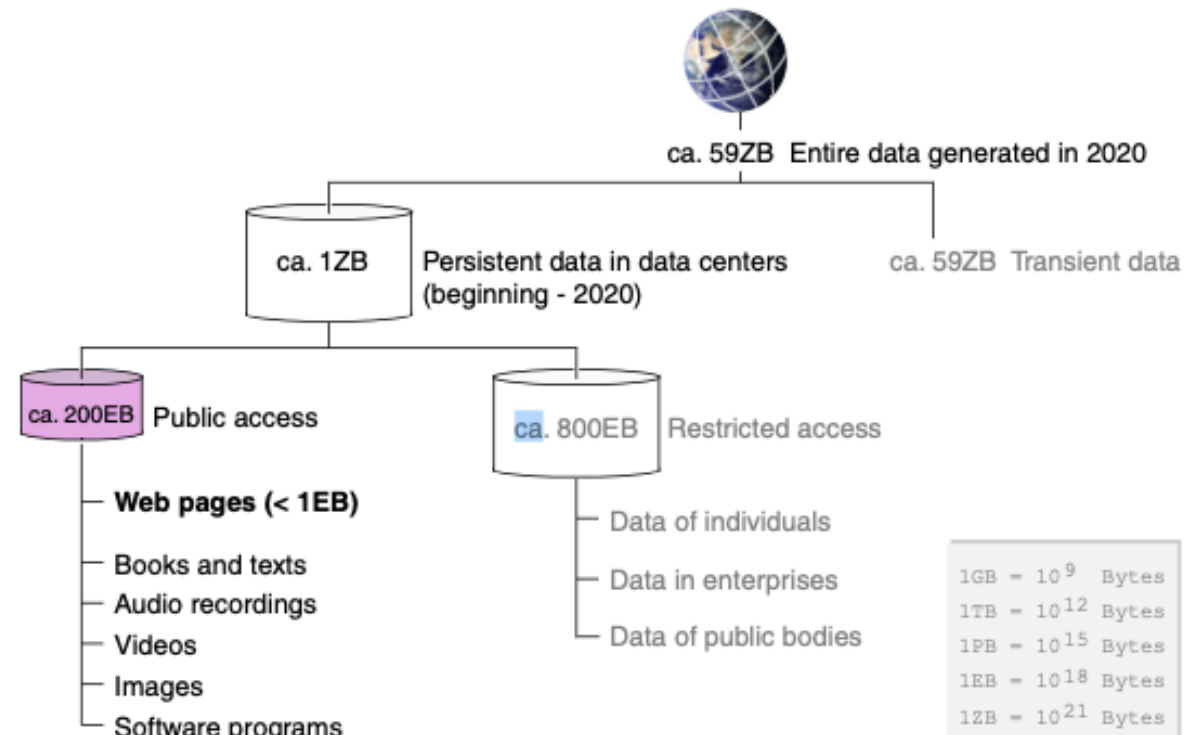
Submit ↻ ↺ ↻

223

OpenAI trained on open crawls like Common Crawls

Working with web data can be challenging and costly: its big & unstructured

- High-demands on hardware resources
- High level of technological skill
  - Infrastructure
  - Big Data computing
  - Data cleaning
  - Natural Language Processing & Computer Vision
- Need only for particular subsets of the data
- Legal and ethical constraints (e.g. GDPR)
- Competitive, partially adversarial environment (e.g. Spam, Link Farms, Security)



Völske, M., Bevendorff, J., Kiesel, J., Stein, B., Fröbe, M., Hagen, M., & Potthast, M. (2021). Web Archive Analytics. INFORMATIK 2020.

## 1. The Web Index

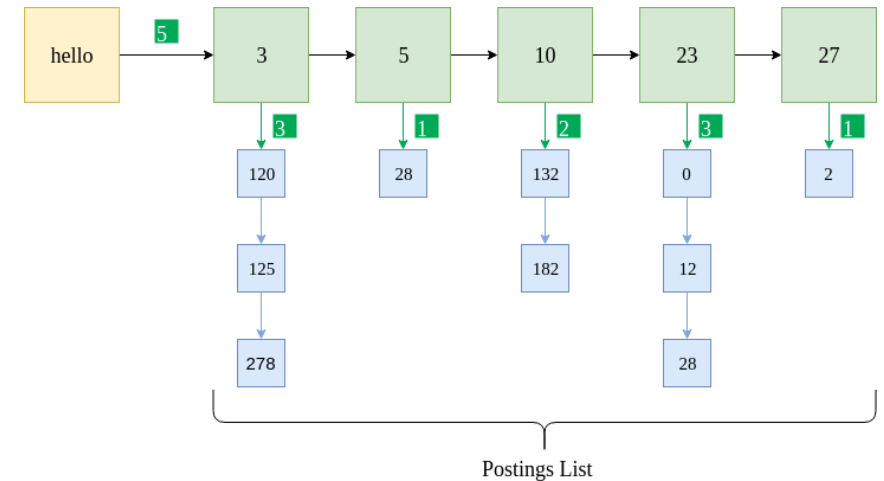
- Data structure for fast access to web documents / sites
- Supports search and ranking criterions

## 2. Preprocessing Pipeline: Hypermedia → Search API

- Crawling the Web and its formats
- Cleaning Web Data
  - Preprocessing HTML at scale
  - Metadata Extraction and Management (e.g. Microformats)
  - Headless Browser support (e.g. SPAs)
  - Dealing with additional formats (e.g. PDF, Doc, PNG...)
- Semantic Enrichment / Extraction
  - Geo-tagging
  - Information Extraction & Linking
  - Knowledge Graphs
- Indexing

## 3. Search UIs

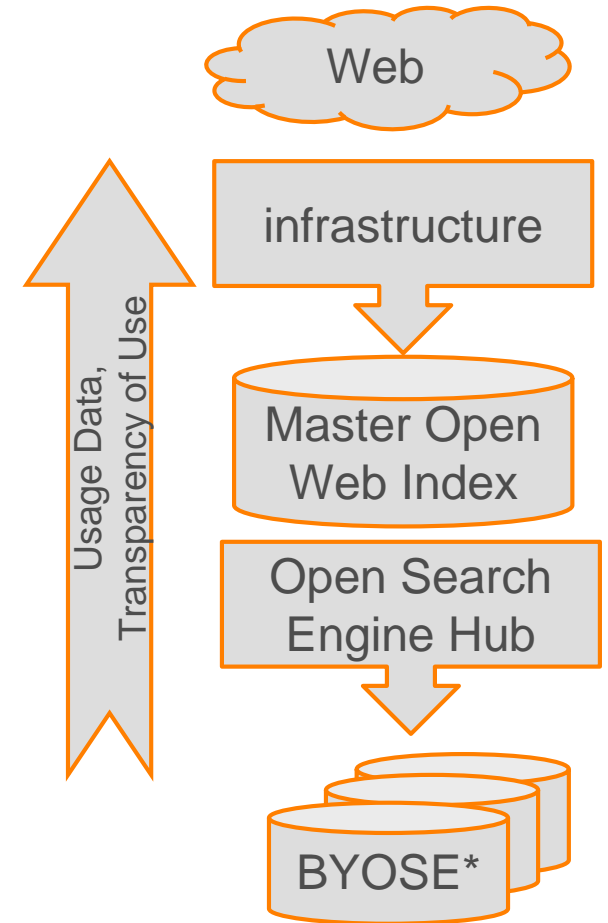
- A single box + ranked list



*Not only technical challenges, but also legal and societal challenges; overcoming technical challenges via crowdsourcing*

# Our Goal: Build an Open Web Index Collaboratively

- Build an Open Web Index including the corresponding pipelines and infrastructure
- Empower users, researchers & innovators to build on top of the Index
- Principles for an Open Web Index
  - Open Data: Slice'n dice the index as needed
  - Open Source / Open Configuration: Know the tech stack and its usage, extend if needed
  - Open Resources: fair-use access and you can bring your own resources
  - Open to contributions from third parties (e.g. semantic Enrichment)
  - Collaborative Management of a Web Index
  - Transparency / control to the content owners – respect legal, societal and ethical frameworks



\* BYOSE: Build your Own Search Engine



# Objectives

# Objectives

## Objective 4: Ecosystem

- Community Building
- Dissemination and Exploitation
- Simulating a competitive search engine market and web-data products
- Third Party Calls



## Objective 2: Added Value



### Vertical Search Engines

(Open Science Search, Mobile location Search, 3rd Party)



### Novel Search Paradigms

(Personal Search, Argumentation search, Conversational Search)



### High-Quality Web Data Collections

(cleaned, preprocessed, annotated)



### Knowledge Representation Models

(Knowledge Graphs, Neural Language Models)

## Objective 1: Technology Stack

- Coordinated Crawling
- Extensible Content Analysis
- Federated Indexing and Search
- Scalable, federated infrastructure



## Objective 3: Infrastructure and R&D Network

- Infrastructure Pilot
- Feasibility Study and Cost Estimation
- Governance Structure
- Platform for providers and consumers of data products and services





## Content Information

- Genres
- Topics / Concepts
- Geo-References
- Information Quality
- Ethics (e.g. Hatespeech).

## Legal Information

- License (CC-\*)
- Personal Information
- Legal Content

## Content Reuse Properties

- Indexing Y/N
- Data Mining Use Y/N...

## Website Usage

- Applied Semantic Enrichment Alg.
- Engines that indexed a site
- Access statistics
- User Ratings
- Blacklists / Whitelists
- Inclusion stats in Search Engines

## Access Information

- Reliability, Return Codes
- Access Time, Change Time
- API availability ..

## Topological Information

- Site structure (sub-sites)
- In- /outlinks to Websites

**Information Sources:** Crawler, Website Owner, Content Creator, Automatic Analysis, Logs, Users/SE Providers

**At Scale:** Cover >60% of the Text Web

# The Approach

Resources /  
Ecosystem /  
Target Stakeholders

Third Party Services  
and Data Products

OpenWebSearch.EU  
Service Infrastructure

OpenWebSearch.EU  
Storage Infrastructure  
/ Types of data products

Provenance chain for  
legal, ethical and societal  
considerations

Two verticals:  
Open Science Search & Mobile Search

Search Paradigms:  
argumentation search, conversational  
search

30-50% of commercial indices (html only)

Entities / Components	Technical Specification
Estimate for storage raw data (replicated 3 times)	1500 TiB
Estimated size of the Open Web Index (replicated 3 times)	500 TiB (Fast Access)
Estimated demands for temporary storage for intermediate results	1000 TiB
Node requirements for storage and analytics computations	25 Nodes a 96 cores & 256 GiB RAM
Node requirements for serving the index	70 Nodes a 48 cores & 256 GiB RAM

Critical: Extensibility and openness

# Example: Argument Search



healthy eating

All Discussions News People

PRO

## Arguments "A 100 gram portion (3.5 ounces) of raw ground..."

► Show full argument

Arguments "A 100 gram portion (3.5 ounces) of raw ground beef contains large amounts of Vitamin B12, B3 (Niacin), B6, Iron, Zinc, Selenium and plenty of other vitamins and minerals " "Studies now show that saturated fat in ...

<https://www.debate.org/debates/Eating-meat-is-healthy/1/> score ▾

## schools should definatly not enforce healthy eating!!!...

► Show full argument

schools should definatly not enforce **healthy eating!!!** it's the childs and parents job to make sure the kid doesn't end up as a fat blob

<https://www.debate.org/debates/schools-should-not-enforce-healthy-eating/1/> score ▾

## i asked some other kids and they said they would HATE...

► Show full argument

i asked some other kids and they said they would HATE school if school enforced **healthy eating**. ... i think the kids are responsible for what they eat and if it's now the schools turn to take over what kids eat.. it's the end of ...

<https://www.debate.org/debates/schools-should-not-enforce-healthy-eating/1/> score ▾

## I will need to see sources for your census to believe its

I am currently in italy and eating a lot of italian food. Is this healthy, what would you recommend me to eat and how much weight will i gain?

I would recommend eating a variety of Italian foods in moderation. Weight gain will depend on how many calories you consume overall.

What is typical healthy italian food?

Some typical healthy Italian foods include lean protein sources such as grilled chicken or fish, vegetables, whole grain pasta or bread, and fruit.

How much wine can i drink?

You can drink up to one glass of wine per day.

Can you cite some scientific papers that proof your statement?

Yes. Here are a few scientific papers that support the health benefits of moderate wine consumption:

1. <https://www.ncbi.nlm.nih.gov/pubmed/17956989>

2. <https://www.ncbi.nlm.nih.gov/pubmed/19335713>

3. <https://www.ncbi.nlm.nih>

# Outcomes (Planned)

# Some Envisioned Key Innovations

---



- Open Management of Website Data
- Open pre-processing and new semantic enrichment for information quality and ethical considerations
- Two search verticals (Open Science and Location-based search) and new search paradigms
- Open Search Engine Hubs - Install a search engine like a virtual machine including personal search engines or augmented enterprise search (e.g. only take the part of the web the interests you or your organisation)
- Ethical, legal and social concerns
- Towards a European open search association: Joining infrastructure organisations, researchers and innovations to bootstrap an infrastructure

*Bootstrapping the ecosystem: 1.3 M  
for third-party funding through open  
calls*

*Key Question: Feasibility of  
such an infrastructure?*



## **Opening up the search market**

- Search engines with very different flavours and purposes
- Choose the search engine you prefer, similar to the choice of your newspaper

## **Support the development of [new] search paradigms at large scale**

- Argumentation search, conversational search, geo-centered search, privacy
- HCI and UI concept at scale

## **Ease the utilization of clean Web Data**

- Neural Language Models, Data Augmentation ...
- Study trends at Web level: changes in end use licenses after GDPR, behavioural data

## **Web Search as a multiplier Service**

- Integration with other Data Spaces (e.g. EOSC, GAIA-X, Enterprise Search, Clouds)

## **Empower users, researchers and innovators at scale**

- **No substitution of major players: (i) we can't and (ii) we do it differently**
- Opening up the search market and tapping the web as resource
- Three Pillars: Tech, Network, Ecosystem
- Collaborative, open approach for building an Open Web Index
- Let's do it together: third party funds for bootstrapping
- Caveat: OpenWebSearch.EU can only bootstrap the approach. More efforts needed to go beyond
- Involve the Community – Funding available for outside parties
  - Public call to contributors: 1.3 Million EUR on the scope of OpenWebSearch.EU
  - Further NGI funds on sister project NGI Search: (<https://www.ngisearch.eu/>)

Thank You.  
Questions?