OpenWebSearch.EU
Towards a European Websearch and analysis system

https://openwebsearch.eu/

Prof. Dr. Michael Granitzer

UNIVERSITÄT PASSAU
Fakultät für Informatik und Mathematik

To be supported by European Union

Partners

Status: entered negotiation European funding, 3 years, 12+2 partners

Agenda

- Web Search and Digital Sovereignity
- Towards a Collaborative and Open Web Search Index
- Three Pillars of an Open Search Index [in Europe]
  - Technology
  - Network of Providers
  - Ecosystem
- Approach, key innovations and impact

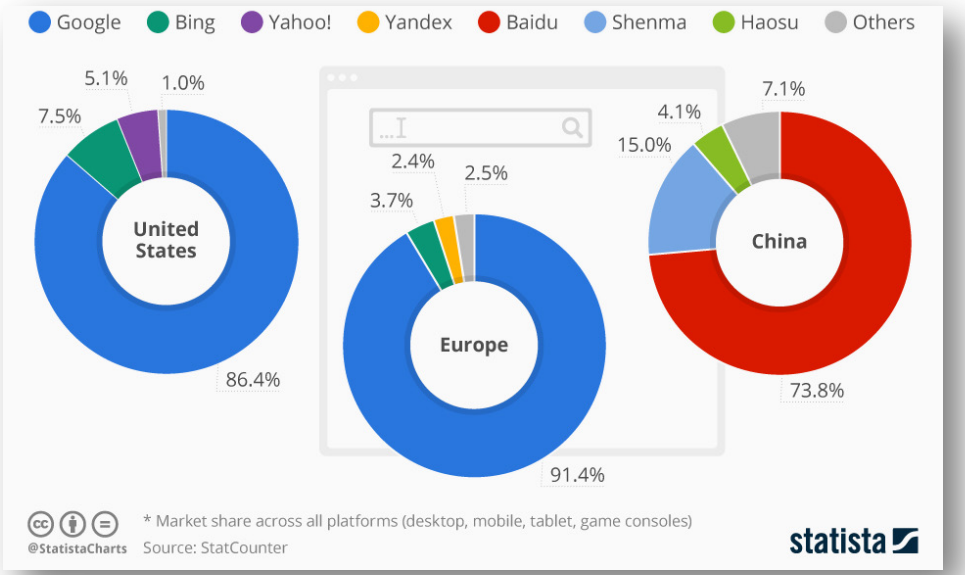
Web Search: Critical Infrastructure + Oligolopy

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)

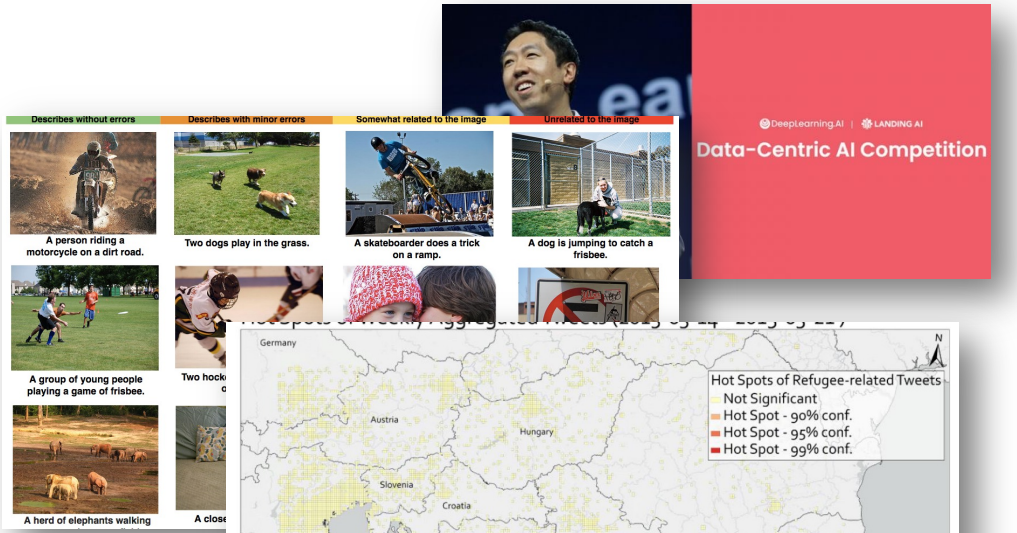
Effects

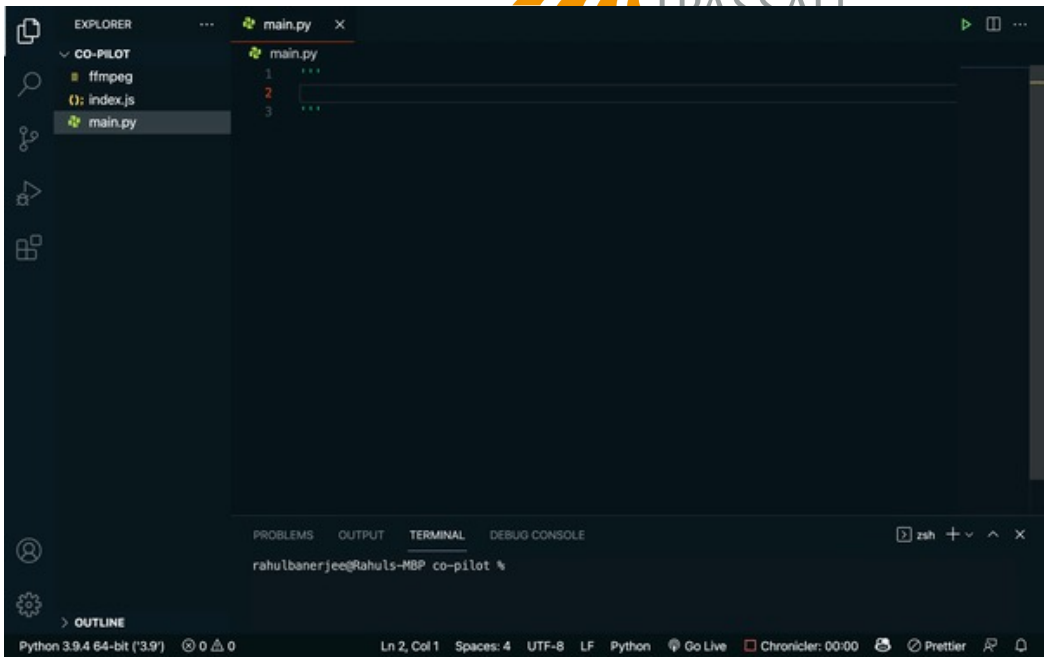
- Reduced User Choice
- SEO optimized ranking vs. best information delivery?
- Rich-gets-Richer effects
- User locked-in despite of "Open" technologies
- Concerning market behaviour (e.g. Jedi Blue)
- Limited buisness models
- ….



Beyond ~~whining~~ oligopolies: The Web as Resource

Web data drives innovation beyond search





Microsoft copilot trained on github data

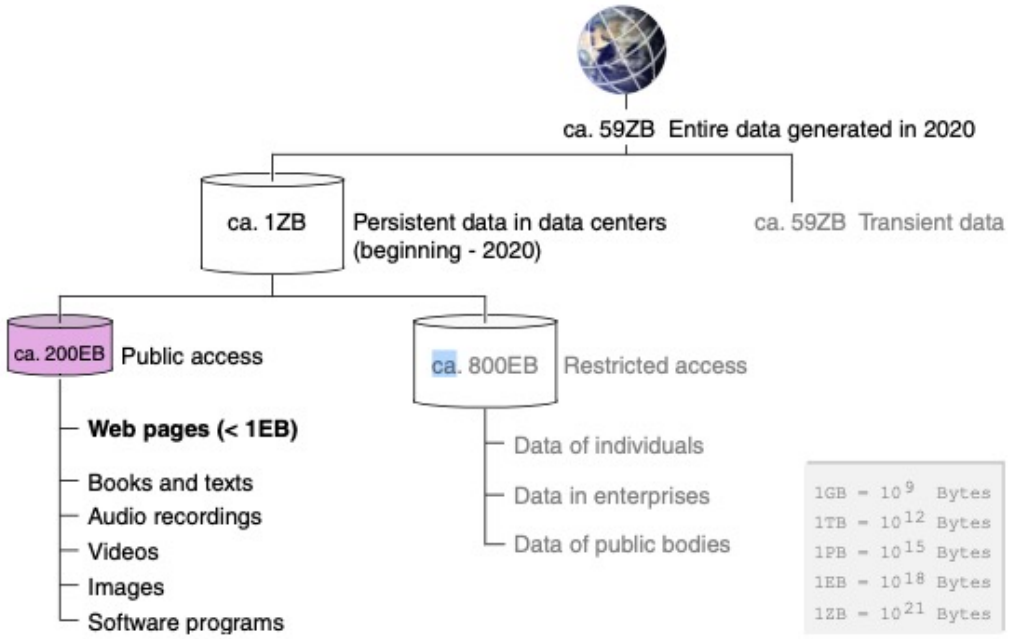
Human behaviour analysis, detecting migration routes

Havas, C., Wendlinger, L., Stier, J., Julka, S., Krieger, V., Ferner, C., ... & Resch, B. (2021). Spatio-temporal machine learning analysis of social media data and refugee movement statistics. *ISPRS International Journal of Geo-Information*, *10*(8), 498.

Tapping the web as resource

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

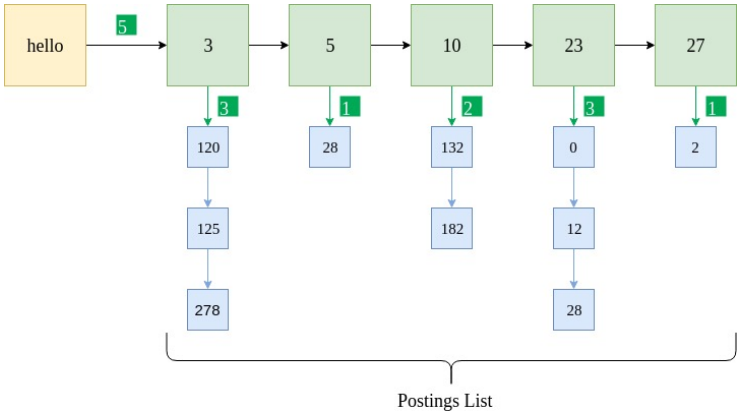
- High-demands on hardware resources
- High level of technological skill
  - Infrastructure
  - Big Data computing
  - Data cleaning
  - Natural Language Processing & Computer Vision
- Need only for particular subsets of the data
- Legal and ethical constraints (e.g. GDPR)
- Competitive, partially adversarial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

Goal: Build an Open Web Index Collaboratively

- Web Index
  - Data structure for fast access to web documents / sites
  - Supports search and ranking criterions
  - Build by crawling the web and preprocessing HTML
  - Enirchment: Geo-tagging, information extraction



- Principles for an Open Web Index
  - Open Data: Slice'n dice the index as needed
  - Open Source/ Open Configuration: Know the tech stack, extend if possible and needed
  - Open Resources: fair-use access and you can bring your own resources
  - Open to contributions from third parties (e.g. content push instead of pull)
  - Collaborative Information Management – Quality instead of Quantity
  - Control to the content owners – respect legal and ethical frameworks

Objectives, Outcomes, Values

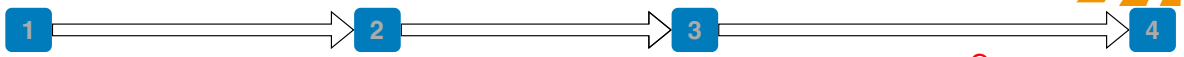
UNIVERSITÄT PASSAU

Symbolic and sub-symbolic Knowledge Representation	Curated High-Quality Web Data Collections	New Search Paradigms	User choice, Transparency, Privacy and Trust

Open Web Index

Pillar 1: Technology	**Pillar 2: Infrastructure Operators**	**Pillar 3: Ecosystem**
An open and extensible technology stack for Coordinated Crawling, Extensible Content Analysis and Indexing, Search Architecture(s) and Paradigms, Data Products and Services	An open and extensible (non-profit) Network of infrastructure hubs, researchers and business organisations providing resources, standardisation, governance etc.	around the Open Web Index and its Data Products / Services for advancing a competitive web search and web data market

The Approach



Resources / Ecosystem / Target Stakeholders

Webmaster

Websites

Data Curator
Data Wrangler

Technology Provider
Researcher

Innovator
Businesses / Service Provider

Open Science Search & Mobile Search

Two verticals:

Third Party Services and Data Products

Third Party Crawlers

Third Party Crawl Storage

Third Party Enrichment Services

Third Party Search Engines

Third Party AI Models and Services

OpenWebSearch.EU Service Infrastructure

Web Site Registry

Crawler Coordination

OWS Crawler

Web Crawl Management

Web Crawl Selection

Enirchment Plugin Management

OWS Semantic Enrichment Services

Indexing & Management Coordination

Search Engine Hub

Data & Knowledge Curation

New Search Paradigms:
argumentation search, conversational search

OpenWebSearch.EU Storage Infrastructure / Types of data products

Open Website Index (OWSI)

Open Crawl Storage Index (OCSI)

Curated and enriched high-quality web content

Open Web Index (OWI)

Open Knowledge Representation Models

Provenance chain for legal, ethical and societal considerations

Website specific Privacy Information

Website specific License

License compliant storage and access

Privacy-aware processing

License compliant processing

Privacy-aware processing

License compliant processing

Some Envisioned Key Innovations

- Open Management of Web**site** Data
- Open pre-processing and. new semantic enrichment for information quality and ethical considerations
- Two search verticals and new search paradigms
- Open Search Engine Hubs - Install a search engine like a virtual machine
- 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

Sister-Project: NGI Search over 6 M for Third-party calls beyond OpenWebSearch.EU

Impact of an Open Web Index

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 …

Web Search as a multiplier Service

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

Empower researchers and innovators at scale

Conclusio

UNIVERSITÄT PASSAU

- 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 – joining efforts and resources
- Let’s do it together: third party funds will be available to support your innovation ideas
- Caveat: OpenWebSearch.EU can only bootstrap the approach. More efforts needed to go beyond

Thank You.
Questions?