Open WebSearch

UNIVERSITÄT PASSAU
Fakultät für Informatik und Mathematik

OpenWebSearch.eu
Current Status of the project after one year

Prof. Dr. Michael Granitzer
Chair of Data Science University of Passau and Coordinator
OpenWebSearch.eu

Open Web Search

Funded by
the European Union

SUPPORTED BY NGI

OpenWebSearch.eu will create an open European infrastructure for internet search, based on European values and jurisdiction

Open WebSearch .eu

What?

Restore an open search ecosystem / market as a basis for a new Internet Search

- → lay a foundation for a new Internet search
- → contribute to Europe's digital sovereignty
- → empower Europe's researchers, innovators and businesses to systematically tap into the Web as business and innovation resource

Why?

1. Web search is dominated and limited by a few gatekeepers like Google, Microsoft, Baidu, Yandex.

Resulting situation:

- → unilateral, biased, opaque access to information
- → locked-in effects

2. Tapping the Web as resource is challenging for innovators and researchers

Who?

14 renowned European universities + institutions will pool their expertise and resources.

- → including some of the largest research and computing centres in Europe
- → e.g. IT4Innovations, Leibniz Supercomputing Centre, CSC, European Organisation for Nuclear Research CERN

How?

Develop the core of a European Open Web Index

Four Objectives

1. Open Technology Stack
2. Resource provision by a network of infrastructure providers
3. Added value services
4. Bootstrapping the ecosystem

14 Partners plus Third Party Calls

Open WebSearch.eu

UNIVERSITÄT PASSAU

Webis.de

UNIVERSITÄT LEIPZIG

Bauhaus-Universität Weimar

Radboud Universiteit

DLR

TU Graz

VSB TECHNICAL UNIVERSITY OF OSTRAVA | IT4INNOVATIONS NATIONAL SUPERCOMPUTING CENTER

-lutions for Brilliant Minds

+ 6 Third Parties from our calls:

- **LISA:** Open Web Search - Legal, Intellectual Property and Cyber-Security Aspect
- **ALMASTIC**, Assessing Legal Risks and Mitigating Challenges in Open Web Indexes **MRC**, Market potential assessment for OpenWebSearch.EU: Quantifying benefits and costs of scaling EU web search
- **LOREN**, Legal Open euRopean wEb iNdex (Germany),
- **LAW4OSAI**, License-Aware Web Crawling for Open Search AI
- **OC**, Open Console

nlnet FOUNDATION

suma-ev

Infrastructure

Research

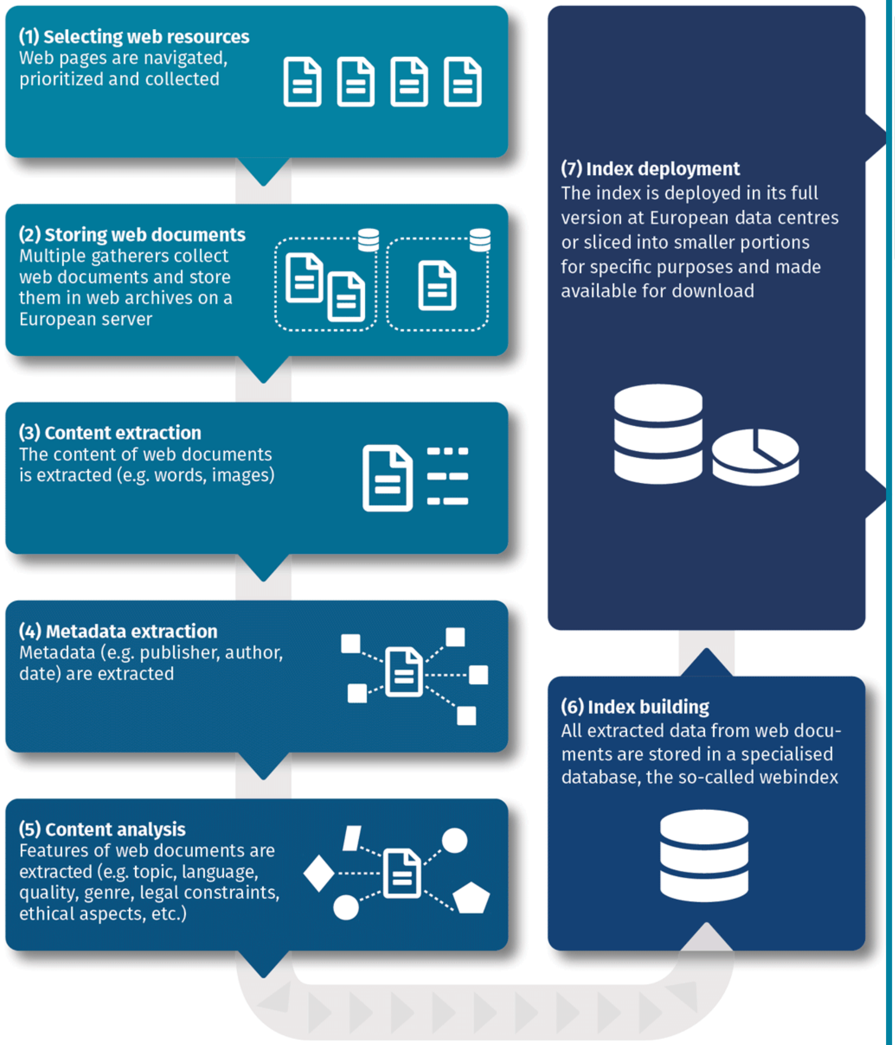
NGOs

Businesses

Our Approach

Index Generation

Web resources are selected and retrieved, their content and metadata are analysed, and all data stored in the index database.



(1) Selecting web resources
Web pages are navigated, prioritized and collected

(2) Storing web documents
Multiple gatherers collect web documents and store them in web archives on a European server

(3) Content extraction
The content of web documents is extracted (e.g. words, images)

(4) Metadata extraction
Metadata (e.g. publisher, author, date) are extracted

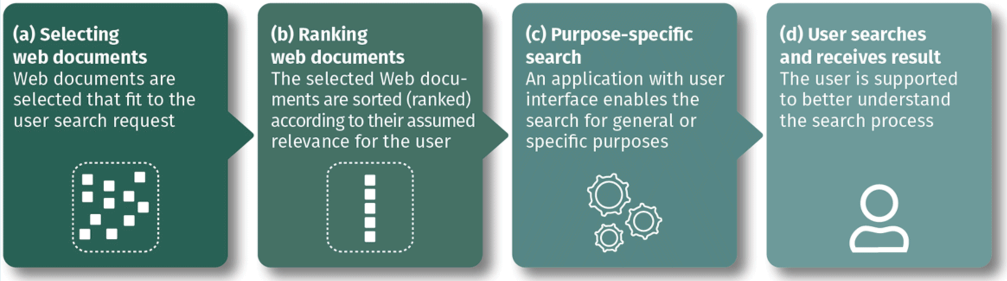
(5) Content analysis
Features of web documents are extracted (e.g. topic, language, quality, genre, legal constraints, ethical aspects, etc.)

(6) Index building
All extracted data from web documents are stored in a specialised database, the so-called webindex

(7) Index deployment
The index is deployed in its full version at European data centres or sliced into smaller portions for specific purposes and made available for download

Search Applications

A user search request will be answered by a search application that makes use of the open web index.



(a) Selecting web documents
Web documents are selected that fit to the user search request

(b) Ranking web documents
The selected Web documents are sorted (ranked) according to their assumed relevance for the user

(c) Purpose-specific search
An application with user interface enables the search for general or specific purposes

(d) User searches and receives result
The user is supported to better understand the search process

Data Products

Knowledge representation models will be created using the open web index, in order to be used by any agent and for many applications

Building knowledge graphs
Using the extracted information from web documents, a knowledge graph is created that supports specific search requests

Building AI Language Models
Creation of different types of language models by using Web documents

Any agent, multiple applications
Language models and knowledge graphs can be used by any agent (or application)

• • •

Open WebSearch.eu

LUMI@CSC

KAROLINA@IT4I

Leibniz Supercomputing Centre
of the Bavarian Academy of Sciences and Humanities

CERN

DLR

Web-scale Platform for heavy-lifting

Applications and Innovations as Multiplicator

Distributed Infrastructure as Enabler

Current Status – OSSYM Talks

Index Generation

Web resources are selected and retrieved, their content and metadata are analysed, and all data stored in the index database.

OWLer

~ 1 TiB / day

~ 5M URLs / day /machine

2 Crawl Centers

1 Central frontier

Spam and Topic Detection

"A Comprehensive Dataset for Webpage Classification"

Mohammed Al-Maamari et a.

Resiliparse Library

Fast WARC parsing

(Character encoding, Main Text Extraction, Language Detection (101 Languages) and Topic Identification)

Indexing and Provision in CIFF Format

Challenges of index exchange for search engine interoperability,

Hiemstra, Hendriksen, Kamphuis, de Vries

Web-scale Platform for heavy-lifting

Search Applications

A user search request will be answered by a search application that makes use of the open web index.

Using CIFF Format in Retrieval Engines

Conceptual Design and Implementation of a Prototype Search Application using the Open Web Search Index, Nussbaumer et al.

Evaluation of IR Engines

Prototyping Open Web Search Applications with TIRA: A Case Study in Research-oriented Teaching, Fröbe e al.

Web Data Analytics

Product Spam on YouTube: A Case Study, Bevendorff et al.

Studying Future Search Paradigms

Commercialized Generative AI: A Critical Study of the Feasibility and Ethics of Generating Native Advertising Using Large Language Models in Conversational Web Search, Zelch et al.

Data Products and Innovations

Cooperate via Open Console, Mark Overmeer

Applications and Innovations as Multiplicator

Governance and Community Building

"Governance Towards an Open Web Index"

Jasmin Tietgen, Maari Alanko

Distributed Infrastructure as Enabler

Conclusion

- Web and Web Search critical for Europe’s digital sovereignty. An Open Web Index to the rescue!
- We have so far
  - A first running pipeline with ~ 1 TiB/day, but download facilities are not yet in place
  - Use of the index in standard retrieval libraries
  - Application scenarios in development
  - ELSA Considerations and Governance Planning
  - Lots of challenges ahead!
- Detailed results presented in the next 4 hours

Thanks. Questions?

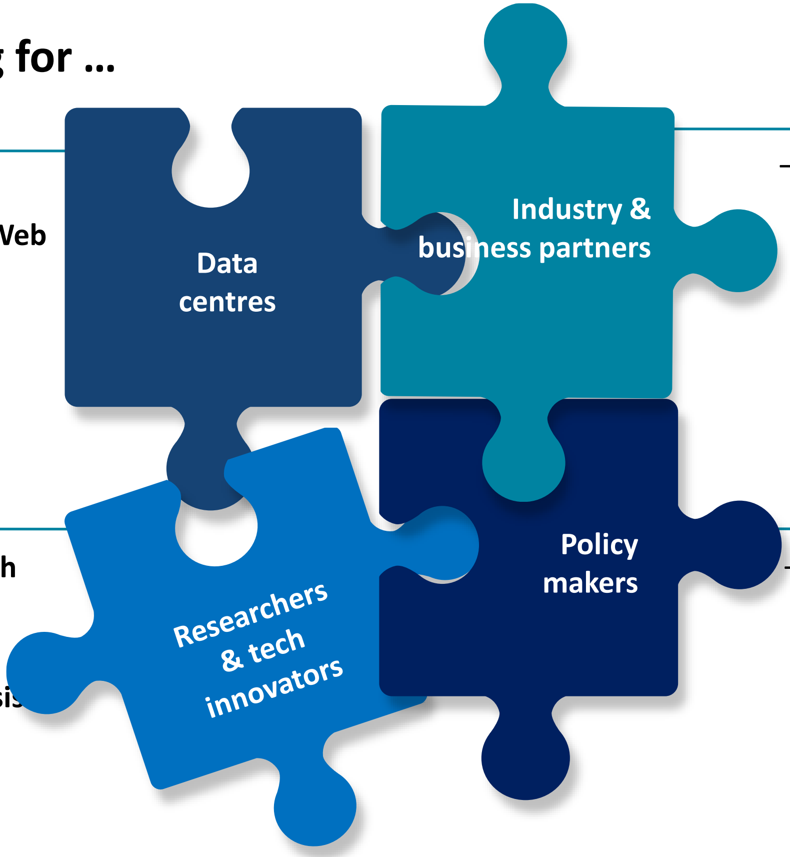
Contact us:

To keep in touch with these possibilities or to join us send an email to **join@openwebsearch.org**

We are looking for …

→ **help hosting a distributed Open Web Index**

→ **develop new search & retrieval paradigms and content analysis algorithms**



Data centres

Industry & business partners

Researchers & tech innovators

Policy makers

→ **discover the business models of an Open Web Index**

→ **help shaping the governance of an open search ecosystem**

→ **We will also offer small grants for potential contributors.**

Watch the upcoming calls for fundings or meet us at our Open Search Symposium at CERN

https://opensearchfoundation.org/en/events-osf/5th-international-open-search-symposium-ossym2023/#ossym-program