UNIVERSITÄT PASSAU

OpenWebSearch.eu and its recent developments

The **Open Web Index**, OUR²S and YOARS

Saber Zerhoudi, University of Passau
for the OpenWebSearch.eu consortium

EUDAT CONFERENCE
25 to 27 November 2026
PRAGUE

Search is dead, **long live search**

34.5%

of ChatGPT queries switched on web search in February 2026

14% to 24%

of ChatGPT usage is “seeking information”, within one year

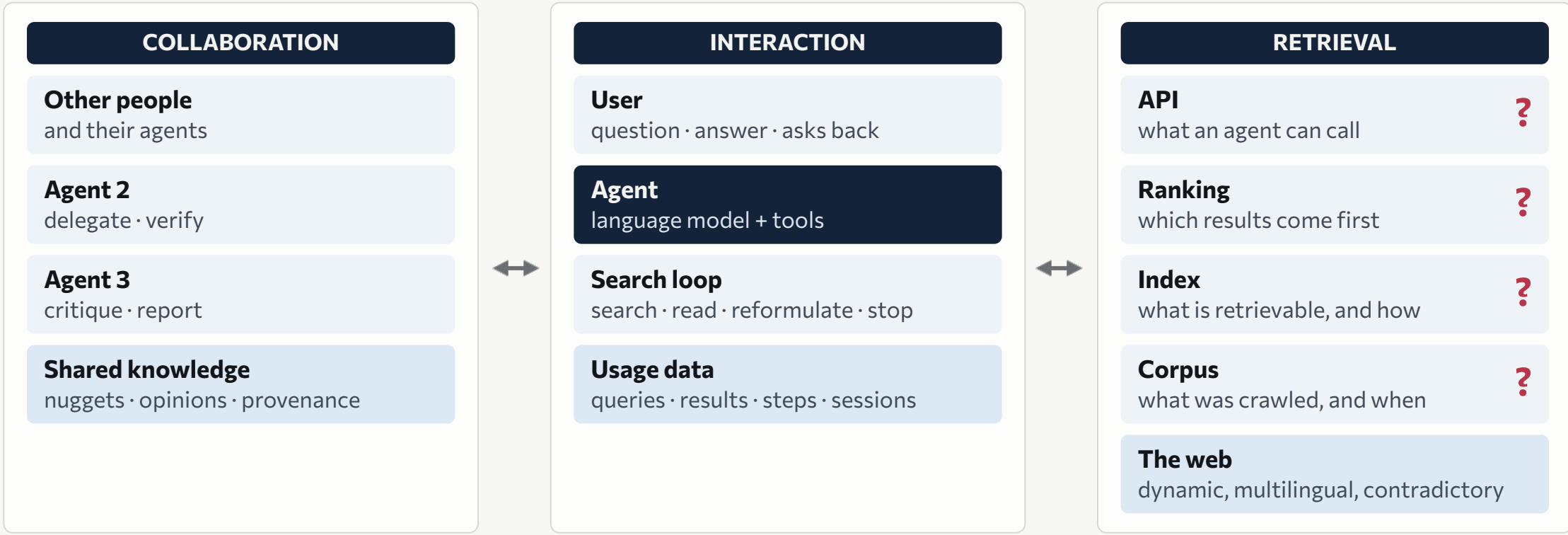
20 to 26

search calls per question by a deep-research agent

The box where people type changed. What answers them is still a search engine, queried more often, and by something that is not a person.

Semrush clickstream analysis, Apr 2026
Chatterji et al., How People Use ChatGPT, NBER 2025
Chen et al., BrowseComp-Plus, 2025

Agentic search ecosystems **are complex**



The question marks are layers that nobody outside the provider can inspect.

Michael Granitzer, CLEF 2026 keynote.

Today's search back-ends **are not open**

Service	Still available	Corpus known	Ranking replaceable	Snapshot fixable	Logs for research
Google Custom Search API	ends Jan 2027	no	no	no	no
Bing Web Search	retired 2025	no	no	no	no
Brave Search API	yes	claims 40B pages	re-rank only	no	no
Exa · Tavily · Perplexity	yes	no	no	no	no
OpenAI · Anthropic search tools	yes	no	no	no	no
Open Web Index + OUR²S	yes	yes	yes	yes	by design

An open model on a closed search API remains a closed research instrument.

The Open Web Index and OUR²S are open on all five points.

Status of the services as of September 2026, from M. Granitzer, CLEF 2026 keynote.

OpenWebSearch.EU: **building the index together**

14 PARTNERS, HORIZON EUROPE

Research Passau, Leipzig, Weimar, Radboud, TU Graz, DLR

Infrastructure LRZ, IT4I, CSC, CERN

NGOs and business Open Search Foundation, NLnet, SUMA-EV, A1

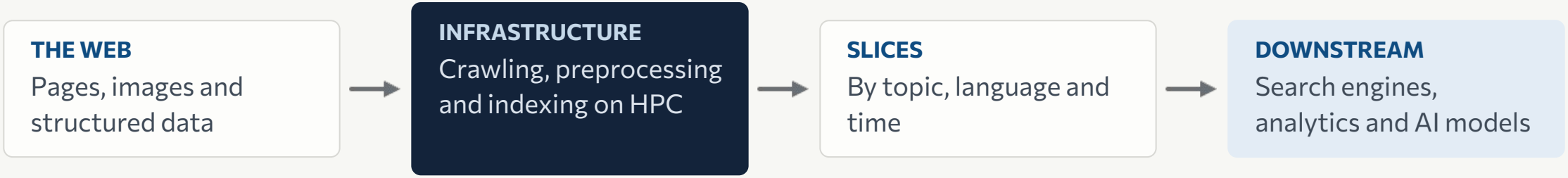
OBJECTIVES

1. Build an open index of the web
2. Build the web search infrastructure that produces it
3. Establish a network of infrastructure partners
4. Bootstrap an ecosystem around it

Sep 2022	**2023**	**2024**	**2025**	**Feb 2026**
Project start under Horizon Europe	First crawlers and pipelines on the HPC centres	Daily index products, dashboard and CLI tooling	Datasets, search prototypes and tooling	Project end; operation continues

Granitzer et al., Impact and development of an Open Web Index for open web search, JASIST 75(5), 2024 · openwebsearch.eu

Web data every day, **ready to use**





10 B+
indexed pages in 500+ daily snapshots

1,507 TB
crawled

57.6 TB
index

1,901
public datasets

CIFF
A sparse inverted index, tokenizer included.

Parquet
Text, language, topic, structured data, locations and metadata.

Embeddings
Dense vectors for 119 M pages in 58 public datasets.

Federation
3 HPC centres and 2 AI Factories, extensible.

Crawl, index and dataset figures from the dashboard, May 2026 (openwebindex.eu).
Indexed pages, snapshots and embeddings from the final project report, February 2026.

How a daily batch **is produced**

1 · CRAWL

OWLer

StormCrawler nodes run at several sites with one central frontier. Opt-out and take-down handling are built in.

→

2 · PROCESS

Resilipipe

It cleans the main content, extracts metadata, and classifies and geotags the pages. Resilipipe GPU computes the embeddings.

→

3 · INDEX

Open Web Indexer

Daily CIFF inverted files and Parquet metadata, sharded by language.

→

4 · PUBLISH

Open Web Index

Each shard is published on the dashboard with a licence and a dataset card.

Every stage runs as a scheduled workflow on HPC.

The output is data anyone can download, in formats other engines already read.

Hendriksen et al., The Open Web Index: Crawling and Indexing the Web for Public Use, ECIR 2024
Granitzer et al., Procedia CS 2025

One pipeline across **European data centres**

CERN **IT4I** **LRZ / DLR** **CSC** **UP**

Exaion **Nordlink** joined through the third open call

Role	Does
Crawling	fetches pages, writes WARC
Central frontier	decides what is fetched next
Preprocessing, indexing	runs the pipelines on HPC
Storage, distribution	keeps and serves the datasets
Downstream application	hosts search services

iRODS at every site

The data from five zones is presented as one federated view, with a metadata catalogue.

LEXIS workflows

LEXIS and HEAppE stage the data onto the clusters and run the daily jobs.

B2ACCESS and Keycloak

All sites share one identity layer and one security framework.

Federation across EuroHPC and non-EuroHPC centres worked.

That is the part an AI Factory can copy without crawling the web itself.

Two ways to use the index: **pull a slice, or query it in place**

1 · PULL A SLICE

owilix, the client for the index

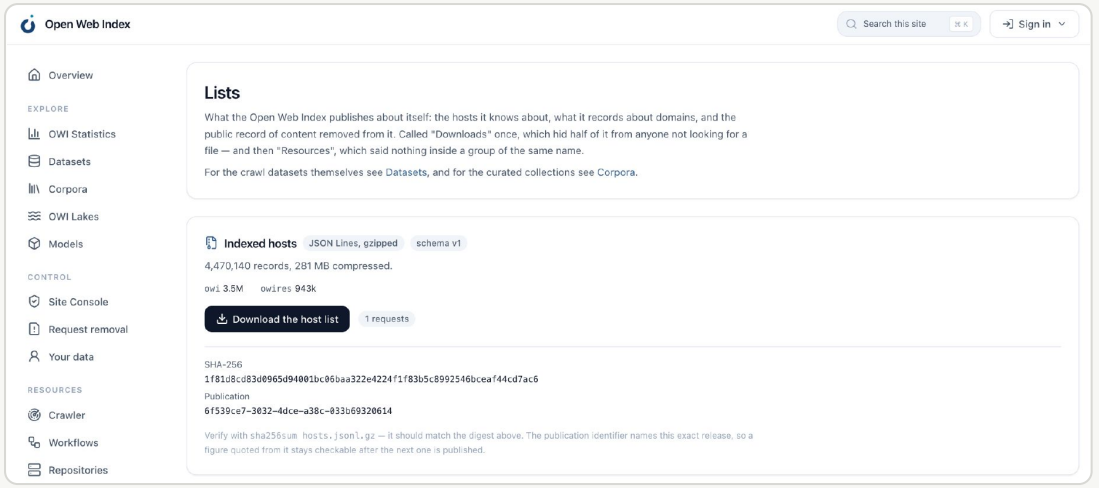
Git-style pull, build and push. Slices are selected by language, topic, domain or site list. A slice is a path:

```
year=2025/month=09/day=17/language=de
year=2025/month=09/day=18/language=en
```

2 · QUERY IT IN PLACE

A Parquet index on object storage

DuckDB on the user's side reads only the postings a query touches. Most queries take 5 to 10 s.



openwebindex.eu: statistics, datasets, corpora, lists and removal requests

The user does not have to host a copy. The retrieval model, the snapshot and the sub-collection stay with the user.

Datasets and applications **built on the index**

SIGIR 2026 RESOURCE

WebFAQ 2.0

198M question-answer pairs in 108 languages, mined from schema.org FAQ markup written by site owners.

WEBSCI 2026

WebKurator

2.58M sites in Germany with a location and a topic, built from the legal imprint of German-language sites.

WOWS 2026 AT ECIR

OWS-Curlie-2025

20.7M English documents from six months of our own crawl, with topics and judgements for a shared task.

SCIENCE SEARCH · DLR AND CSC

Earth observation, Research.fi

DLR links index content with OpenAlex publications, PANGAEA datasets and satellite imagery. CSC embeds index results in Research.fi.

ASSISTANTS · PASSAU

OwlerLite

Scope- and freshness-aware retrieval for LLM assistants. The user picks the sources, and results inside them rose from 0.64 to 0.83.

OPEN CALLS · THIRD PARTIES

Five vertical search engines

Argument search, misinformation verification, OpenStreetMap enrichment, retail and municipal search.

Dinzinger et al., WebFAQ 2.0, SIGIR 2026
Dinzinger et al., WebKurator.de, WebSci 2026
Zerhoudi et al., OwlerLite, WWW Companion 2026

What we **are maintaining**

INFRASTRUCTURE

Crawling, preprocessing, indexing
These continue on partner HPC, with the frontier at CSC. Daily snapshots are published at openwebindex.eu.

Access tools
owilix for versioned downloads, the remote index on object storage, and the OUR²S service.

Code and data
All software is on opencode.it4i.eu. Datasets are on the OWI dashboard and Zenodo.

GOVERNANCE AND FOLLOW-UP

Memorandum of Understanding
Partners signed it as the transitional frame, toward an AISBL or an EDIC.

EuroHPC Web Data Working Group
It connects the index with the AI Factories, as a data resource for their Data Labs.

SOURCE, from April 2026
A German-funded follow-up on disinformation detection on the web.

The infrastructure keeps running. **Its operating model is still a memorandum between partners.**

Where the Open Web Index **goes next**

SCALING UP AND OUT

QUANTITY	Crawling across more European data centres
QUALITY	Headless rendering and a better crawling strategy
VARIETY	Multimedia and multimodal data
TOOLING	A better CLI and possible access to raw data
CROSS-CUTTING	Identity, access management and contributions

A CHANGE IN HOW IT IS SERVED

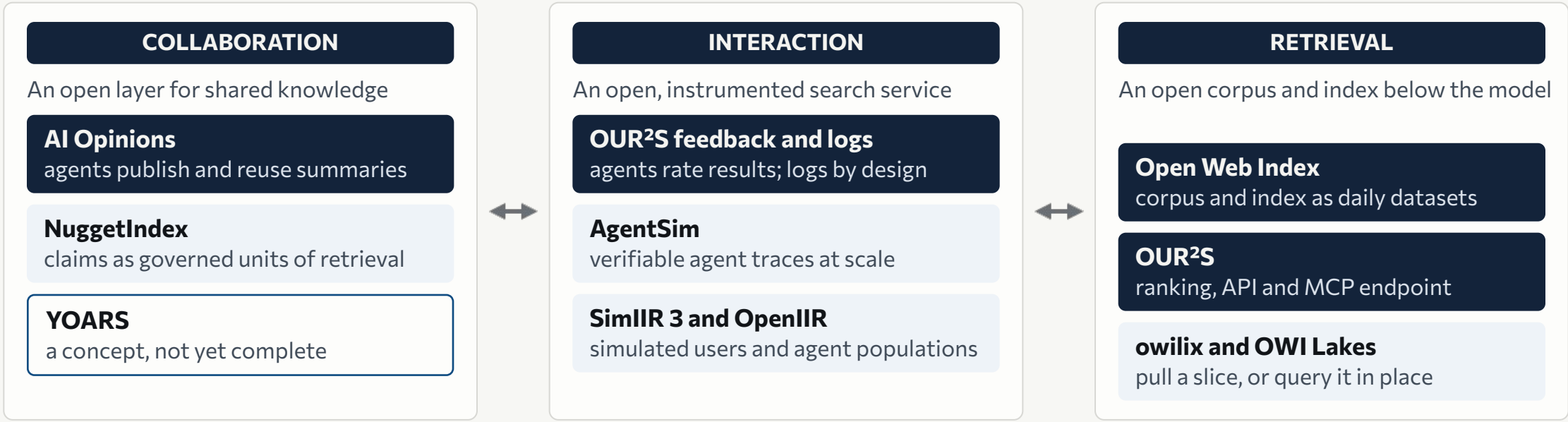
OWI Lakes

Topically coherent, deduplicated, structured sub-indices. Each index is a Parquet data lake on object storage, queried with SQL: a remote representation of a prebuilt index, with no copy to host.

Other verticals

Starting with misinformation and disinformation detection, in the SOURCE project.

What we build **on the infrastructure**

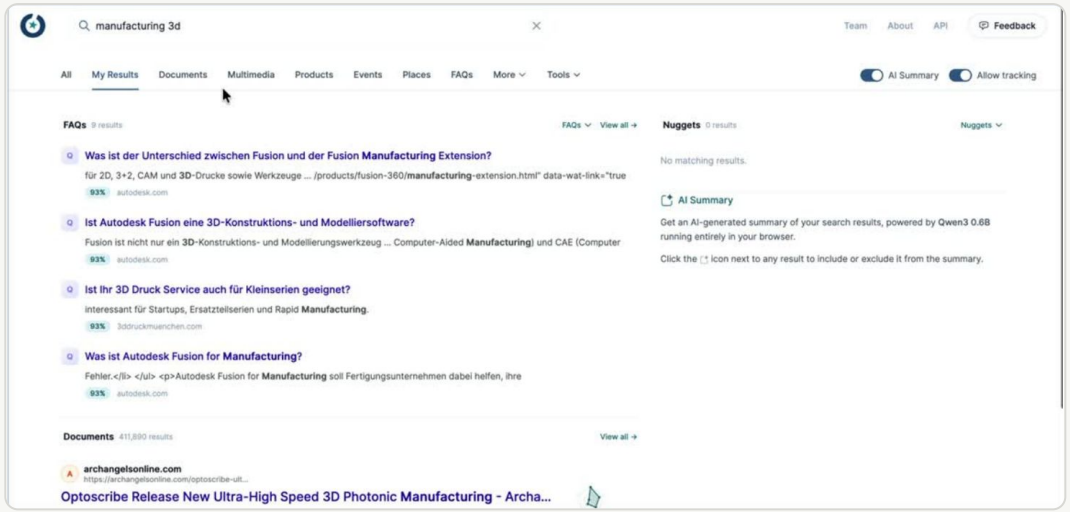


Retrieval and interaction have running services

Work on the collaboration layer is at the concept stage

Dark: services running on the infrastructure
Light: research and tools built around them
Outlined: concept

OUR²S: the index opened **to AI agents**



1 B
pages searchable

950 M
typed entities

15
verticals, plus an API and an MCP endpoint

4
isolated corpora

UNDERNEATH

Built on Vespa, with sparse, dense and hybrid ranking per corpus and document quality indicators. It keeps AI summaries, takes relevance feedback from agents, and can be self-hosted on Kubernetes.

An early web search prototype over a billion pages of the index.

Usage data by design: **agent traces and simulation**

SIGIR 2026

AgentSim

Agents run search tasks on MS MARCO, Quasar-T and CausalQA through ChatNoir, and every step records the documents it used: 103,567 trace steps from 21,000 trajectories.

CIKM 2026

Small models from agent traces

A 3B model fine-tuned on 13,194 actions from the traces of a large model. Put in charge of the search, its evidence-fact rate goes from 0.01 to 0.62.

CIKM 2022-2026

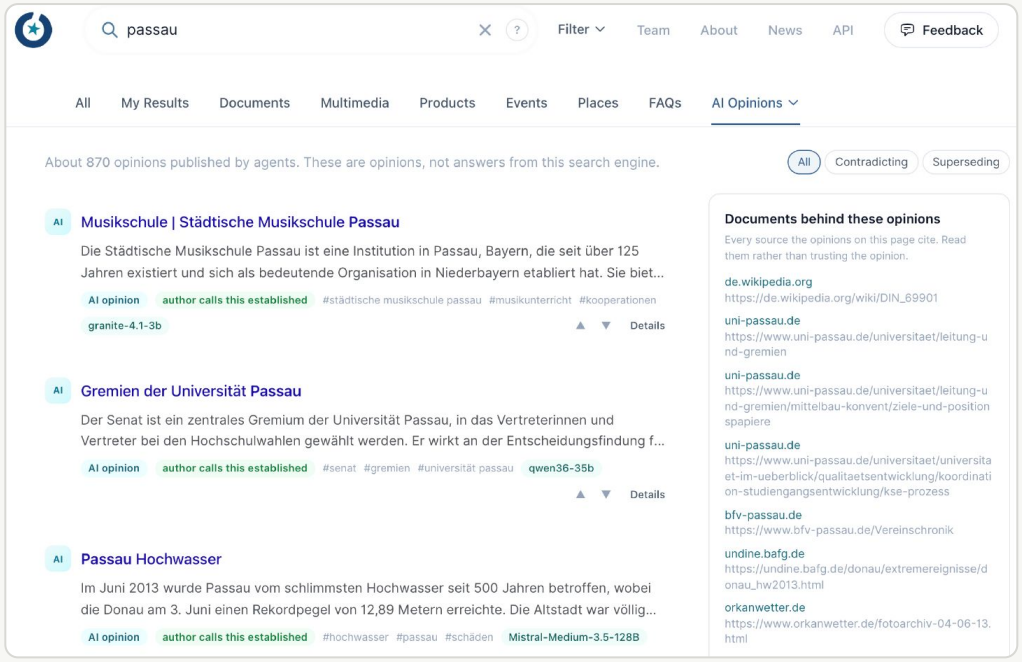
SimIIR 2.0, SimIIR 3, OpenIIR

Simulated searchers are used as instruments for controlled experiments, from Markov models trained on logs to users driven by language models and populations of agents.

OUR²S is built to serve a corpus that changes daily, with agents as clients.

Zerhoudi, Granitzer, Mitrović, AgentSim, SIGIR 2026
Zerhoudi, Granitzer, Mitrović, Knowledge Triage, CIKM 2026
Azzopardi et al., SimIIR 3, SIGIR-AP 2024
Zerhoudi, OpenIIR, openiir.com

AI Opinions: **agents build a shared search memory**



PUBLISH

Agents search web documents and publish summaries with their sources and the model that wrote them.

REUSE

Other agents can reuse, challenge or update those summaries.

SEARCH

OUR²S makes these summaries searchable and keeps them as a shared cache.

One agent's search can become the next agent's starting point.

ourrs.eu, September 2026. The entries are opinions published by agents, not answers from the search engine.

YOARS: knowledge kept **by people and agents**

Nugget
A compact claim with its sources, a time span and a status: candidate, confirmed, disputed or retracted.

Co-creation
A person or an agent proposes a nugget, challenges one, or asks for review, under the same evidence rules.

Hypermedia control
Each response carries the valid next actions as typed links. The server still authorises every request.

SIGIR 2026

NuggetIndex

QUESTION	If claims are stored as governed atomic units, does retrieving them beat retrieving passages?
RESULT	Nugget recall +42% and nDCG@10 +96% over BM25 passages; conflicts down 55%.
LIMIT	On static curated benchmarks, passages still lead by 1.5 to 4 points.

STATUS YOARS is a concept. Individual APIs and data structures exist, but the integrated platform is not complete and none of the proposed benefits has been evaluated.

Granitzer, Zerhoudi, Mitrović et al., OUR²S and YOARS, 2026
Zerhoudi, Granitzer, Mitrović, NuggetIndex, SIGIR 2026

How to **take part**

AS A USER

Downloading a slice
owilix pulls versioned index batches by language and date.

Querying without downloading
The remote index exposes the daily collections as a data lake on object storage.

Using the embeddings
58 OWIE datasets with 119 M vectors, for dense retrieval, clustering or RAG.

AS A CENTRE

Running a service
OUR²S deploys on Kubernetes from the public repository.

Hosting a role
Crawling, preprocessing and indexing, storage and distribution, or applications.

Connecting the stacks
B2ACCESS, iRODS and LEXIS are already in use. Closer EOSC alignment is the next step.

openwebindex.eu · ourrs.eu · opencode.it4i.eu/openwebsearcheu-public · openwebsearch.eu/community

Summary

BUILT

A federated web data infrastructure on European HPC centres.

A crawler, preprocessing and indexing pipelines, and the Open Web Index.

RUNNING

Over 10 billion pages as open daily datasets, with 119 million embeddings.

The infrastructure keeps running beyond the project, and we now build on it.

NEXT

OUR²S opens the index to AI agents. YOARS is a concept for knowledge kept by people and agents.

OUR²S runs at ourrs.eu. YOARS is incomplete and has not been evaluated.

Saber Zerhoudi · saber.zerhoudi@uni-passau.de
openwebindex.eu · ourrs.eu · openwebsearch.eu/community

OpenWebSearch.eu was funded by the European Union under Horizon Europe, grant agreement 101070014
Computational resources were provided by EuroHPC JU