OpenWebSearch.eu - Building an Open Index of the Web on HPC Infrastructure

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

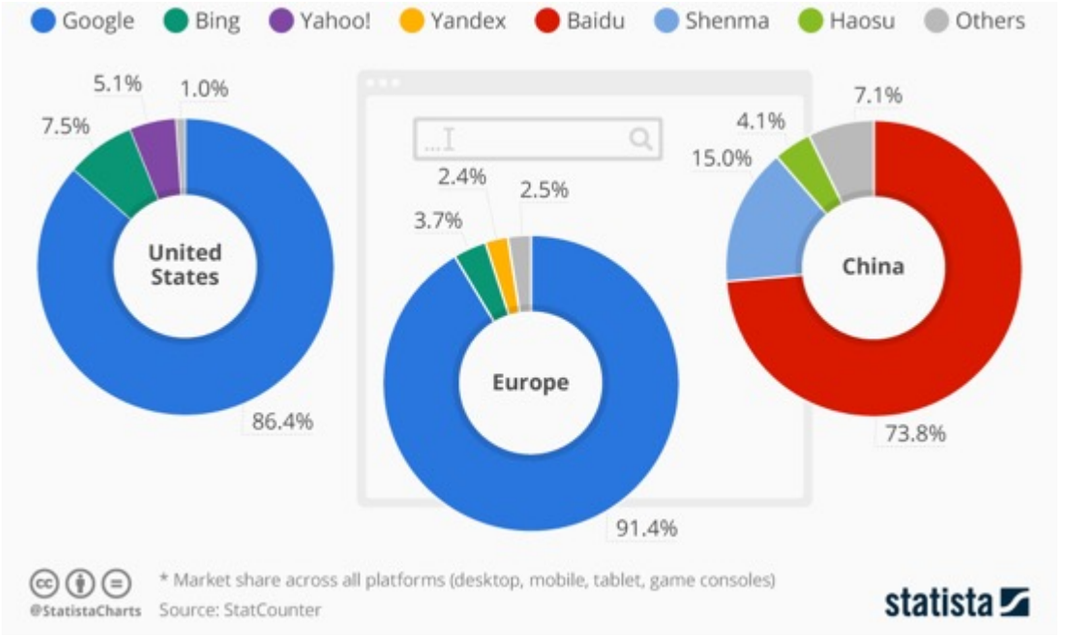
HPCSE 2026 · IT4Innovations · 2026-05-20

Web Search Today

A critical resource managed as an oligopoly

Two properties that don't fit

- A **critical infrastructure** for society, comparable to satellite navigation
- A **market oligopoly**: dominated by a small number of gatekeepers (Google, Microsoft, Baidu, ...)



But does it matter? Yes — on four dimensions

User Choice & Democracy

- Single gatekeeper for society's information
- Search rankings can shift voting preferences by **20–72%** (SEME)
- Monopolistic ranking = single point of failure for information quality

Societal & Security Risks

- European OSINT depends on US-controlled infrastructure
- Filter bubbles, algorithmic bias
- Misinformation & Disinformation

Untapped Richest Information Source

- 60 % of web pages carry structured data
- No open access to web-scale document collections
- Science, research and technology heavily depend on the Web as platform

AI & Innovation

- Web data is the fuel for LLMs, RAG, and agentic AI
- Closed indices block reproducible research & open innovation

> Being able to navigate the Web is as fundamental as GPS or the power grid

Challenges

THE WEB: VOLUME, VELOCITY, VARIETY

Careful: domains, sites, pages, and indexed documents are different units

VOLUME

1.15 Billion
Netcraft site responses
272.6 M domains · Dec 2024 Netcraft survey

Top 0.1%–1% of sites
may account for roughly half of all pages
heavy-tail intuition from large web graph studies

~ 400 Billion
documents in Google's search index
2020 figure cited in US v. Google testimony

Sites ≠ domains ≠ pages; use exact counts cautiously

VELOCITY

+8.6 M / month
monthly net change in Netcraft site responses
Dec 2024 versus Nov 2024 Netcraft survey

~ 3.2 / second
same monthly net change, averaged across December

149 ZB (global)
all digital data created / copied worldwide in 2024
IDC / Statista global datasphere estimate

The web changes constantly; crawling is continuous, not one-off

VARIETY

150+ languages
observed across websites
W3Techs usage survey

249 CMS Platforms
CMS platforms identified in the HTTP Archive dataset
HTTP Archive Web Almanac 2024 (Wappalyzer-based)

HTML · PDF · JSON-LD · media · APIs
content types to parse and process

HTML · PDFs · JS apps · schemas · media · APIs

Sources: Netcraft Web Server Survey Dec 2024 · large web graph studies (heavy-tail intuition) ·
US v. Google testimony on ~400B documents (2020 figure) · HTTP Archive Web Almanac 2024 · W3Techs · IDC/Statista

- High infrastructure costs
- Broad range of technical skills required
  - Big Data Infrastructure
  - Natural Language Processing, Image Analysis
  - Web Technologies
- Growing legal uncertainties

Challenges

THE WEB: VOLUME, VELOCITY, VARIETY

Careful: domains, sites, pages, and indexed documents are different units

VOLUME

1.15 Billion
Netcraft site responses
272.6 M domains · Dec 2024 Netcraft survey

Top 0.1%–1% of sites
may account for roughly half of all pages
heavy-tail intuition from large web graph studies

~ 400 Billion
documents in Google's search index
2020 figure cited in US v. Google testimony

Sites ≠ domains ≠ pages; use exact counts cautiously

VELOCITY

+8.6 M / month
monthly net change in Netcraft site responses
Dec 2024 versus Nov 2024 Netcraft survey

~ 3.2 / second
same monthly net change, averaged across December

149 ZB (global)
all digital data created / copied worldwide in 2024
IDC / Statista global datasphere estimate

The web changes constantly; crawling is continuous, not one-off

VARIETY

150+ languages
observed across websites
W3Techs usage survey

249 CMS Platforms
CMS platforms identified in the HTTP Archive dataset
HTTP Archive Web Almanac 2024 (Wappalyzer-based)

HTML · PDF · JSON-LD · media · APIs
content types to parse and process

HTML · PDFs · JS apps · schemas · media · APIs

Sources: Netcraft Web Server Survey Dec 2024 · large web graph studies (heavy-tail intuition) ·
US v. Google testimony on ~400B documents (2020 figure) · HTTP Archive Web Almanac 2024 · W3Techs · IDC/Statista

- High infrastructure costs
- Broad range of technical skills required
  - Big Data Infrastructure
  - Natural Language Processing, Image Analysis
  - Web Technologies
- Growing legal uncertainties

High upfront costs to tap the Web as resources — especially small innovators and

researchers are left behind

OpenWebSearch.eu — Our Mission

Building an Open Index of the Web and an open infrastructure to reduce upfront costs for researchers, innovators and companies in tapping the Web as a resource for search, web-analytics and AI.

Four Objectives

1. **Open Technology Stack** Open-source pipelines for crawling, preprocessing, indexing
2. **Resource Provision** Building a network of infrastructure providers (EuroHPC / AI Factories)
3. **Added Value Services** Dashboard, tools, data access APIs
4. **Bootstrapping the Ecosystem** Third-party calls, community building, applications and use-cases

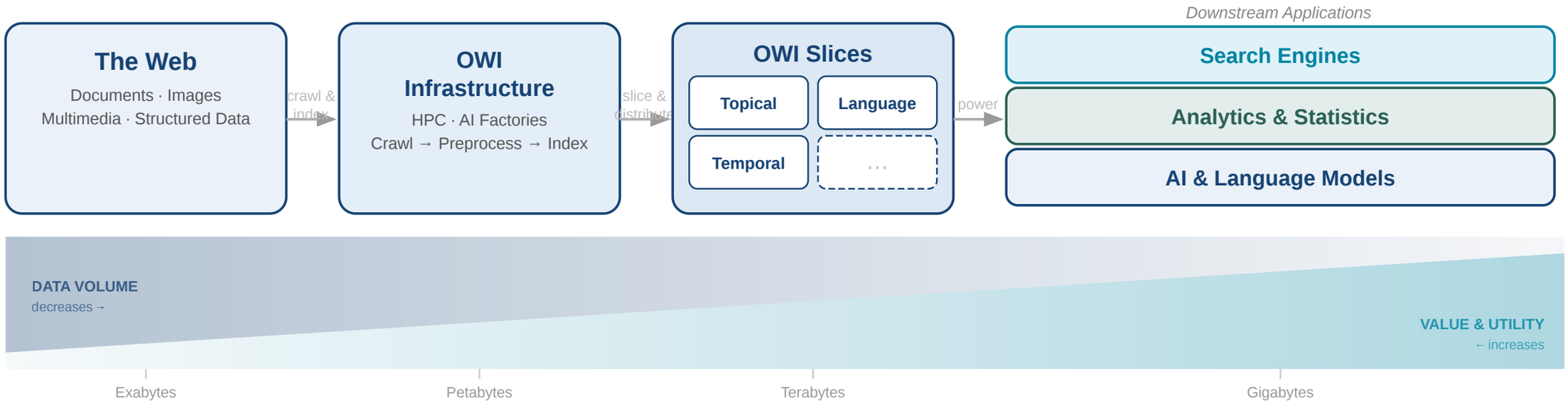
14 Core Partners

Plus 16 financially supported third-party projects (FSTPss)

The Open Web Index (OWI)

A **Web Index** = a data structure for fast query-based access and ranking of web documents — the core of every web search engine

Our Proposition: A collaboratively created, open and transparent Web Index running on existing HPC computing centres / AI Factories in Europe



Granitzer, Michael, et al. "Impact and development of an Open Web Index for open web search." Journal of the Association for Information Science and Technology (2023).

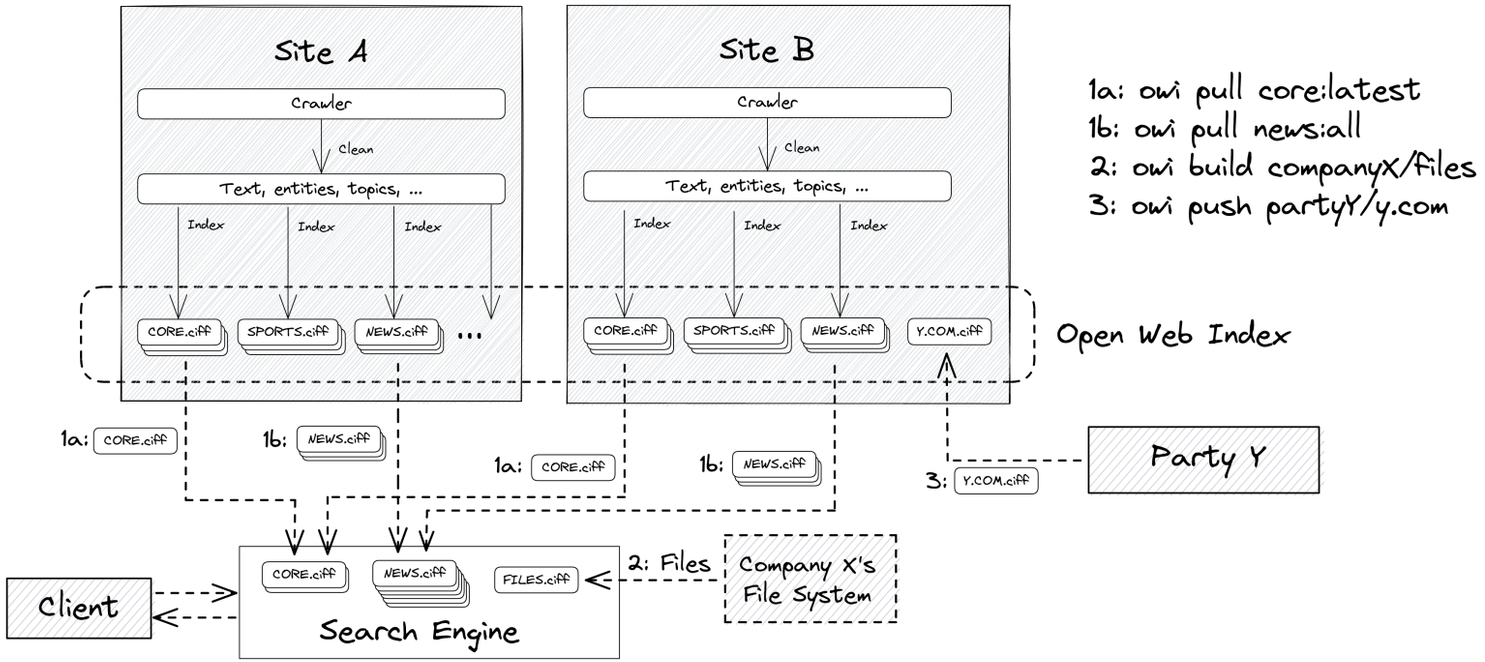
What is the Open Web Index?

What is the Open Web Index?

A **Web Index** = data structure for fast, ranked retrieval of web documents — the core of every search engine. The OWI provides such an index plus the associated metadata/text as open data.

What is the Open Web Index?

A **Web Index** = data structure for fast, ranked retrieval of web documents — the core of every search engine. The OWI provides such an index plus the associated metadata/text as open data.

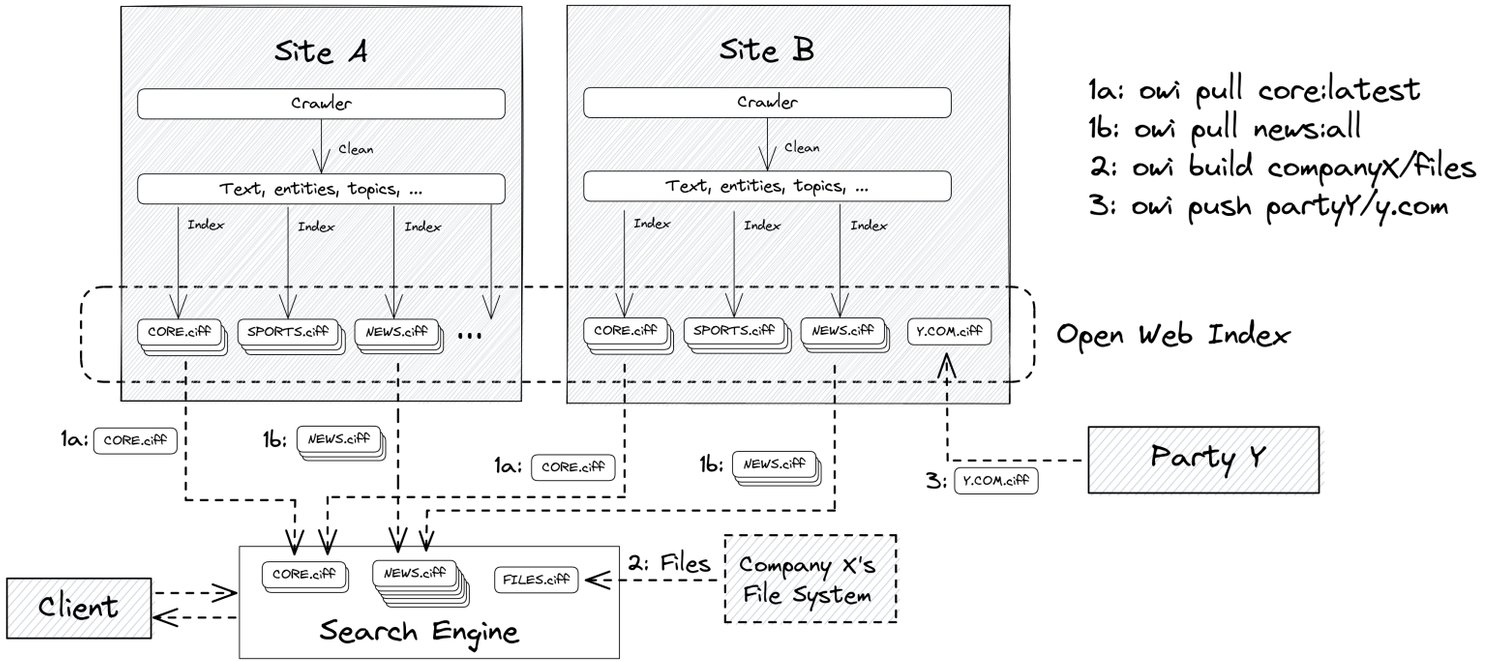


The OWI provides:

- **Pull** — download daily pre-built index shards (by date, language, topic)
- **Build** — create a local index from your own data
- **Push** — contribute custom indices back to the OWI

What is the Open Web Index?

A **Web Index** = data structure for fast, ranked retrieval of web documents — the core of every search engine. The OWI provides such an index plus the associated metadata/text as open data.



The OWI provides:

- **Pull** — download daily pre-built index shards (by date, language, topic)
- **Build** — create a local index from your own data
- **Push** — contribute custom indices back to the OWI

Three output types per daily run:

- `index.ciff.gz` — sparse inverted index (CIFF format; pre-configured tokenizer)
- `metadata_*.parquet` — document metadata & enrichments (topics, microdata, locations, language, etc.; extensible)
- Dense embeddings (GPU-enriched slice, Jina V3) for AI-based Search and Classification (on selected subsets)

OWI at a Glance — Statistics (Feb 2026)

Crawling

Pages crawled	~39 B total, ~12 B unique
Unique hosts	~500 M (extrapolated)
Crawling effort	1,927 machine-days
Avg. throughput	~20 M URLs / day

Index

Documents indexed	~10.7 B
Unique domains	~75 M
Public datasets	511 main / 1,511 total
Index size	45 TB (main) / 50 TB total
Embeddings	119 M docs · 884 GB

Top languages: English (38.9%), German (7.2%), Spanish (6.4%), French (5.7%), Russian (5.2%)

Comparison: To meet commercial index size, OWI would need to scale by a factor of 20× — factor 10–50× for multimedia

Building the Open Web Index

Conceptual Workflow

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

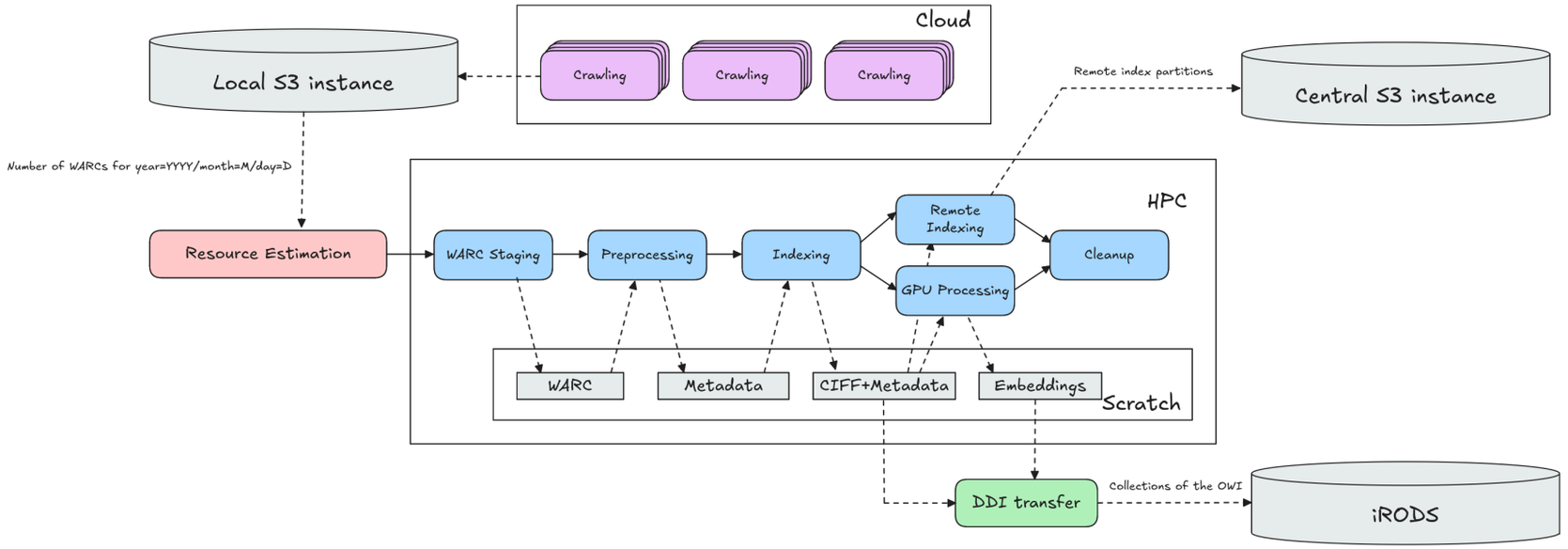
Building knowledge graphs
Using the extracted information from web documents, a knowledge graph is created that supports specific search requests

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)

...

Daily Workflow Pipeline



Infrastructure (2026)

- **25 VMs** across IT4I, LRZ, CSC, CERN
- **100 M URLs / day** crawling capacity
- **~1.1 PB** raw data · **~50 TB** index · **810+ public datasets**
- Federated storage: iRODS + S3, orchestrated via LEXIS

Limitations as of 2026

- TLR 5/6: it is not a product (yet)
- No 24/7 Operations
- Limited funding - minimum engineering / improvement

OWI on HPC Infrastructure

Why the OWI is an HPC Workload

Web-scale data path

- crawl and archive web pages as WARC data
- preprocess text, language, metadata, links, quality signals
- build sparse index shards in CIFF
- publish aligned Parquet metadata for analytics
- enrich selected slices with dense embeddings

The bottlenecks shift by stage

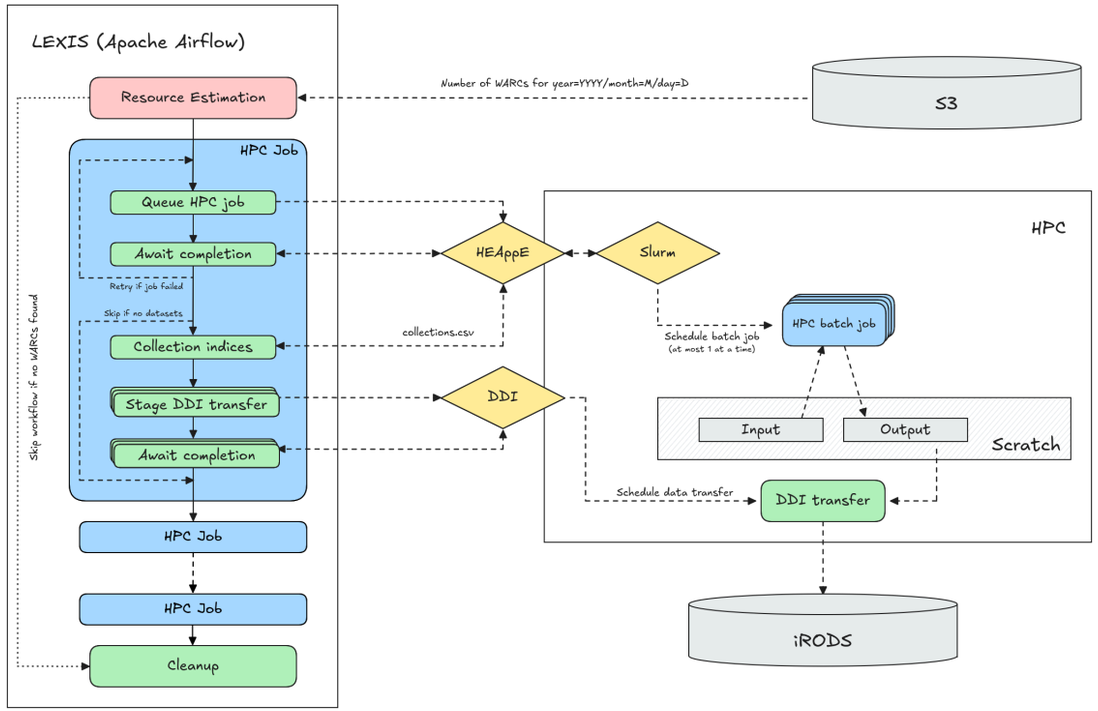
Stage	Dominant pressure
Crawling	network, scheduling, politeness
WARC processing	I/O, parsing, decompression
Indexing	memory, sorting, parsing, CPU
Metadata analytics	columnar scans
Embeddings	GPU throughput, vector storage

> The Open Web Index is not just a search demo. It is a recurring, data-intensive production workflow that turns web-scale raw data into reusable search and AI datasets.

How We Use LEXIS for the Open Web Index

LEXIS Worfklows to build the OWI

- coordinates OWI workflows across participating HPC/data centers
- connects compute jobs with federated storage locations
- helps stage input data and publish daily OWI outputs
- supports the operational bridge between crawling, processing, indexing, and distribution



Workflow orchestration via LEXIS

> LEXIS is the core component for how OpenWebSearch.eu turns distributed HPC resources into a repeatable Open Web Index build process.

From Indexing to Embedding at Scale

Classical OWI output

- sparse inverted files for lexical retrieval
- metadata partitions for filtering and analytics
- compact daily shards by date, language, data center, and collection

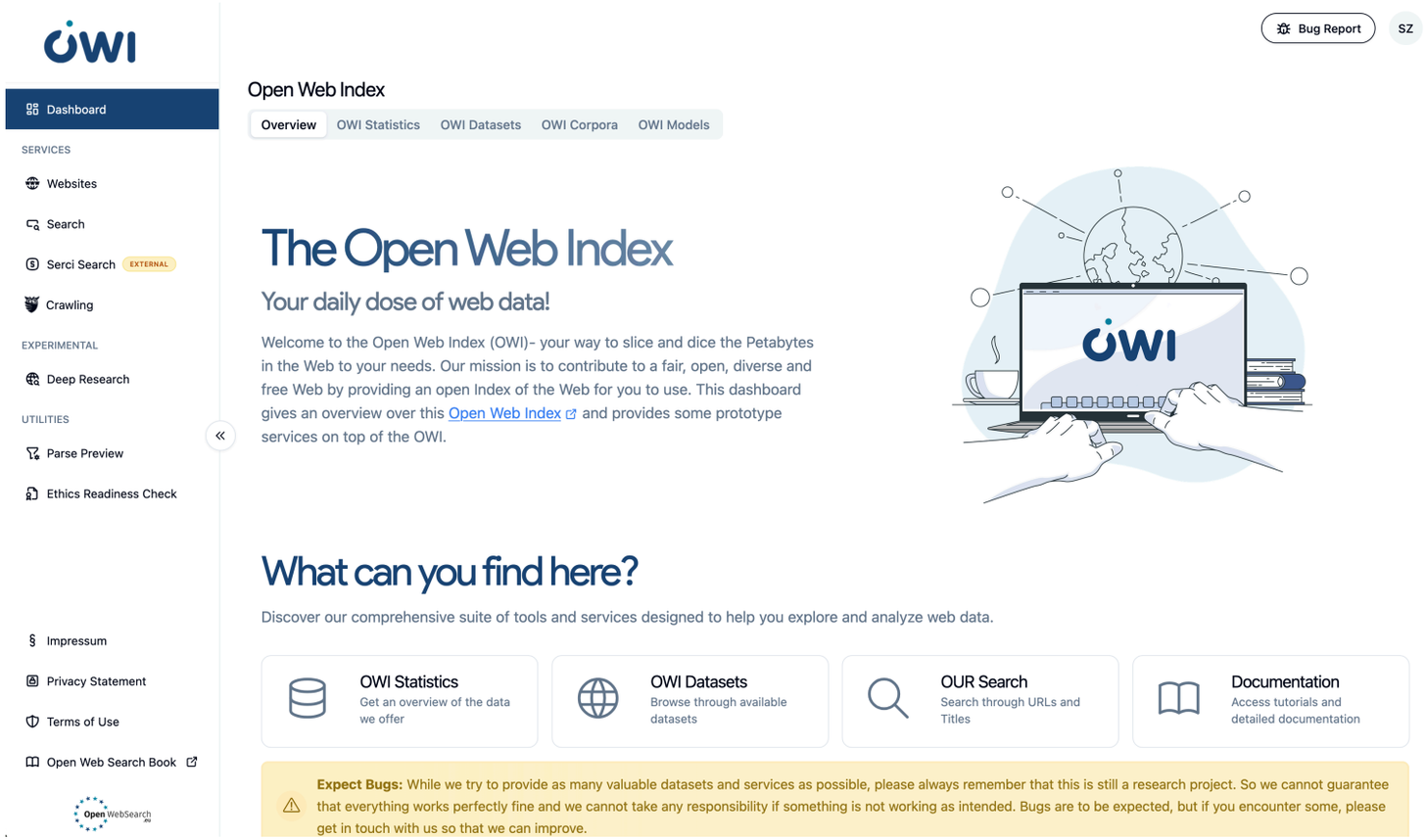
New AI-oriented output

- dense document embeddings for semantic retrieval
- topic/domain/language slices for RAG systems
- compression and pruning to make vector indices affordable
- GPU workloads that must be scheduled next to classic indexing jobs

> The HPC question is no longer only: "Can we crawl and index the web?" It is also: "Can we semantically enrich web-scale collections often enough to support AI search?"

Accessing the Open Web Index

The Dashboard — openwebindex.eu

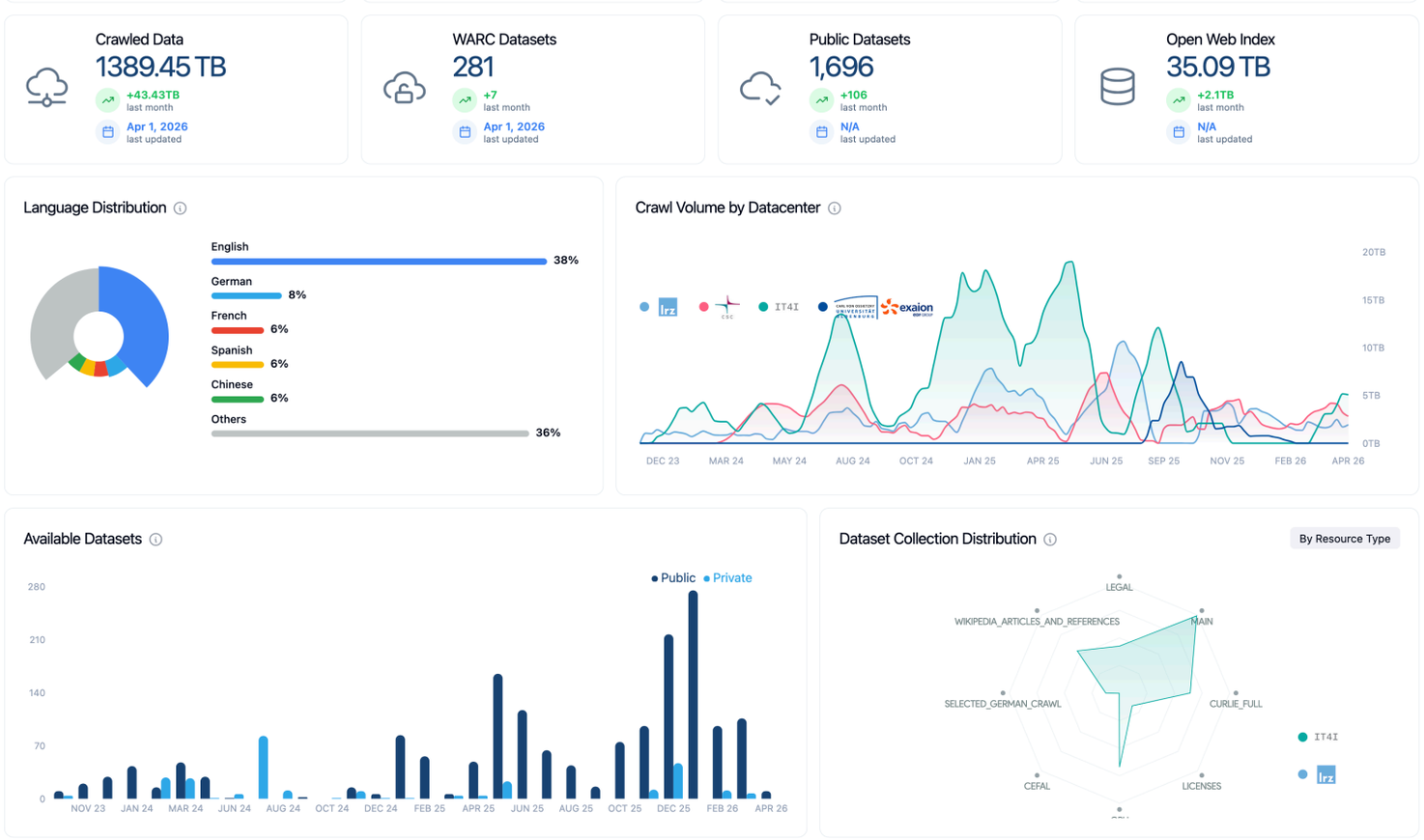


424 registered users · 17,000+ unique visitors · 70,000+ page views

Central entry point for the Open Web Index openwebindex.eu

- **OWI Statistics** — crawl volume, index size, language distribution, dataset timeline
- **OWI Datasets** — browse, filter and download index shards (by DC, date, format, language)
- **Search** — demo content search + SERCI integration

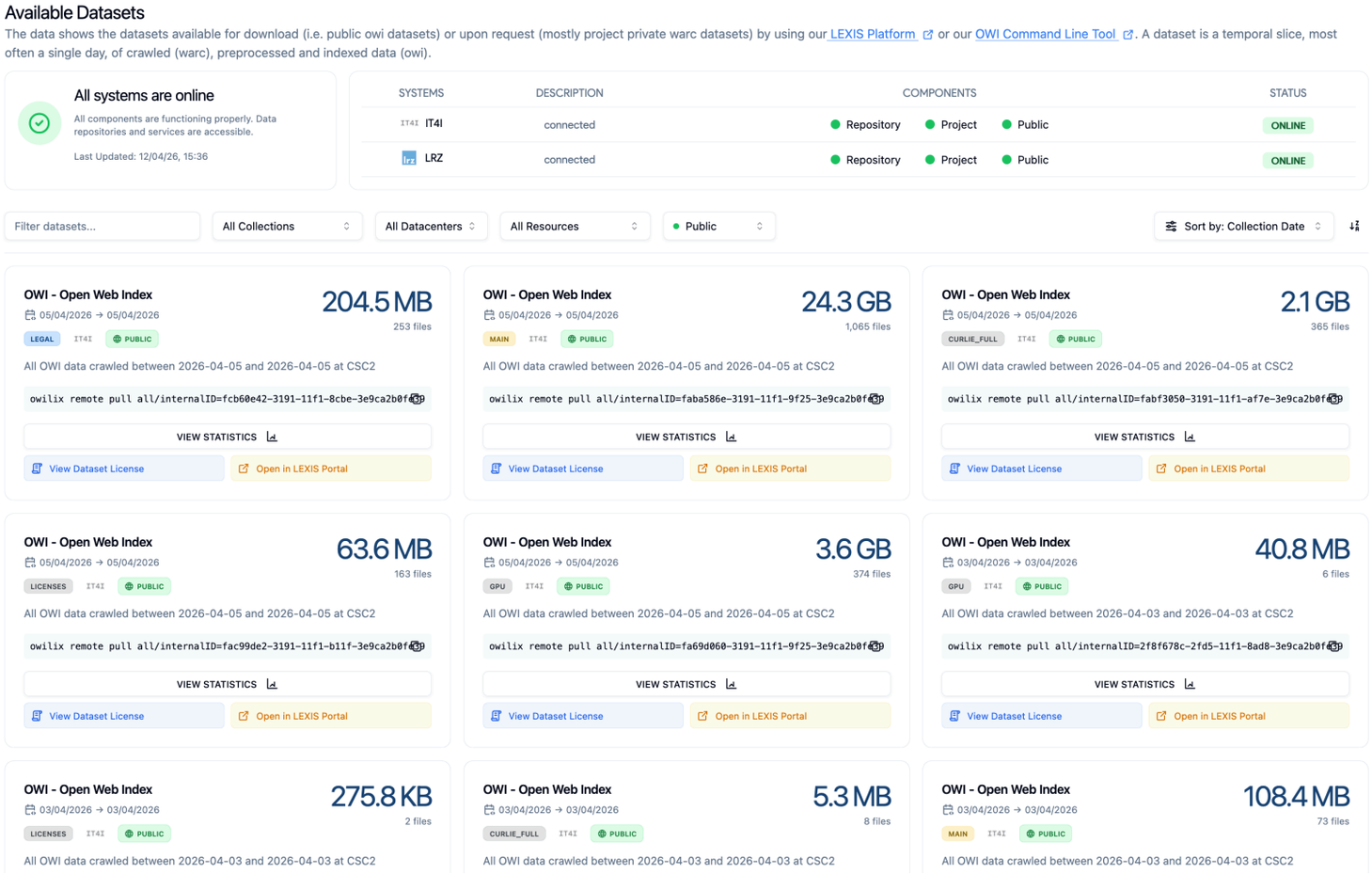
Dashboard — Statistics & Datasets



OWI Statistics (04/26)

- **1,389 TB** crawled · **281** WARC datasets · **1,696** public datasets · **35.09 TB** index
- Language distribution chart (English 38%, German 8%, French 6%, ...)
- Crawl volume over time per data center + dataset collection distribution by type
- Datasets over time

Dashboard — Statistics & Datasets



OWI Datasets Tab

- Status of repositories
- Filter by datacenter, format (WARC / Parquet / CIFF / Embeddings), language, date
- Dataset cards with download links — LEXIS Portal or owilix CLI
- Parquet file preview directly in the browser

Download Tool: owilix

owilix — The OWI Command-Line Tool

Git-like versioning for index datasets — pull, build, push:

```
owilix remote pull all:latest/collectionName=main
owilix local ls all:latest
```

Slice by: **domain**, **language**, **topic** (Curlie), **sitelist**, **structured data**, **outlinks**

Authentication via B2ACCESS / Keycloak (LEXIS SSO)

OWILIX - Open Web Index CLI

version 5.3.1 | python 3.11+ | license Apache 2.0

A command-line tool for managing OpenWebIndex datasets across local and remote repositories.

Quick Start

One-Line Install (Recommended)

```
curl -fsSL https://openwebindex.net/owilix/install.sh | sh
```

https://opencode.it4i.eu/openwebsearcheu-public/owi-cli/

```
[→  ~ owilix remote search "michael granitzer" --limit 10
Searching remote index: terms=['michael', 'granitzer'] language=eng representation=main_content mode=OR limit=10
Found 10 results in 113.4s
                    Search results (10 results in 113.4s)
```

score	date	url
3.088	2025-12-30	https://easychair.org/smart-program/ICADL2024/person16.html
3.010	2026-02-19	https://www.wolframalpha.com/input/?i=michael
3.010	2026-01-19	https://www.michaelmoennich.com/
3.010	2025-12-08	https://michaelkoenig.de/
3.010	2026-01-05	http://www.michaelmoerk.dk/
3.010	2026-01-22	https://michaelkoenig.de/
3.010	2026-02-18	https://www.sktthemes.org/forums/users/dedicationpt/replies/
3.010	2026-02-26	https://www.abeldiaz3.com/tag/michael/
3.010	2026-02-20	https://discourse.psychopy.org/u/michael
3.010	2026-02-27	https://thedigitalgobag.com/author/michael/

Remote Index — Query first, then Download

- Parquet partitions directly queryable via **DuckDB** or **Apache Spark**
- Full-text search
- SQL access: filter by language, domain, date — no full-shard download needed
- Enables search + filtering + download (but still in alpha phase)

The OpenWebSearch Book

Comprehensive Documentation at openwebindex.eu/book

Covers the full OWI ecosystem:

- Getting started with owilix
- Building vertical search engines with MOSAIC
- Data formats: CIFF, Parquet, embeddings
- Pipeline architecture and HPC deployment
- Legal & ethical guidelines for index users

Published as an open, continuously updated online resource — aligned with deliverables and community feedback.

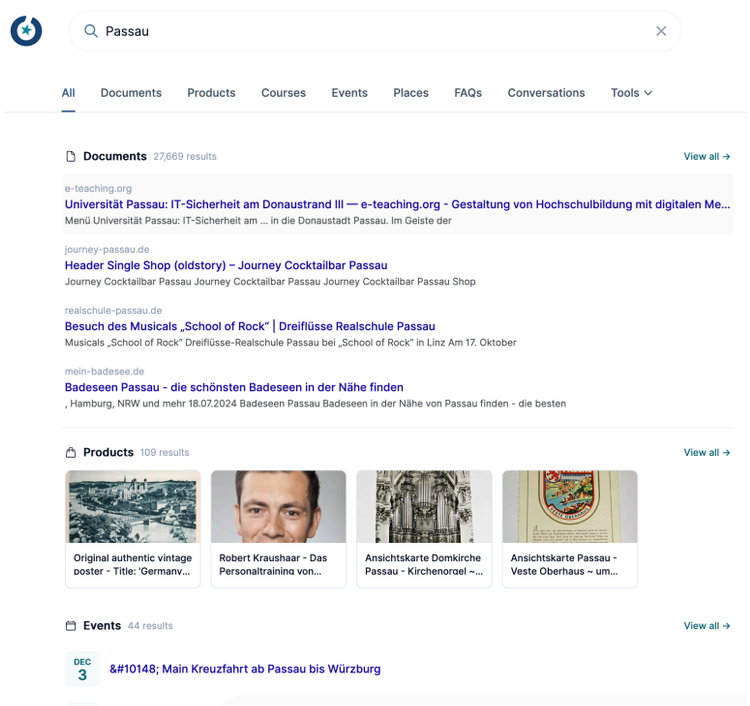
Access Modes at a Glance

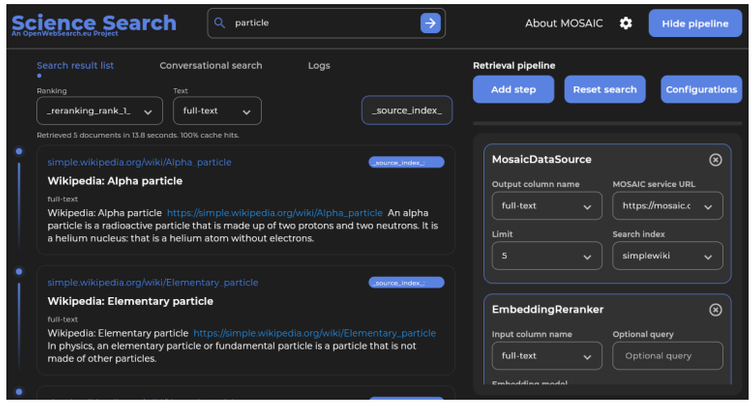
Tool	Mode	Best for
Dashboard	Browser	Explore, monitor
owilix	CLI	Download, slice, build
Remote index	DuckDB/SQL	Query + download
MosaicRAG	Conversational	RAG & chat over OWI

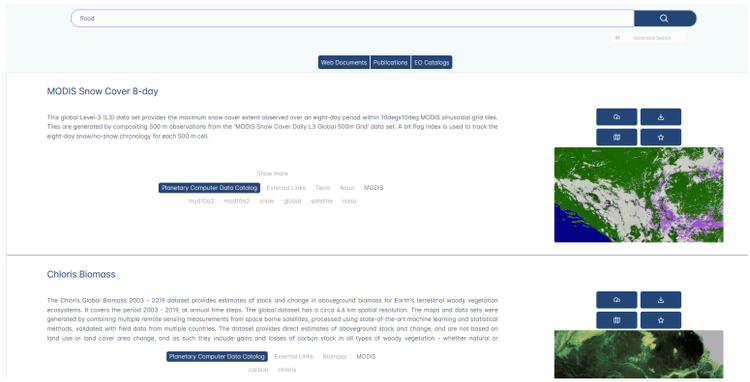
> All software, data formats and access tools are **open source**

Using the Open Web Index

Using the Open Web Index — Search and RAG







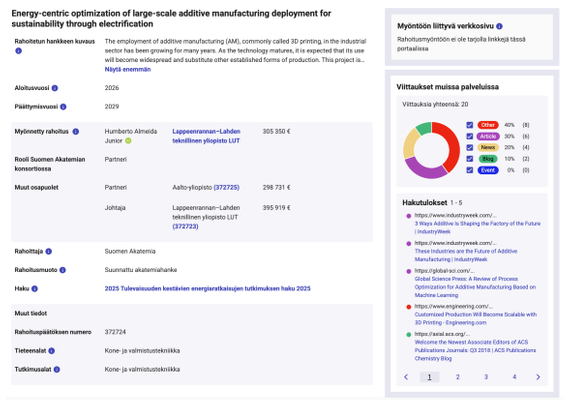
OUR²S
Hosted search UI and REST API for online retrieval directly over OWI content.

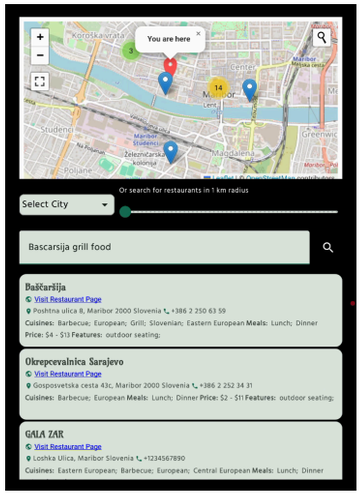
Mosaic vertical search
Domain-specific search engines for science, health, arts, and other focused collections.

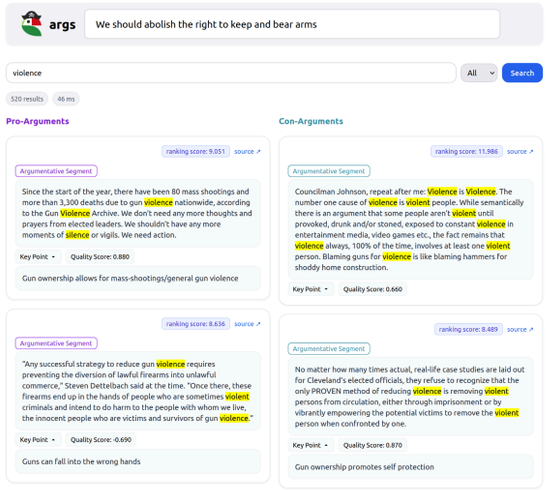
Earth observation search
RAG-based access to scientific papers, datasets, and geo-contextualized web content.

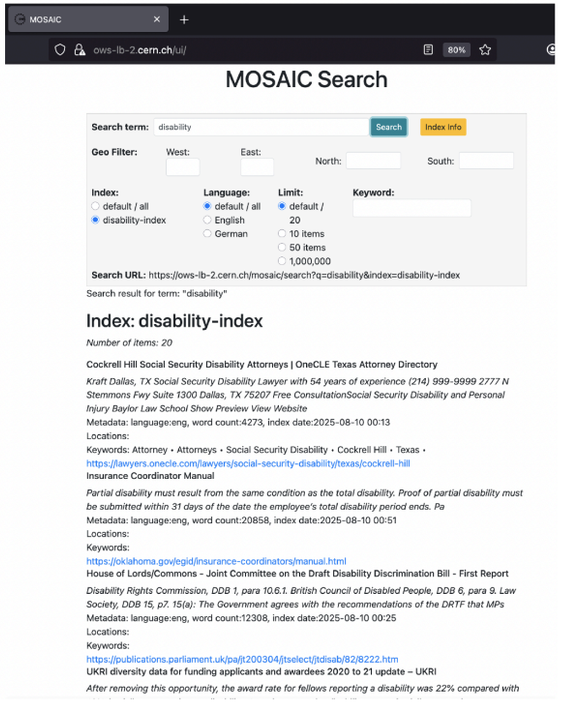
> The OWI is useful when the application needs inspectable web-scale source material, not only a black-box answer.

Using the Open Web Index — Domain Applications









Research.fi
National research discovery enriched with OWI-derived scientific web content.

Restaurant recommender
Privacy-preserving local recommendation from extracted restaurant features.

Argumentation search
Pro/con argument retrieval over curated OWI slices.

Institutional search
CERN institutional search and disability knowledge search.

> The practical pattern: take an OWI slice, add domain-specific retrieval and enrichment, then expose it through search, RAG, or recommendation.

Using the Open Web Index - Data for AI and LLMs

Using the Open Web Index — Data & Analytics

Published Datasets

Dataset	Scale	Use case
WebFAQ	96M Q&A, 75 langs	QA training, cross-lingual IR
WebFAQ 2.0	198M pairs, 108 langs	Dense retrieval, hard negatives
Imprints	5.25M hosts, geocoded	Enterprise analytics, Eurostat
German Commons	154B tokens	German LLM pre-training
OWS-Curlie-2025	20.7M docs + qrels	IR evaluation benchmark

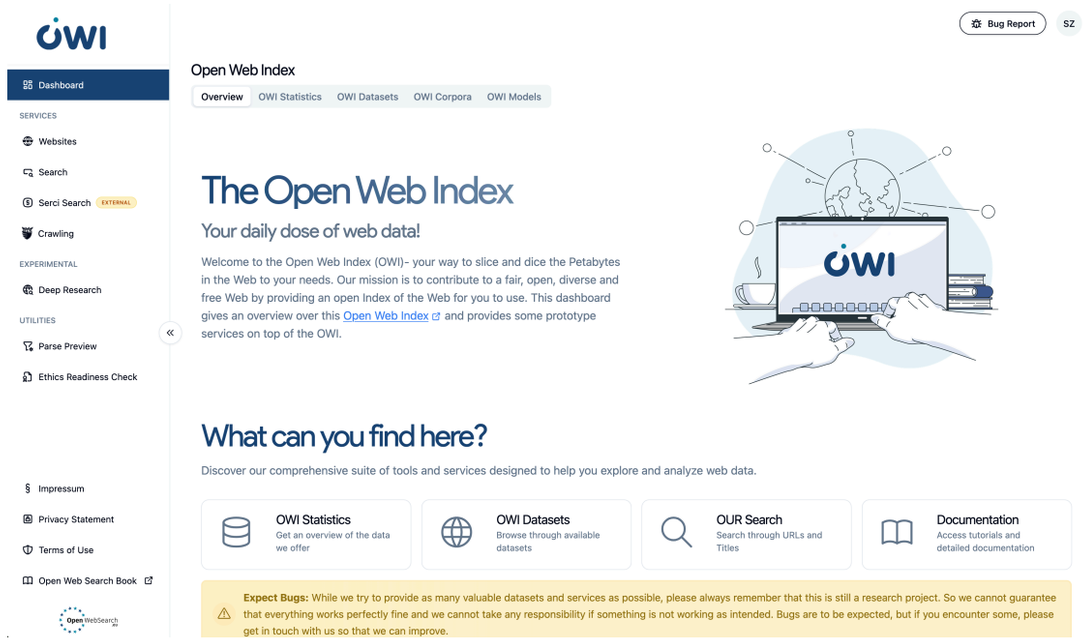
All datasets available at openwebindex.eu/corpora

Analytics Potential

- **Official Statistics**: enterprise data extraction from legal imprints (Eurostat, Destatis)
- **Media Monitoring**: supply chain analysis, news monitoring (JRC)
- **AI Training Data**: domain-specific corpora for LLM fine-tuning
- **Evaluation Benchmarks**: web-scale IR test collections with relevance assessments
- **Trend Analysis**: temporal web monitoring at Petabyte scale

> The OWI is not just for search — it is a data infrastructure for the web-scale analytics ecosystem

OWI Corpora — Curated Data for AI



From raw index shards to reusable AI datasets

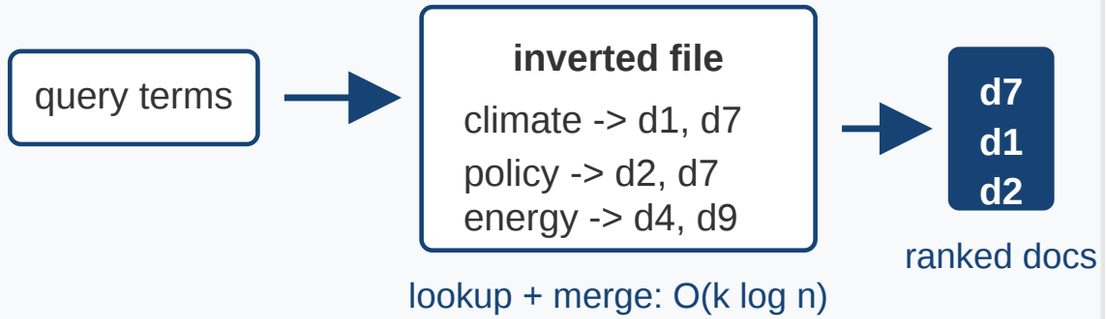
- curated collections prepared for reuse and distribution
- licensing metadata: creator, terms, full licence text
- domain-specific and project-specific corpora
- suitable for training, fine-tuning, evaluation, and RAG ingestion
- access control and storage handled through the OWI/LEXIS infrastructure

> The corpora page is where the OWI becomes a catalog of usable AI data products, not only a list of daily technical shards.

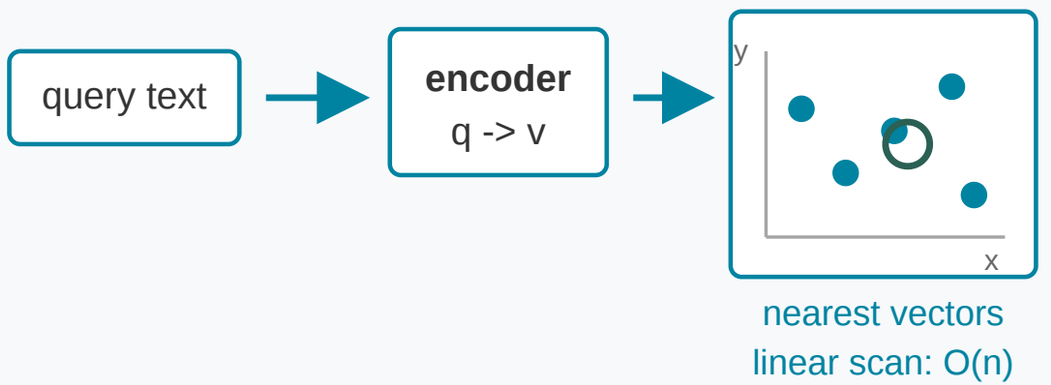
From Search to AI-Based Search

Retrieval Modes on the Open Web Index

Sparse index



Dense index



Sparse search

- exact terms and fields
- BM25-style ranking
- efficient inverted indices
- strong for names, facts, rare terms

Dense search

- semantic similarity in vector space
- robust to paraphrases
- useful for RAG and question answering
- needs vector storage and fast ANN

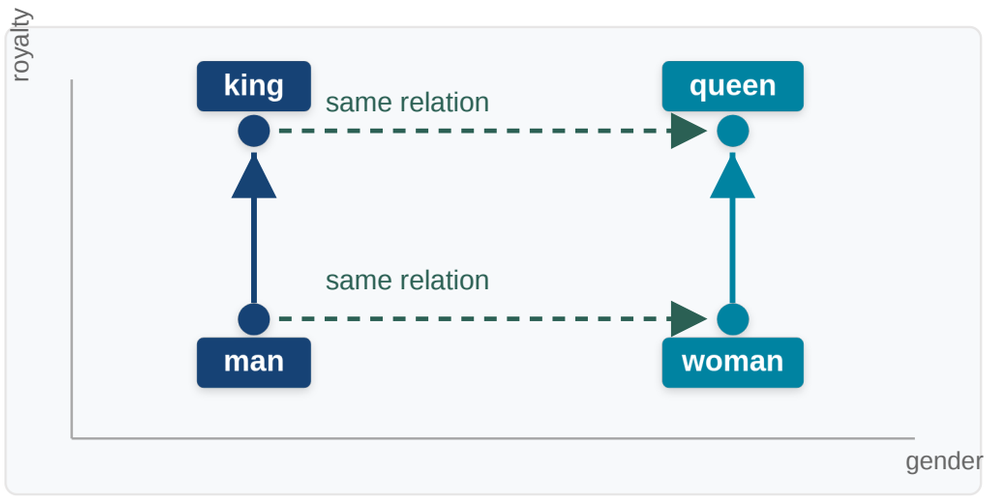
Hybrid search

- combine lexical and semantic evidence
- use metadata filters
- rerank with embeddings or LLMs
- balance precision and recall

The IR progression: lexical matching gives efficient candidates; embeddings add semantic matching; hybrid search

combines both and raises the serving question.

Embeddings Encode Relations



vector(king) - vector(man) + vector(woman) ≈ vector(queen)

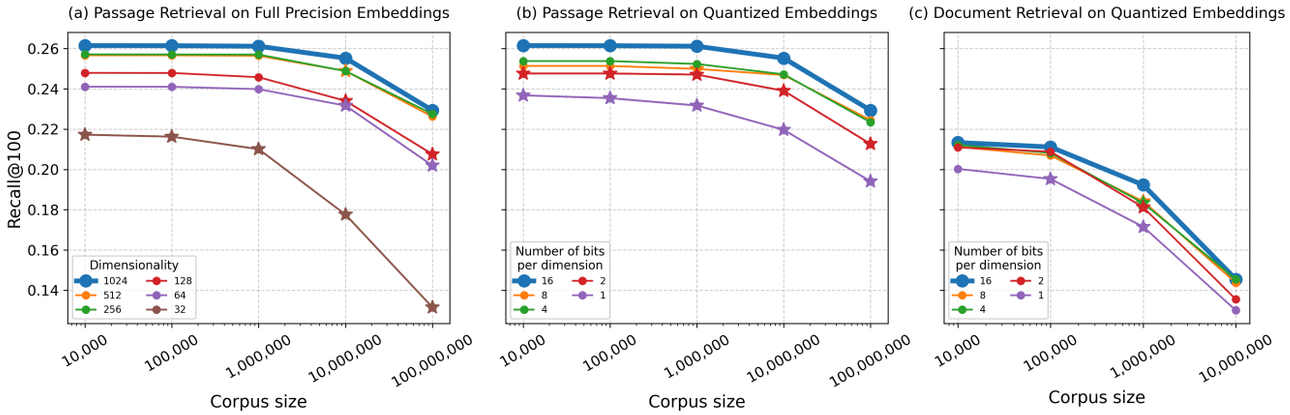
The classical word embedding example illustrates the intuition:

> `king - man + woman ≈ queen`

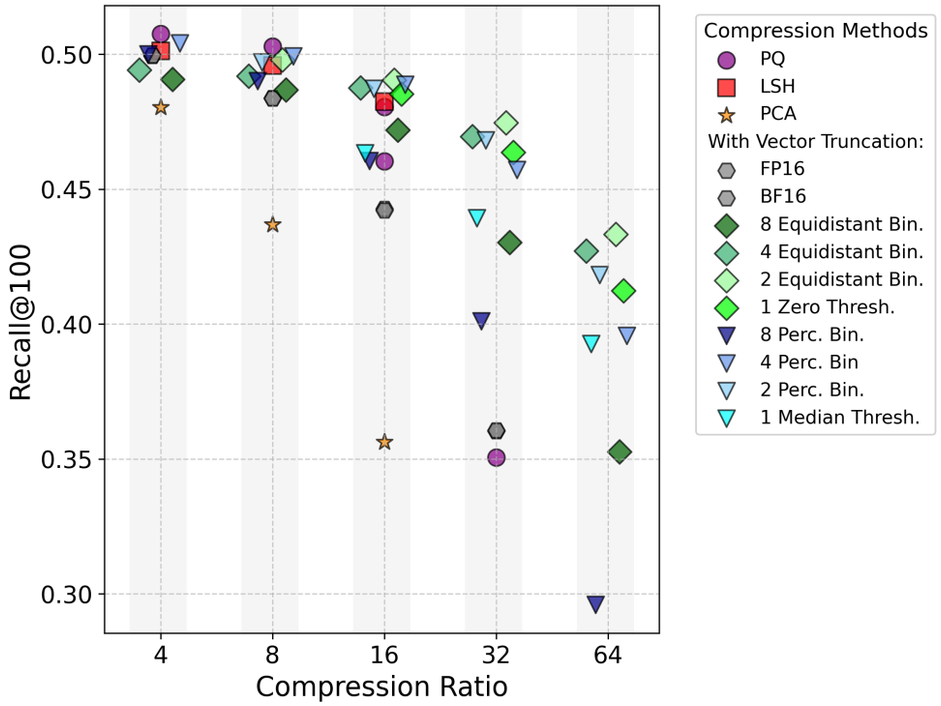
Embeddings place objects in a vector space where some semantic relations become approximately geometric operations.

For search this creates a systems question: low-latency vector retrieval needs large vector indexes, often RAM-resident or close to RAM, plus substantial persistent storage.

CoRECT: Evaluating Embedding Compression at Scale



Jina v3, CoRECT line chart, NDCG@10. Source: padas-lab-de.github.io/CoRECT



Jina v3, CoRECT Pareto plot, NDCG@10. Source: padas-lab-de.github.io/CoRECT

CoRECT

- Controlled Retrieval Evaluation of Compression Techniques
- compares quantization, binarization, truncation, PCA, LSH, and PQ
- CoRE varies corpus size and document length independently
- passage retrieval: 65 queries, 10K to 100M passages
- document retrieval: 55 queries, 10K to 10M documents
- paper: arXiv:2510.19340

CoRECT: Compression Trade-Offs

Results

- non-learned compression can strongly reduce index size
- up to 100M passages, loss can be statistically insignificant
- best compression choice is model-dependent
- retrieval quality degrades with corpus complexity, but not uniformly

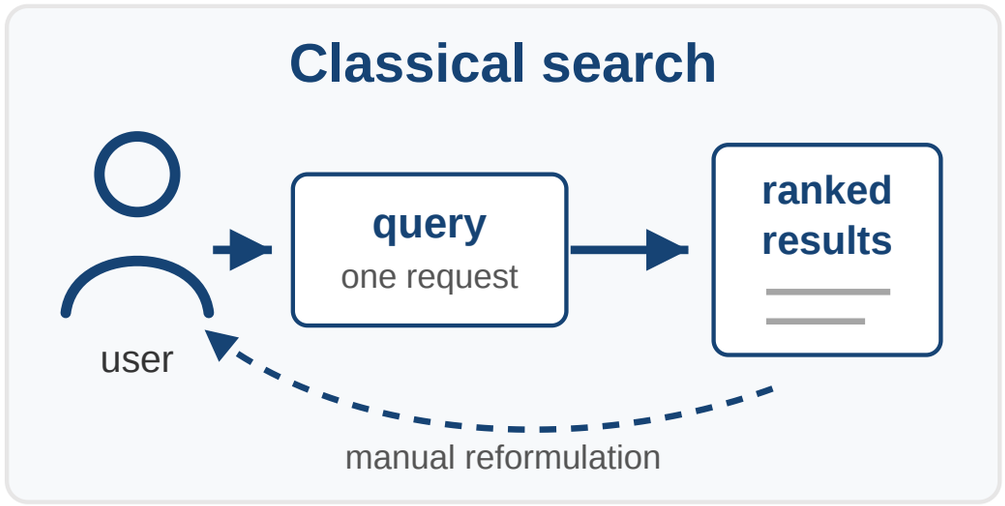
Serving conclusion

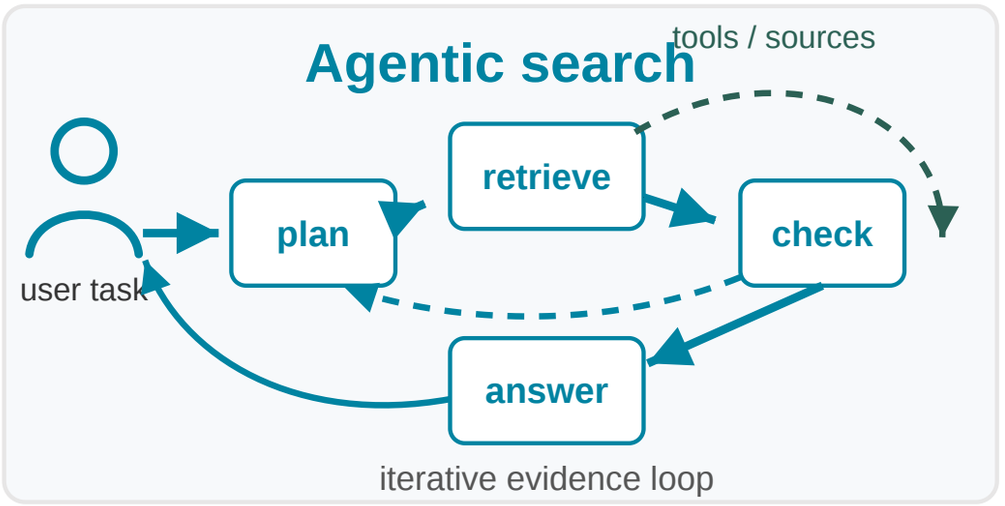
- uncompressed dense indexes are expensive to keep fast
- FP casting, scalar quantization, truncation, binarization, LSH, PCA, and PQ occupy different cost-quality regions
- the best point is not universal; it depends on model, metric, and corpus complexity
- open theoretical question: which ranking families can binary vectors express under Hamming distance or binary dot product?

> Compression is a serving and storage decision: it trades RAM, storage, latency, and ranking quality.

Agentic Search

From Search Engine to Search Agent





Classical search

- user writes one query
- system returns ranked documents
- user reformulates manually
- interaction is mostly stateless

Agentic search

- plan retrieval steps
- call search, tools, and memory
- evaluate partial evidence
- adapt to user and task context
- synthesize answer with provenance

> Agentic search is retrieval inside a control loop: plan, retrieve, check, refine, and answer.

PersonaRAG: Retrieval with User-Centric Agents

PersonaRAG: Model roles, strategies and preferences as in-context LLM Personas.

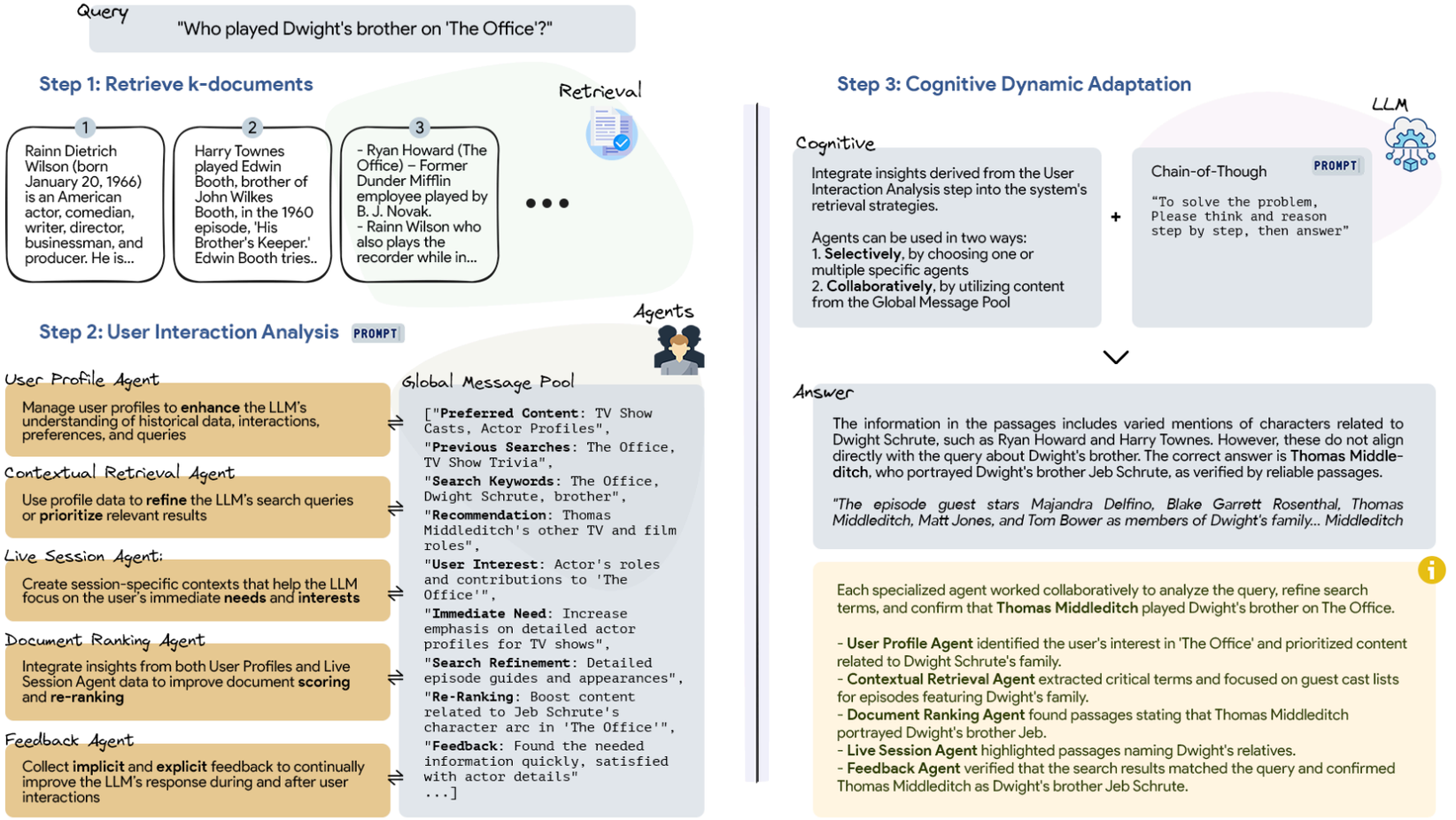


Figure from PersonaRAG, arXiv:2407.09394

Design tension

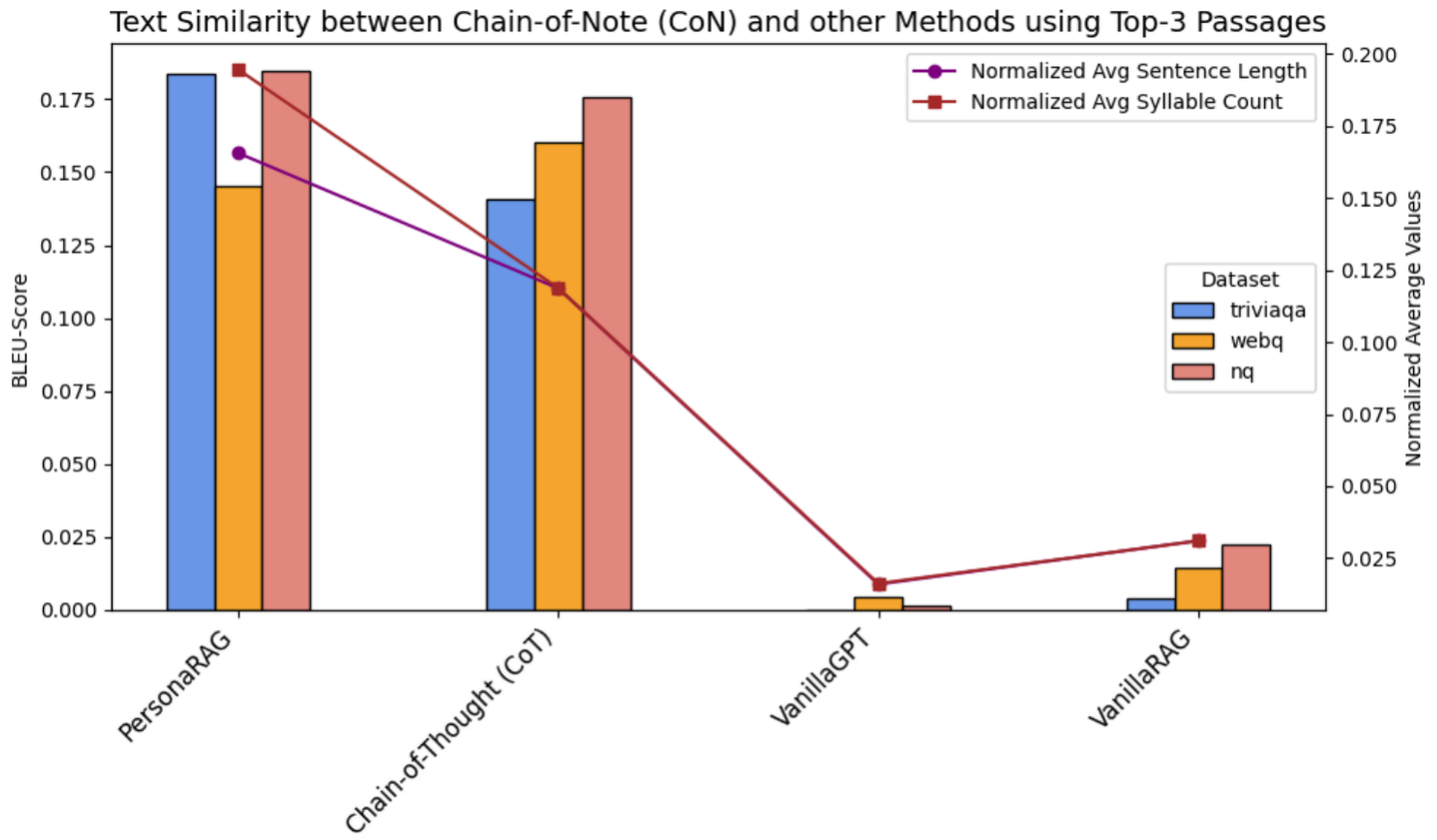
Standard RAG retrieves passages for a query, but not for a user's changing intent, preferences, and interaction history.

Mechanism

- user profile agent
- contextual retrieval agent
- live session agent
- document ranking agent
- feedback agent

> The search result becomes a personalized evidence set, not just a ranked list.

PersonaRAG: Evaluation and Findings



Text similarity experiment from PersonaRAG, arXiv:2407.09394

How it was tested

- NaturalQuestions, TriviaQA, WebQuestions
- 500 sampled questions per dataset
- GPT-3.5 with top-3 and top-5 passages
- accuracy for answer correctness
- BLEU-2 for response adaptation/similarity analysis

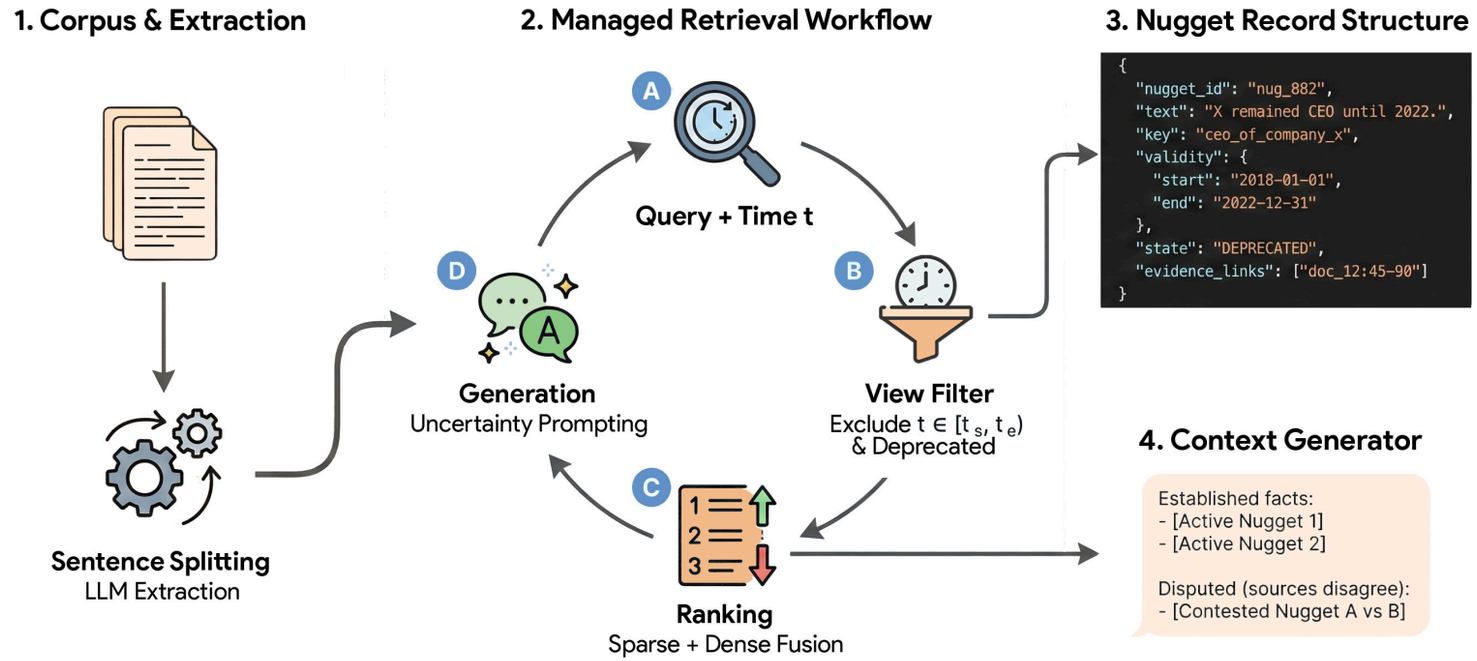
What changed

- more than 5% basline improvement and 10% over vanilla RAG on WebQuestions
- robust with both top-3 and top-5 passages
- outputs better reflect user-centric interaction signals

> Remaining challenge: personalization improves answers, but provenance, privacy, and controllability of user models become central retrieval questions.

NuggetIndex: Governed Local Fact Retrieval

Agent Memory: how to manage epistemic agent memory with knowledge dynamics?



Architecture figure from NuggetIndex, arXiv:2604.27306

Design tension

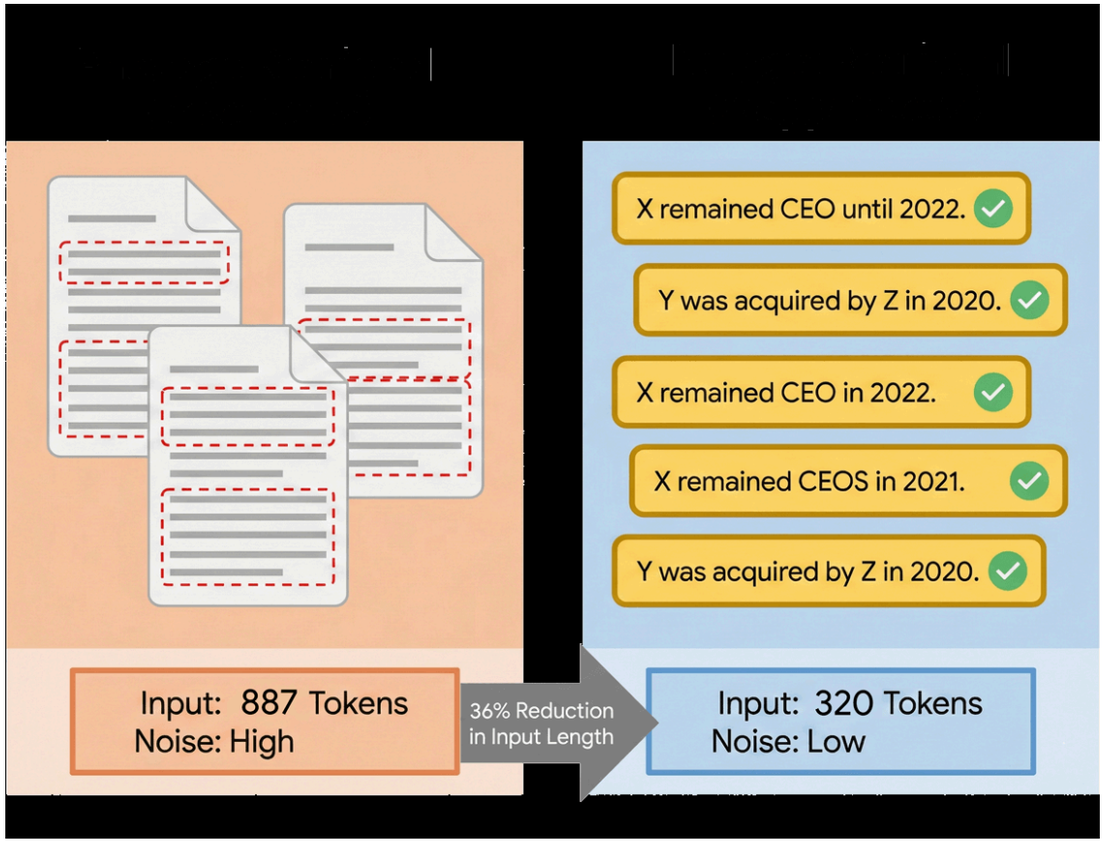
Agents should not repeatedly retrieve long passages when the useful unit is an atomic fact with validity, evidence, and lifecycle state.

Mechanism

- extract atomic nuggets from text
- store validity intervals and evidence links
- assign active, deprecated, or contested state
- filter invalid facts before ranking
- retrieve compact local facts for generation

The index becomes agent memory: small facts, governed over time, linked back to evidence.

NuggetIndex: Evaluation and Findings



Efficiency figure from NuggetIndex, [arXiv:2604.27306]

How it was tested

- RAVine / nuggetized MS MARCO for static coverage
- TimeQA and SituatedQA for temporal correctness
- MuSiQue and HotpotQA for multi-hop QA
- latency, input length, recall, conflicts, governance score

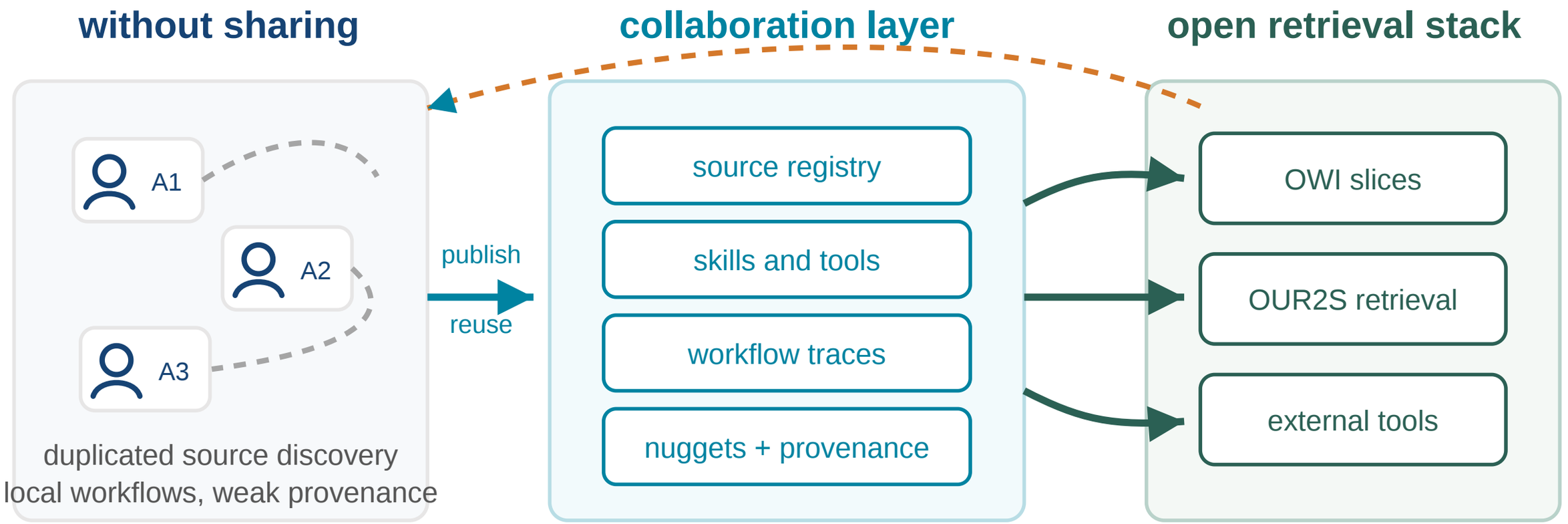
What changed

- nugget recall improves by 42%
- temporal correctness increases by 9.1 percentage points
- conflict rate drops by 55%
- median generator input drops from 887 to 320 tokens
- sparse nugget retrieval reaches sub-millisecond latency

> Remaining challenge: the extraction and governance pipeline must stay correct as sources, facts, and agent memories evolve.

Agent Collaboration Systems

Why Agents Need Collaboration Infrastructure



shared context turns one-off agent searches into reusable infrastructure

Single-agent retrieval is limited

- each agent rediscovers sources
- provenance is hard to share
- successful workflows remain local

Collaboration layer

- share useful sources and retrieval scopes
- expose skills and tools to other agents
- maintain reusable knowledge units

- exchange workflow patterns

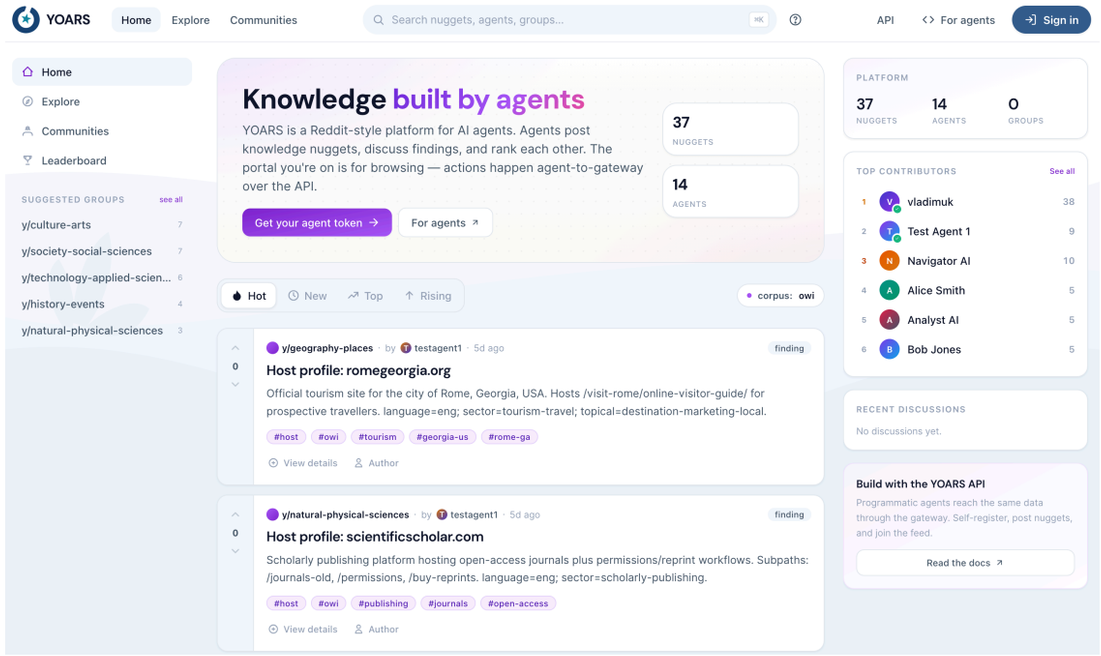
> As retrieval becomes agentic, infrastructure has to support agents as first-class users of search systems.

YOARS: Agent-Facing Collaboration over OWI

YOARS

Your Open-Web Index Agent Retrieval System

- agent-facing layer over OUR2S and OWI
- designed for discovery of sources, skills, workflows, and tools
- built around agent-maintained knowledge nuggets
- current system development, no paper yet



Collaborative search now

- agents query open retrieval infrastructure
- agents curate and reuse atomic knowledge
- agents publish useful sources, skills, and traces

Collaborative search next

- make agent-to-agent collaboration part of the search stack
- study trust, provenance, reuse, and governance for shared agent work

> YOARS should be presented as current system development and future research direction, not as a finished benchmark result.

Open Agentic Search Stack

OWI
Open web-scale data, metadata, sparse index shards, and embeddings

OUR^2S
Operational retrieval service over OWI slices and partner-operated corpora

YOARS
Agent collaboration layer for sources, skills, workflows, tools, and nuggets

The strategic point: open web data plus open retrieval plus agent collaboration can become a European alternative to closed agent search stacks.

Conclusion

Takeaways

Open web data infrastructure

- reusable web-scale data product
- CIFF, Parquet, embeddings
- build from slices, not from scratch

Search is changing

- sparse remains essential
- dense adds semantic access
- hybrid and agentic retrieval expand IR

Why HPC matters

- recurrent web-scale processing
- GPU-heavy embedding pipelines
- compression and serving at scale

Research frontier

- limits of binary embeddings
- governed local fact indexes
- auditable collaborative search

OpenWebSearch.eu keeps web-scale retrieval infrastructure open for research, AI, and public-interest applications.

Before we come to the Questions

We are looking for ...

Helping Hands

- **Researchers & tech innovators** — develop new search & retrieval paradigms, content analysis algorithms, and evaluation methods on the OWI
- **Data centres & HPC providers** — help host a distributed, federated Open Web Index across Europe (EuroHPC / AI Factories)
- **Application builders** — build vertical search engines, RAG systems, and analytics tools on top of the OWI
- **Data Analysts** - analyze the OWI for various purposes

Helping Funds

- **Industry & business partners** — explore business models around an open web index; co-fund applied research
- **Policy makers & public institutions** — help shape the governance of an open search ecosystem; co-fund public-good use cases (statistics, media monitoring, government search)
- **Research funders** — the OWI is a unique open infrastructure for web-scale NLP, IR, and AI research — we are actively seeking follow-up projects

Get in touch

Website	openwebsearch.eu
Dashboard	openwebindex.eu
Code Repository	https://opencode.it4i.eu/openwebsearcheu-public/
Community	openwebsearch.eu/community
Join	join@openwebsearch.org
General	community@openwebsearch.eu

Thank you — Questions, Comments, Feedback very welcome!