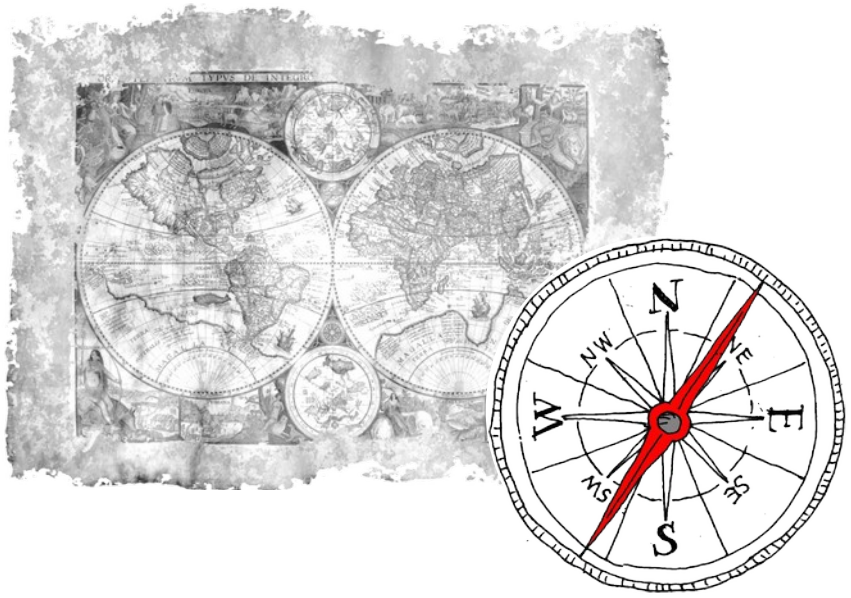

Open WebSearch
Towards an open and federated European Web Data Infrastructure for Search, Analytics and AI
How Europe can regain Sovereignty in the Web
Stefan Voigt, Open Search Foundation / DLR
Open Web Search
Funded by the European Union
Brno, 21. May 2025
SUPPORTED BY
NGI

Orientation = Power + Control
(economic, strategic, geographic, cognitive, political …)

Orientation in the geographic sphere used to be exclusive knowledge and a tool of power



This is still the case in the digital sphere

Illustration: www.atelier-anne-rieken.de

→ Europe needs a Programme like „Galileo" or „Copernicus" for sovereignty in Web Search and Web Data Services

Why does Europe need an independent “Navigation System” for the web?

➜A web index is the key element of web search.

➜Globally, only four comprehensive indices exist. They are either purely commercially driven or state-controlled.

➜Europe does not have its own web index. More than 90% of all web search is done via Google.

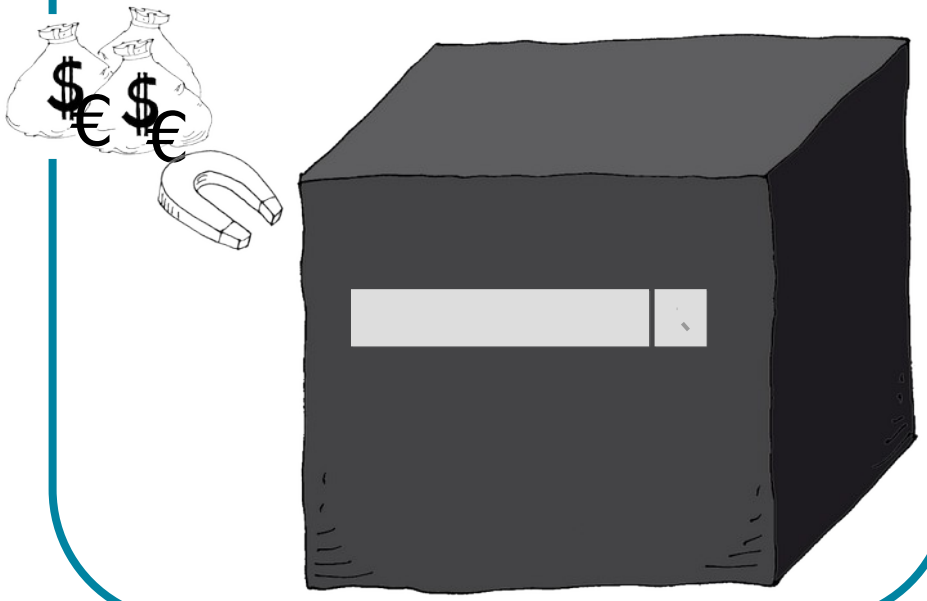
➜Europe depends completely on US-American search/webservice providers and their commercial interests.

➜An Open Web Index for Europe will

- strengthen the strategic sovereignty and technological autonomy through an independent and transparent web access and
- essentially contribute to the European digital targets for 2030 by building a sustainable digital infrastructure

An Open Web Index will enable transparent and unbiased access to Web Content

From a closed and opaque internet search ...



... to an open, transparent and auditable search and webdata ecosystem

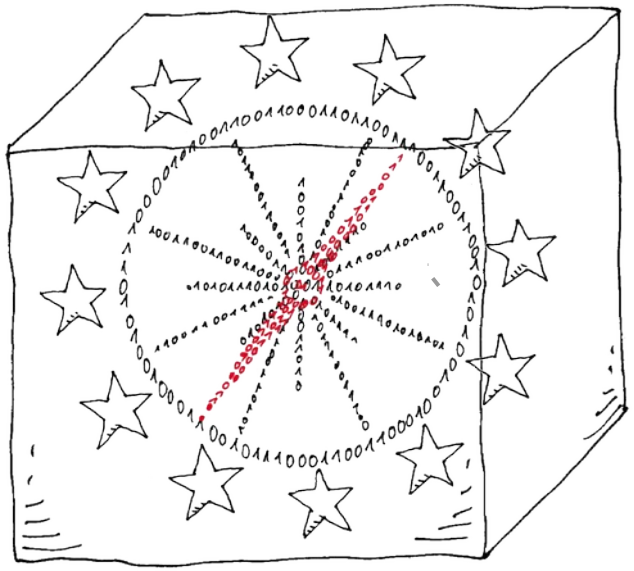


Illustration: www.atelier-anne-rieken.de

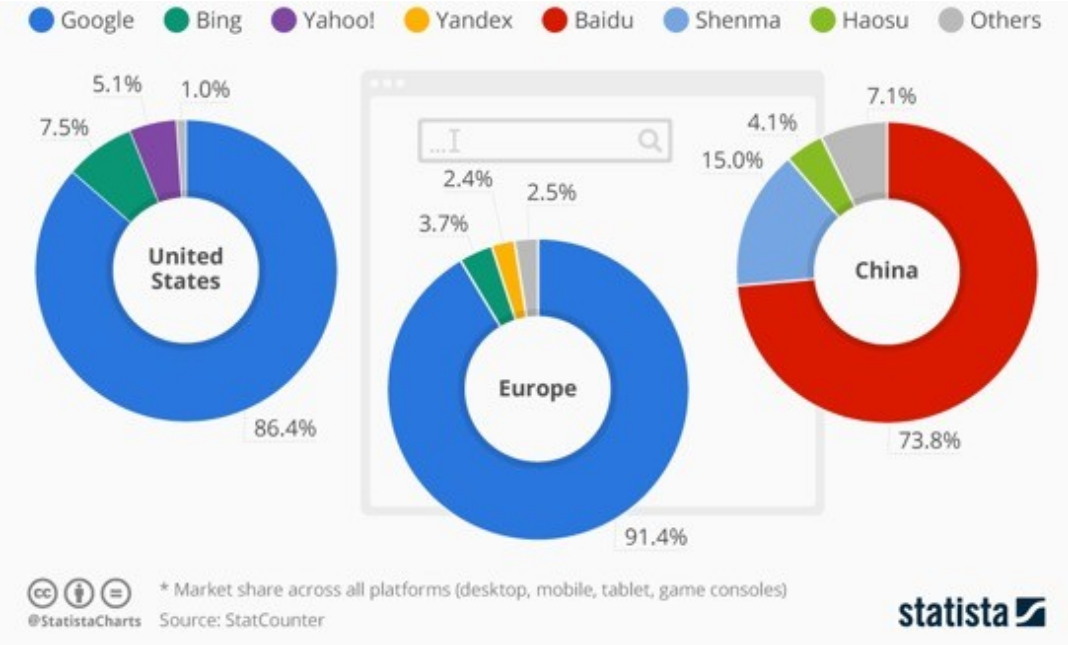
A Critical Infrastructure managed by an Oligopoly

Two current 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

- Missing Digital Sovereignty
- Reduced User Experience (limited choice, lock-in ...)
- Limited Innovation Potential
- No large scale access to Web-data as driver for AI Innovation and beyond



OpenWebSearch.EU's Proposition

Goal: Building an **Open Index of the Web** and a **federated Open European Web Data Infrastructure as a basis for a Web Search, Analytics and AI in Europe – in order to**

- empower Europe's researchers, innovators and businesses to systematically tap into the Web as scientific, business and innovation resource at scale – Petabyte-Scale!
- contribute to Europe's tech and digital sovereignty
- support Web-data analytics and AI / RAG systems across Europe
- build a federated infrastructure across existing European cloud, data and HPC centres

The piloting, currently funded by the EC (HE/NGI), GA:101070014, is carried out by 14 core partners plus 9 additional third party projects and a large ecosystem of early adopters and supporters

Current Core Partners

Open WebSearch .eu

Research

UNIVERSITÄT PASSAU

Webis.de

UNIVERSITÄT LEIPZIG

Bauhaus-Universität Weimar

Radboud Universiteit

DLR

TU Graz

Infrastructure

lrz

CERN

VSB TECHNICAL UNIVERSITY OF OSTRAVA | IT4INNOVATIONS NATIONAL SUPERCOMPUTING CENTER

CSC ICT Solutions for Brilliant Minds

NGOs

osf open search foundation

nlnet FOUNDATION

suma-ev

Businesses

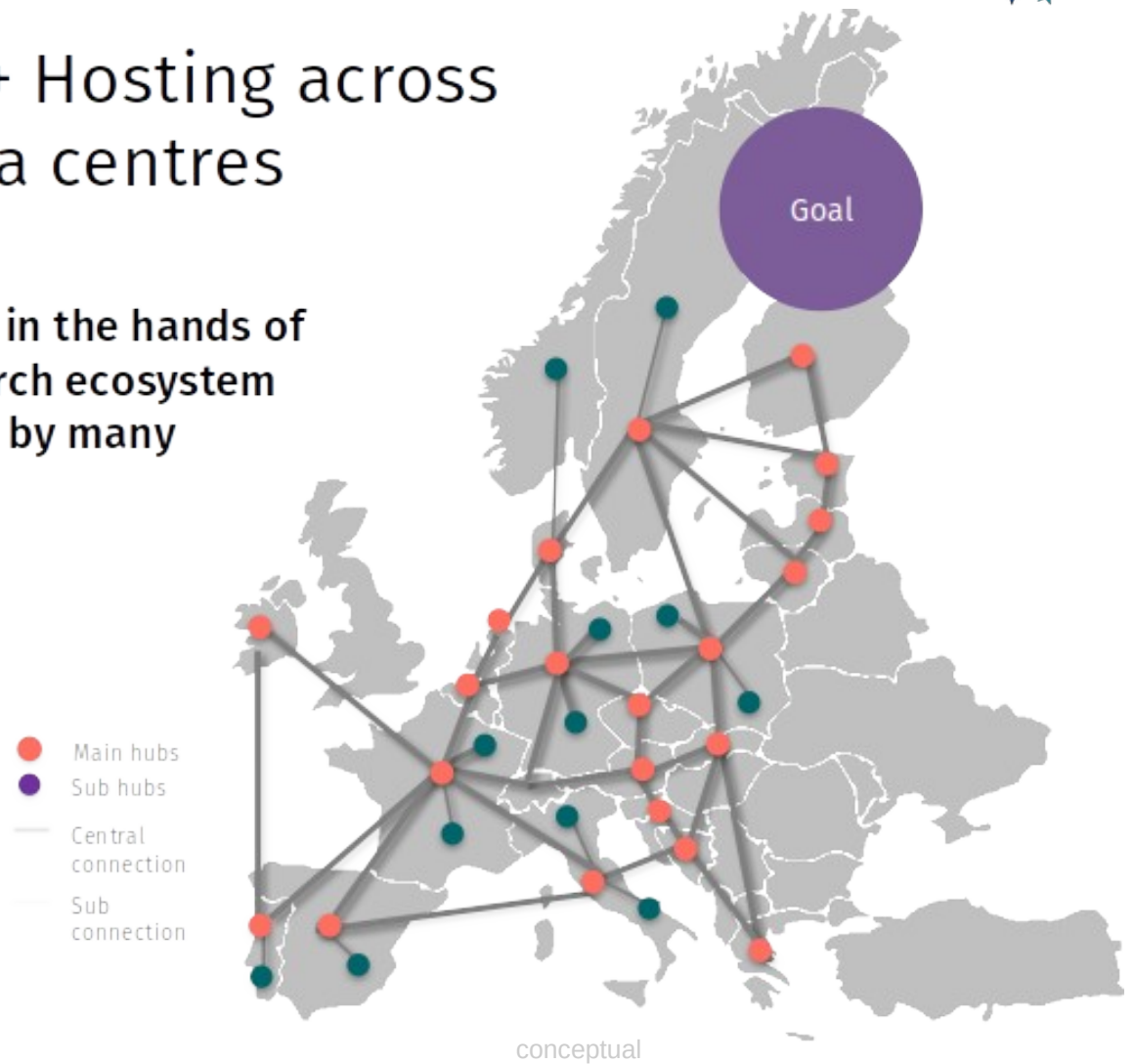
A1

A federated European Web Data Infrastructure enables a large variety of new public and private Web services, boosting innovation in Europe

Federated Computing + Hosting across European HPC and Data centres

From a centralized server landscape in the hands of one company to a decentralized search ecosystem that is shared and collaborated with by many

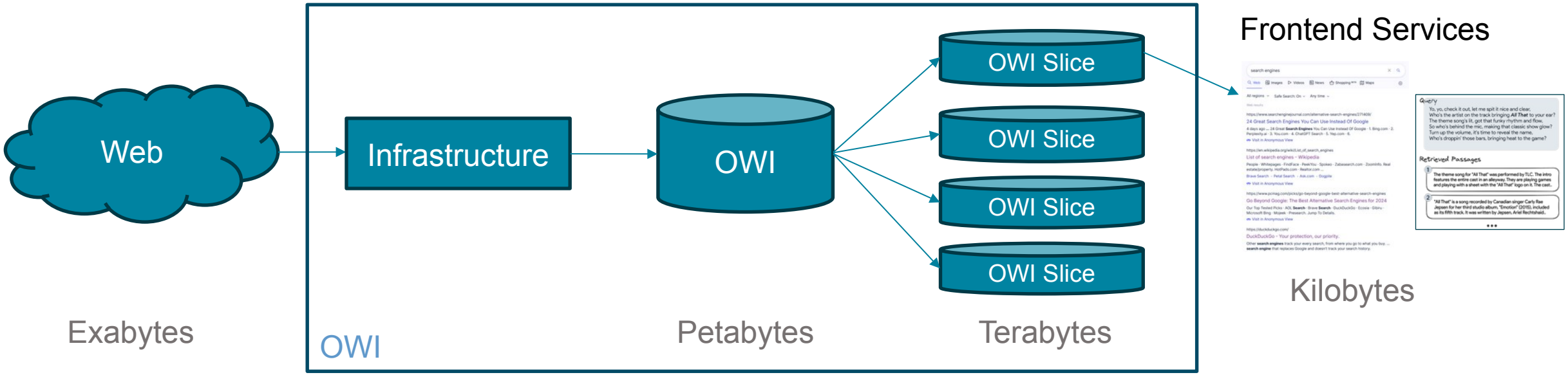
Together: existing data centers Faster, lower costs



The Open Web Data Infrastructure – a booster for Web Sovereignty in AI, Search, Analytics and more

A federated Web Index (OWI) and Web data infrastructure is a data infrastructure for fast query-based access and ranking of web documents at scale for a large variety of web-data-driven services

OpenWebSearch.EU: Piloting a collaboratively created, federated and transparent European Web Index for empowering scientists and innovators and creating an ecosystem for Web Search, Web-data Analytics and AI



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).

Core Elements of the Web Data Infrastructure and Index

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

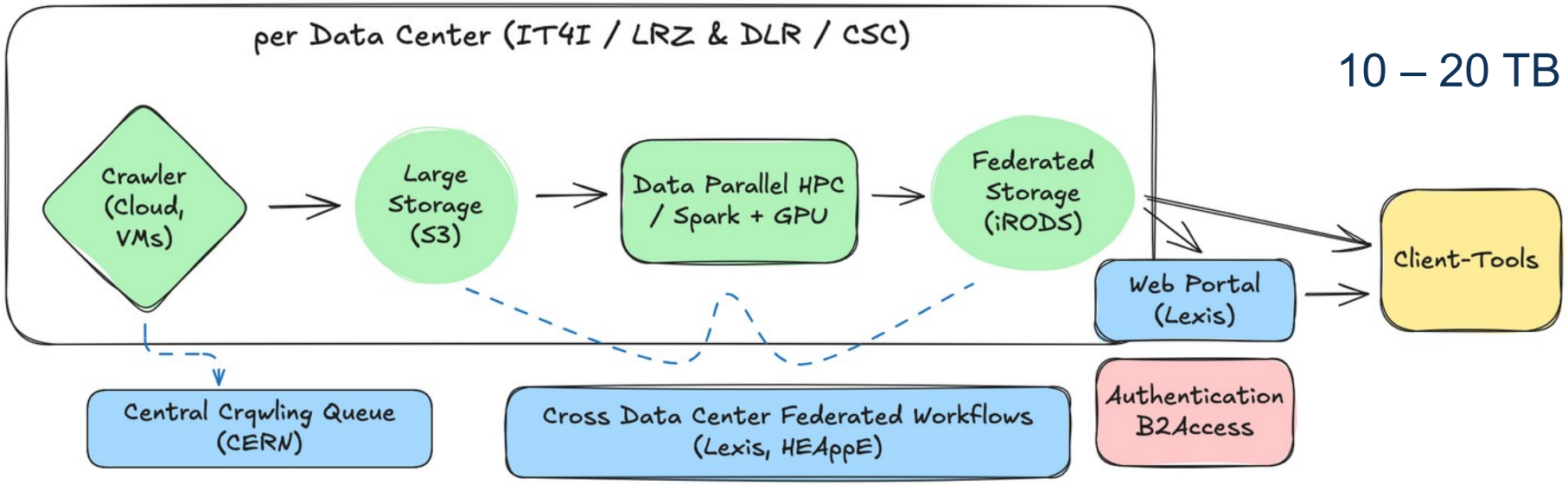
Applications and Innovations as Multiplicator

Distributed Infrastructure as Enabler

Our current OWI setup

18 VMs
100 Million URLs/day
100-200 TB per month (text only)

3-5 TB / Day / data center
IO + memory bound
GPU bound when including AI methods



10 – 20 TB index data / month

Cross Data Center Workflow Execution + Federated Data Storage

- approx. factor 20 for commercial indices
- Approx. factor 10-50 when including multimedia

Tech setup:
Open Web Crawler

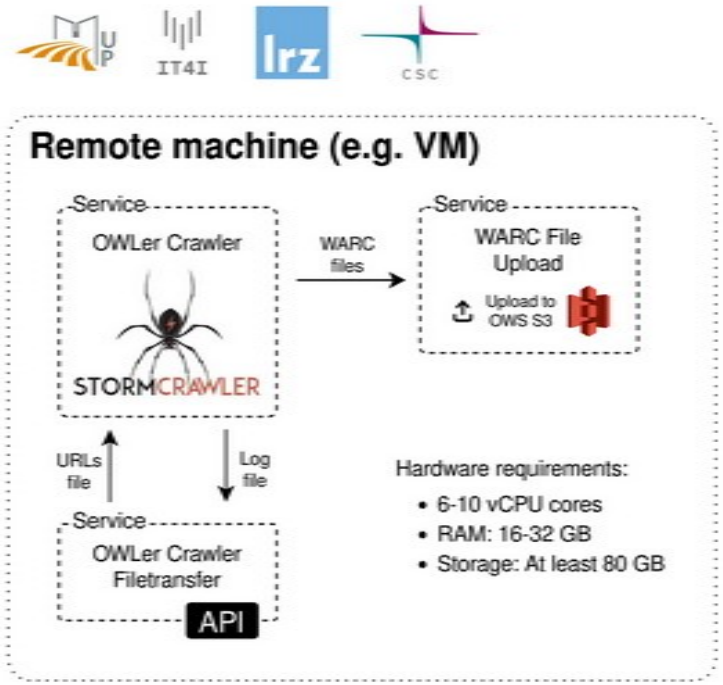
https://opencode.it4i.eu/openwebsearcheu-public/open-web-crawler

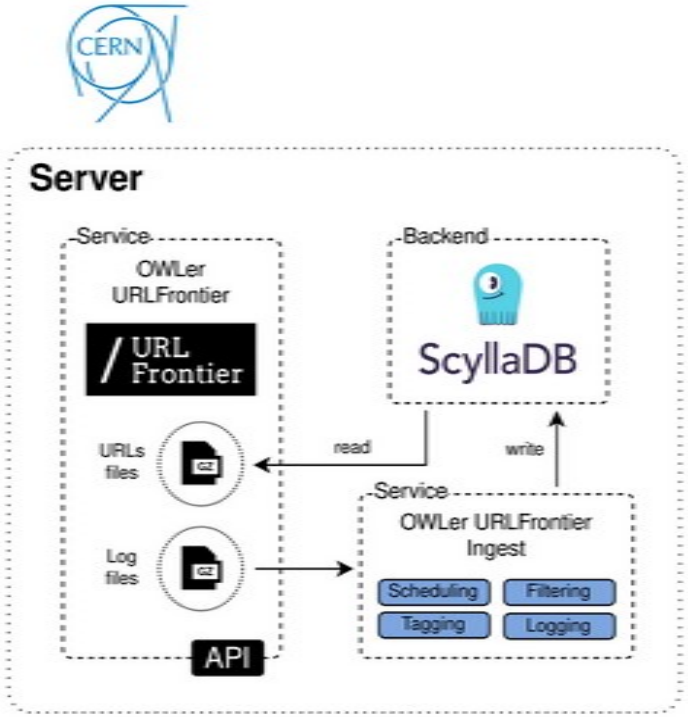
2 Servers (+ 2)

5 VMs

5 VMs (+ 5)

8 VMs (+ 2)





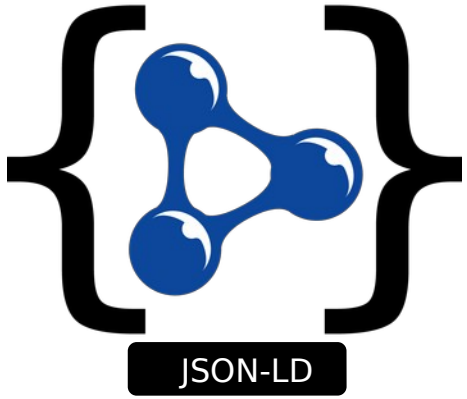
Collecting structured Meta Data
Microdata

Addresses

Phone numbers

FAQs

Opening hours

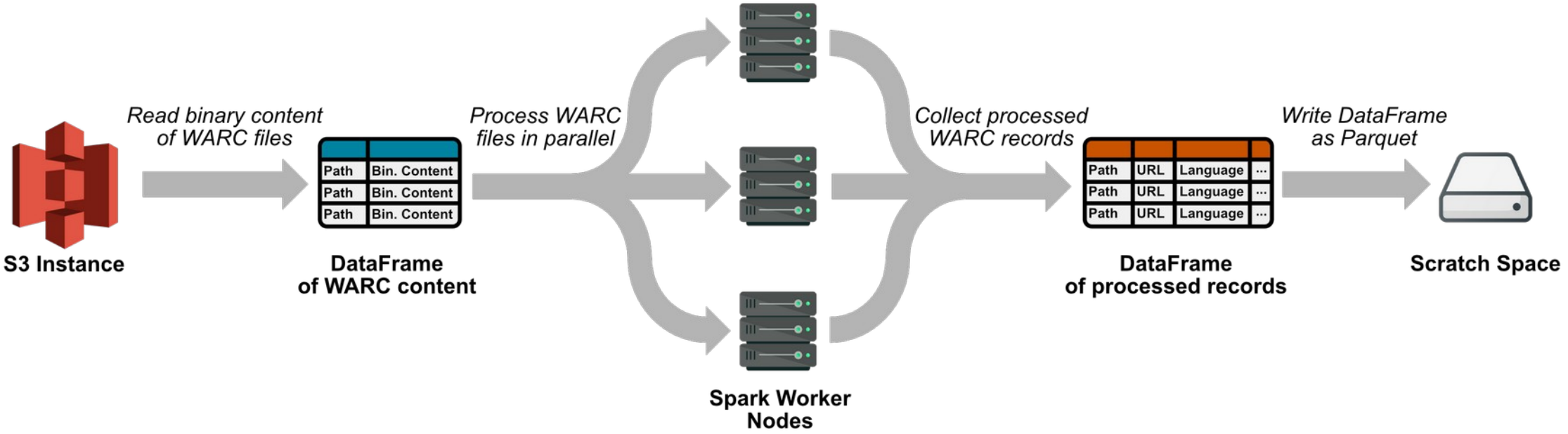



```
{
  ...
  "telephone": "(425) 614-3256",
  "address": {
    "@type": "PostalAddress",
    "addressCountry": "US",
    "addressLocality": "Bellevue King",
    "addressRegion": "WA",
    "postalCode": "98007",
    "streetAddress": "1410 156th Ave NE"
  },
  "openingHours": ["Mo 08:00-22:00", "Tu 08:00-22:00", "We
    08:00-22:00", "Th 08:00-22:00", "Fr 08:00-22:00", "Sa
    09:00-22:00", "Su 09:00-22:00"]
  "@type": "FAQPage",
  "mainEntity": [{
    "@type": "Question",
    "name": "How can I place a Subway Catering order?",
    "acceptedAnswer": {
      "@type": "Answer",
      "text": "To place an order, visit us online at
        catering.subway.com or call your local restaurant."
    }
    ...
  }]
  ...
}
```

All Webpages are processed on CPU-Queues

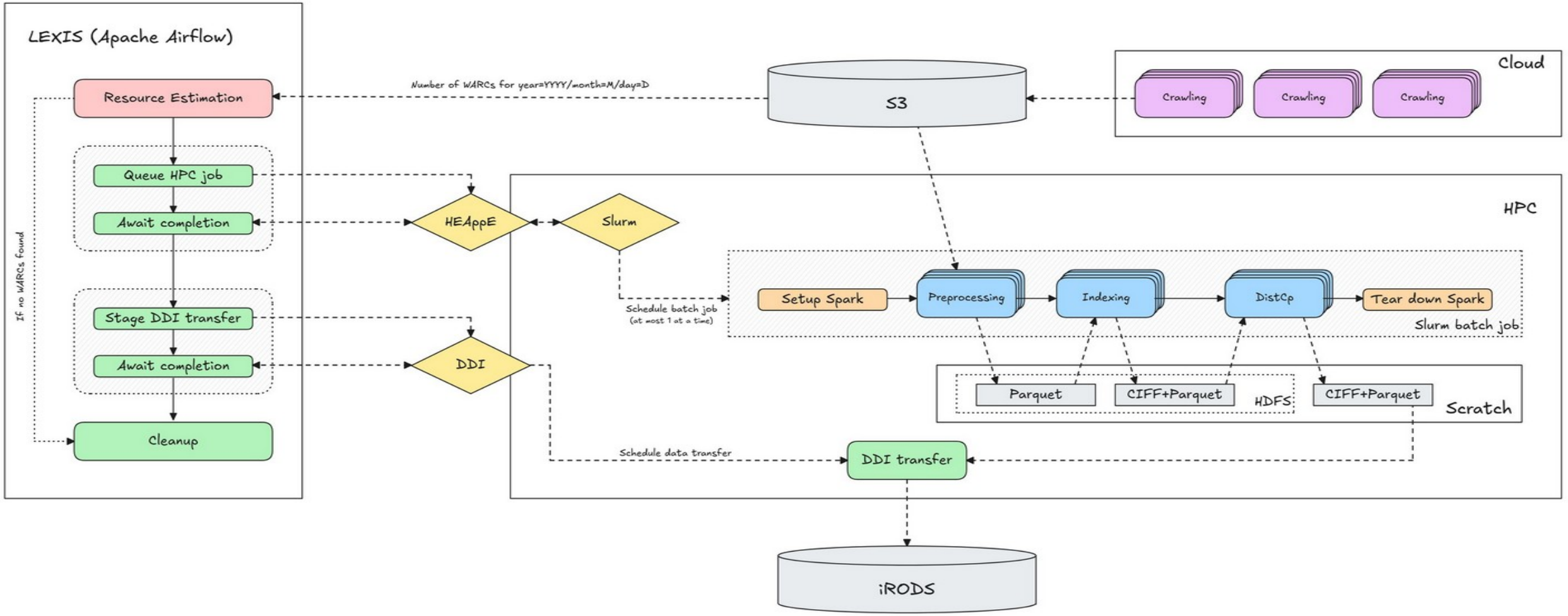
Resilipipe [OpenCode.it4i.eu]

- Extracts meta data from the daily crawls and saves it in Parquet files
- Runs on PySpark and uses Resiliparse to parse the WARC files
- Processes all crawled web pages with lightweight modules on CPU queues



Extension: Select subset of pages to be processed on GPUs

LEXIS workflows: architecture

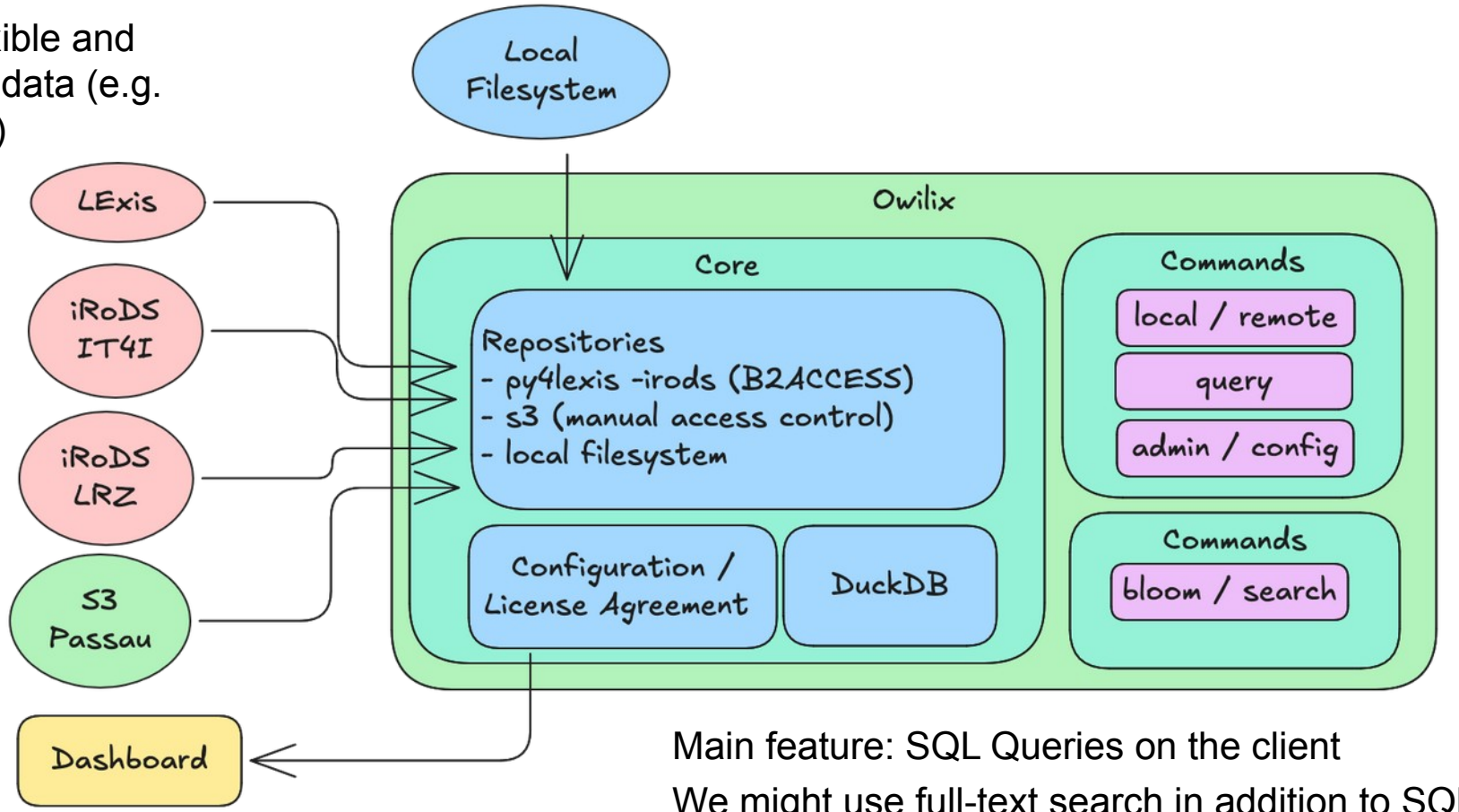


Owilix – The Open Web Index Client

Goal: Access, synchronize (pull/push/slice), query etc. index shards

Main purpose: pushing / pulling / slicing datasets

We can mage repositories flexible and mirror/ partition data (e.g. topic, language)



Flexible command structure

Plugins are somehow possible

Main feature: SQL Queries on the client

We might use full-text search in addition to SQL!

We might be able to integrate owilix as Web-App, i.e. Command Line in your browser

Open Web Index Dashboard => OpenWebIndex.EU

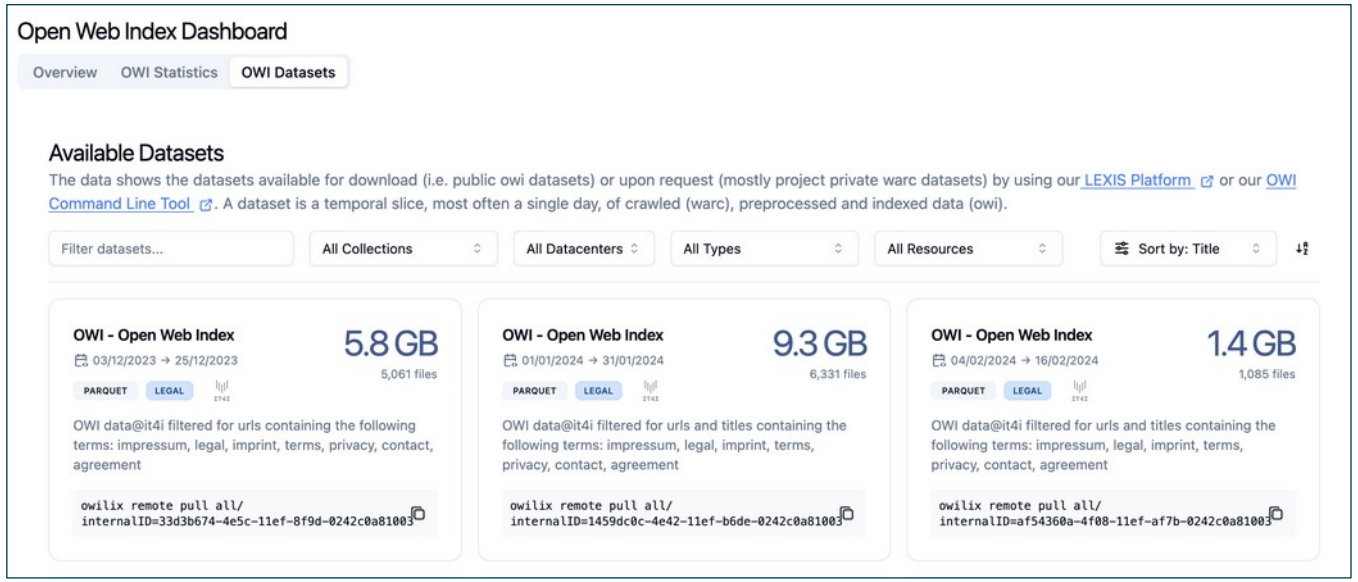
Open WebSearch .eu

Crawled Data
964.92 TB
+101.18TB last month
May 1, 2025 last updated

WARC Datasets
151
+4 last month
May 1, 2025 last updated

Public Datasets
493
+49 last month
N/A last updated

Open Web Index
24.94 TB
+2.1TB last month
N/A last updated



Pipelines
- Flexible, cross data center pipelines
- Daily index shards are available
- Tooling for index access still needs improvement
  - Currently working on search based data selection
  - New: query url lists

Index Details (V 0.2.0)
- Sources: Crawls, Mastodon, Wikipedia
- Collection indices: curlie, legal, main
- Features:
  - Plain text, url, id
  - Content-Metadata: json-ld, Microdata, opengraph, curlie-label(s), links, address.list (=geo microdata), language
  - HTTP-Metadata: http-server, crawler-source, charset, mimetype etc.
  - Process-Metadata: warc reference (file+offset), genai flag, index flag, canonical links

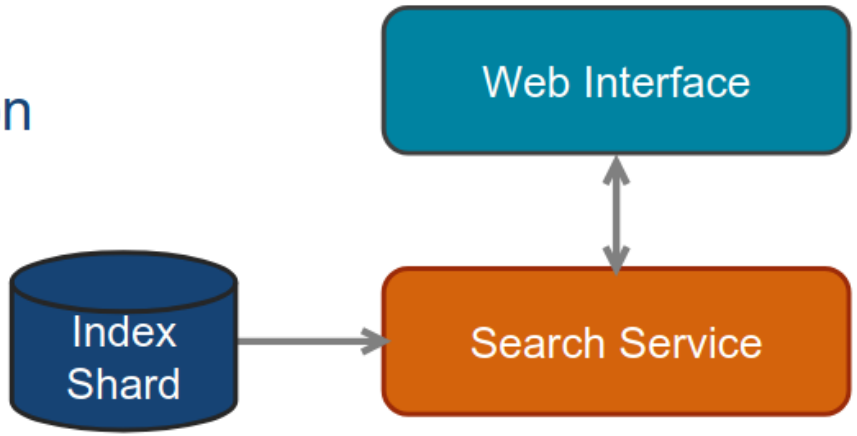
Index launch event on June 6

openwebindex.eu

Some applications built on the Open Web Index
Modular Search Application Builder (MOSAIC)

MOSAIC

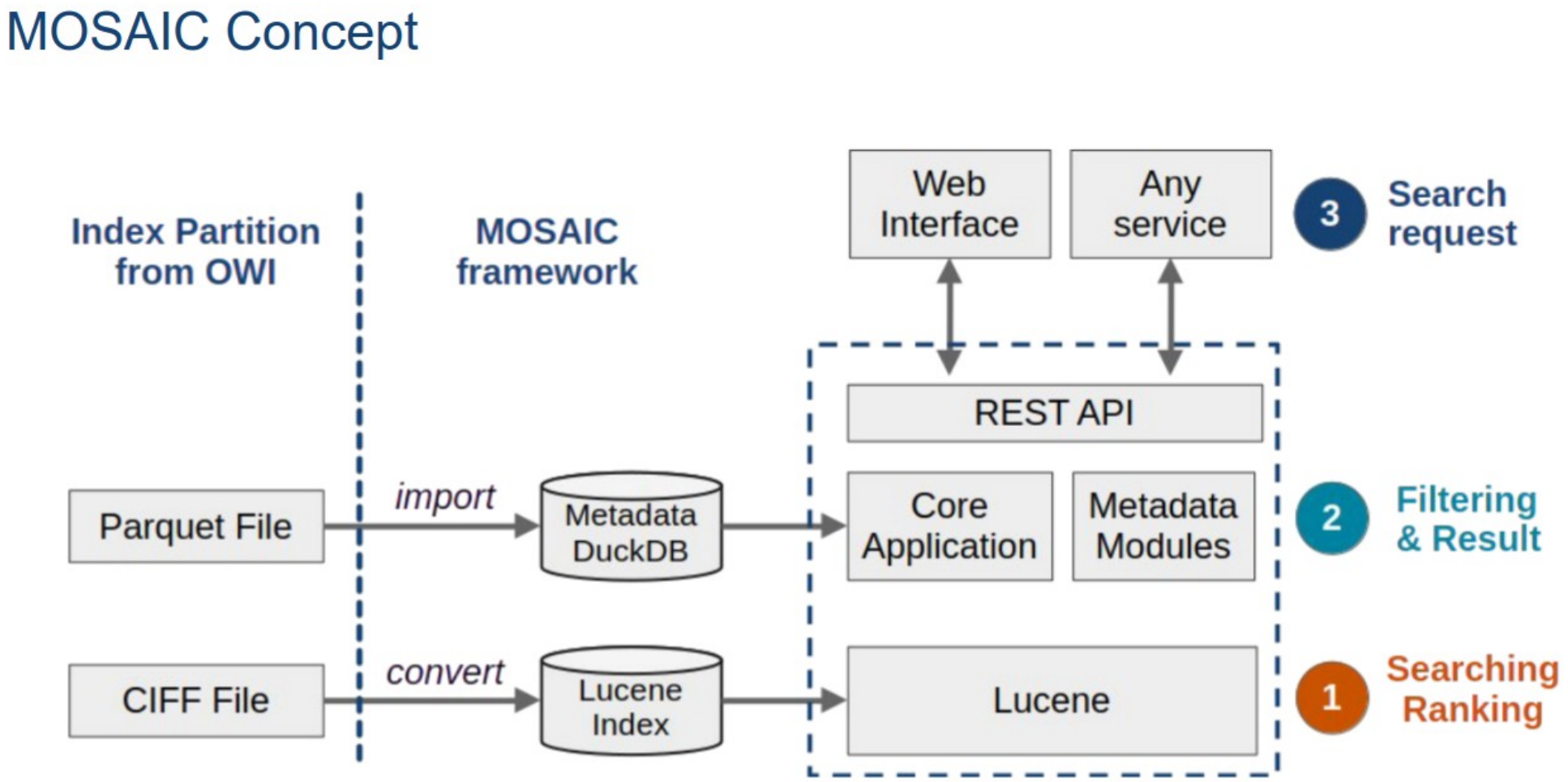
- **Mo**dular **S**earch **A**pplication based on **I**ndex Fra**c**tions
- Generic implementation of an OWS.eu vertical search engine
  - Demonstration of the concept of an OWS.eu vertical engine
  - Out-of-the-box search engine
  - Toolbox for an own search application
- Uses index shards from the OWI



More information:

- OWS GitLab: https://opencode.it4i.eu/openwebsearcheu-public/mosaic
- OWS Book: https://openwebsearcheu-public.pages.it4i.eu/ows-the-book

Some applications built on the Open Web Index
Modular Search Application Builder (MOSAIC)



Some applications built on the Open Web Index
Modular Search Application Builder (MOSAIC)

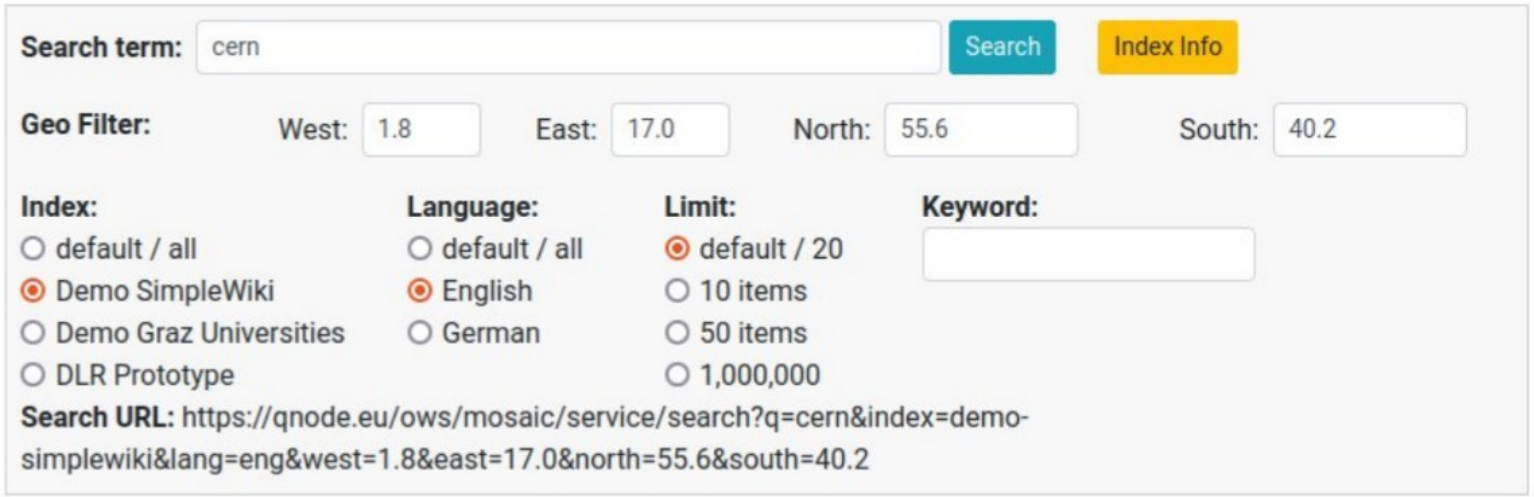
MOSAIC Front-end (for Developers)

Search term

Location filter

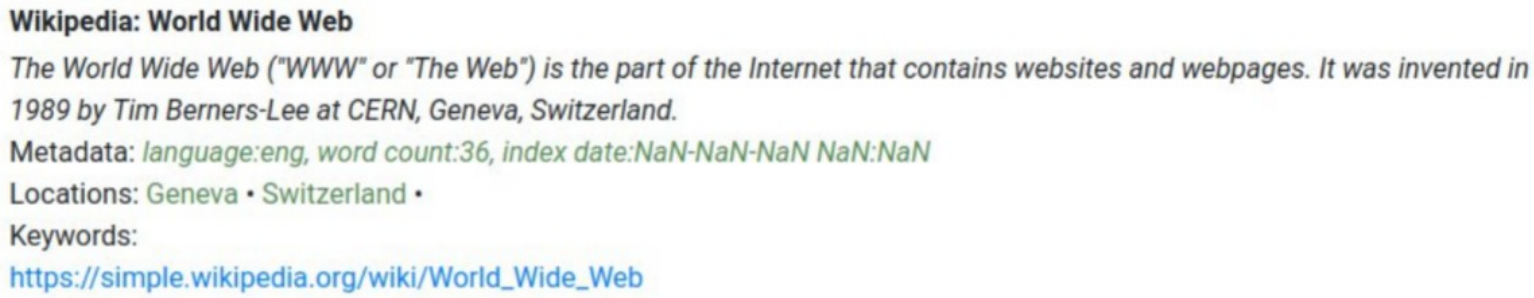
Language filter

Index selection



Text snippet

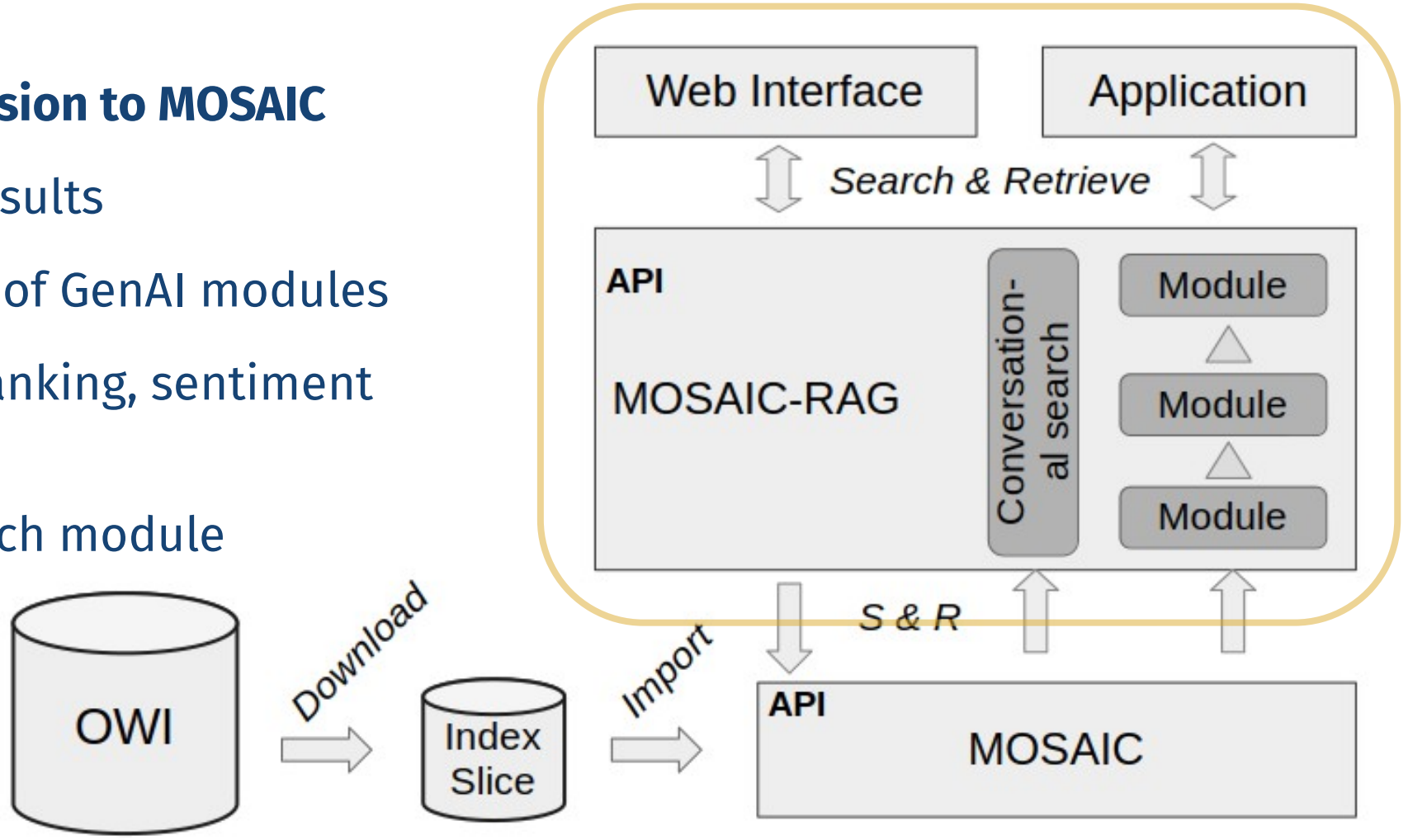
Metadata



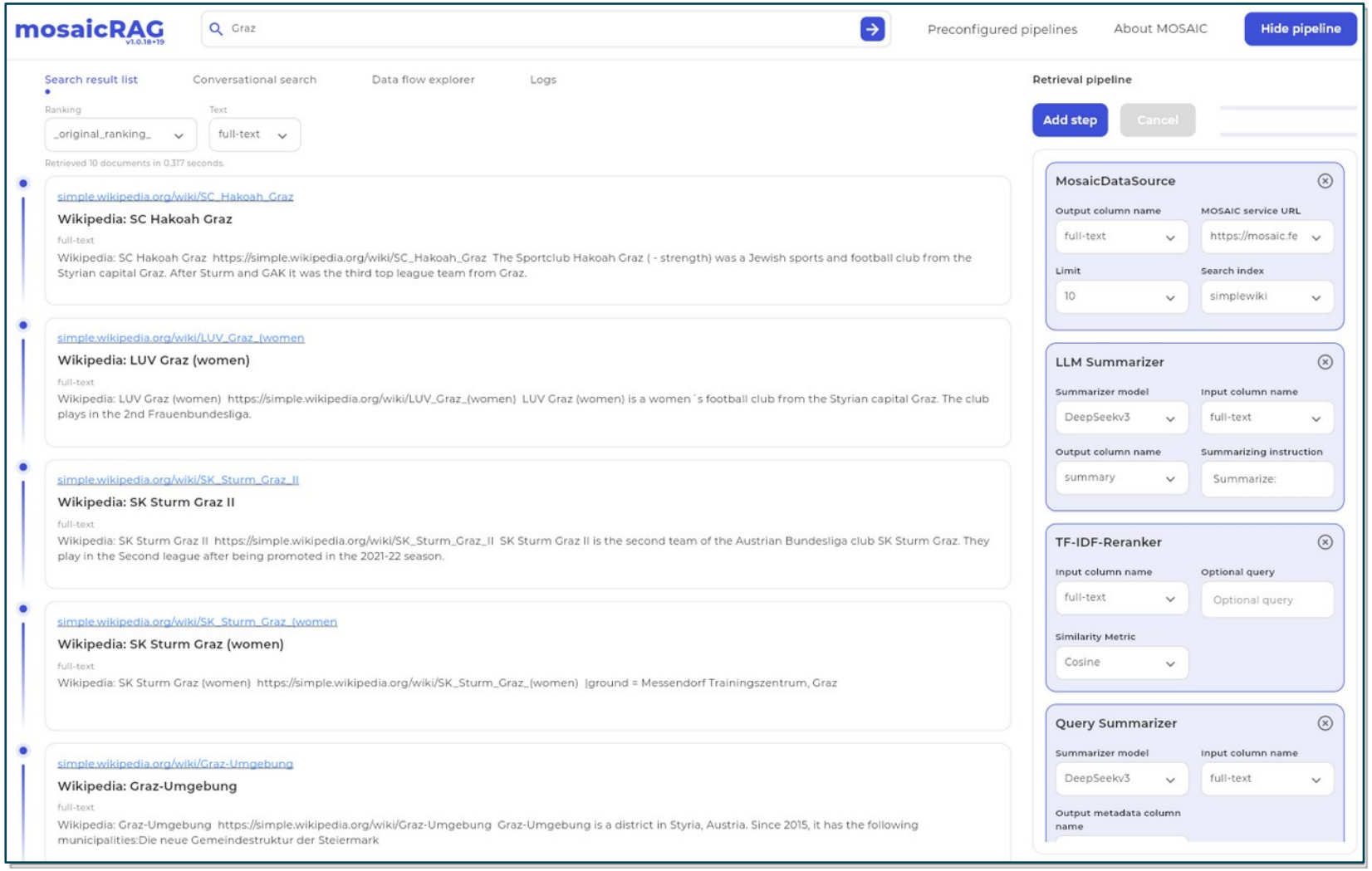
Some applications built on the Open Web Index
MOSAIC-Retrieval Augmented Generation (RAG)

RAG approach as extension to MOSAIC

- Based on MOSAIC results
- Processing pipeline of GenAI modules
- Summarisation, reranking, sentiment analysis,
- Conversational search module



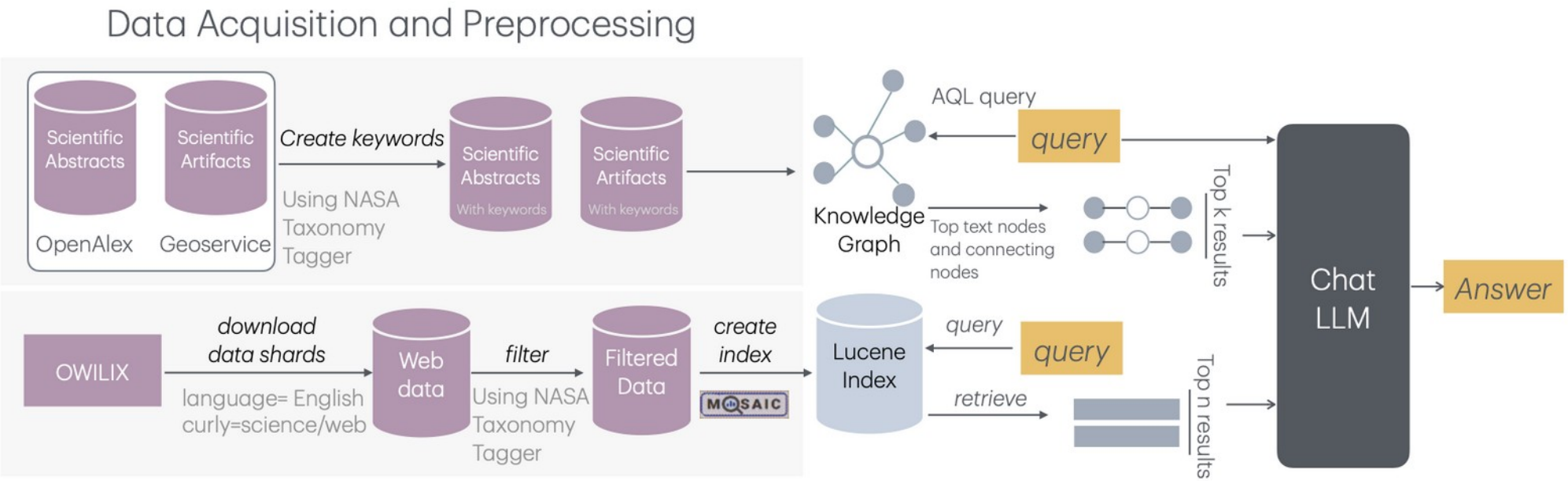
Some applications built on the Open Web Index
MOSAIC-Retrieval Augmented Generation (RAG)



Some applications built on the Open Web Index
Open Science Search: DLR Prototype

Current state

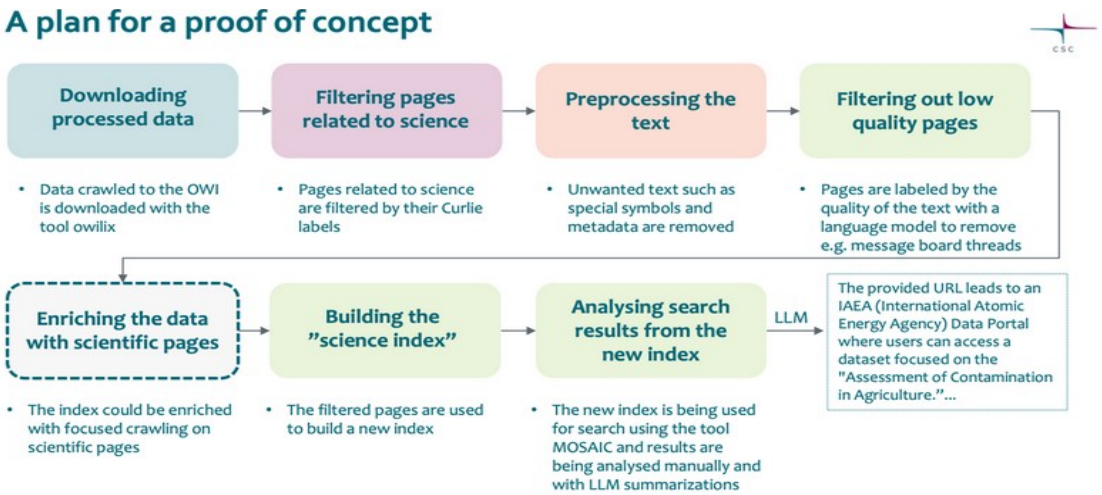
- Working on the *data acquisition and Preprocessing* for multi-genres: scientific abstracts and artefacts, web-data from owilix.
- Developed LLM-based (with multiagents validation) and human-based evaluation plan.
- The *Taxonomy Tagger* was implemented.

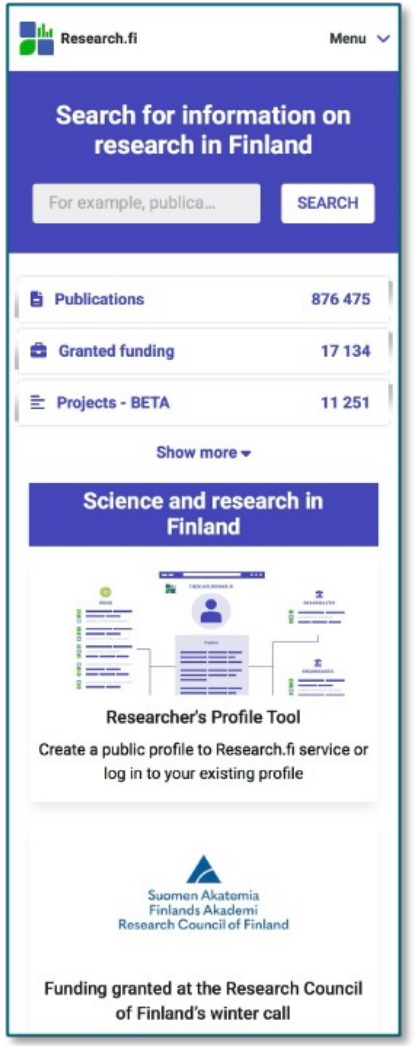


Some applications built on the Open Web Index
Proof-of-concept: Open Science Search by CSC

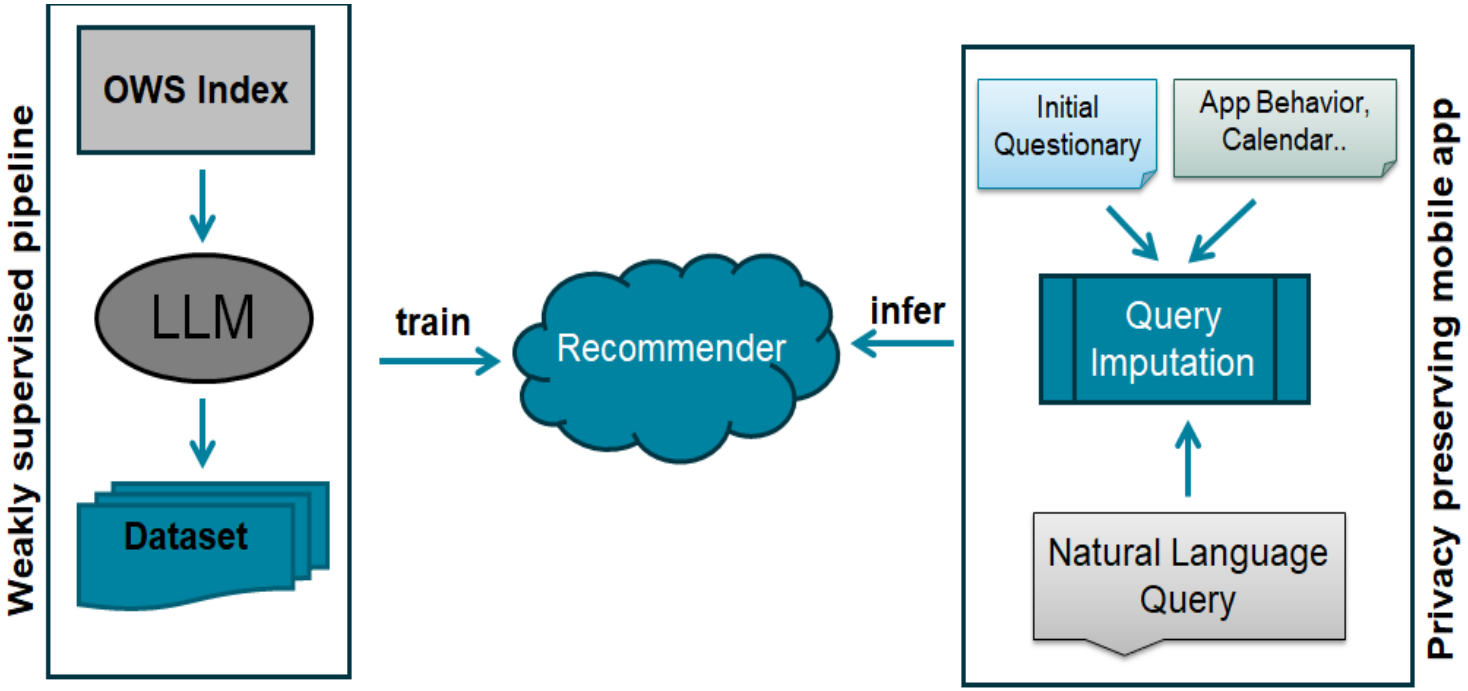
- Research.fi is a Finnish portal for national research outputs, provided by the Ministry of Education and Culture and developed by CSC – IT Center for Science
- Research Information Hub: national aggregator of research-related data in Finland
- Information on research conducted in Finland including publications, grants, organizations and infrastructures.

- **=> Studying how the Open Web Index can be exploited to create a multilingual science index and new features to Research.fi**





Some applications built on the Open Web Index
Mobile privacy-preserving, personalised recommendation of geo-entities by A1

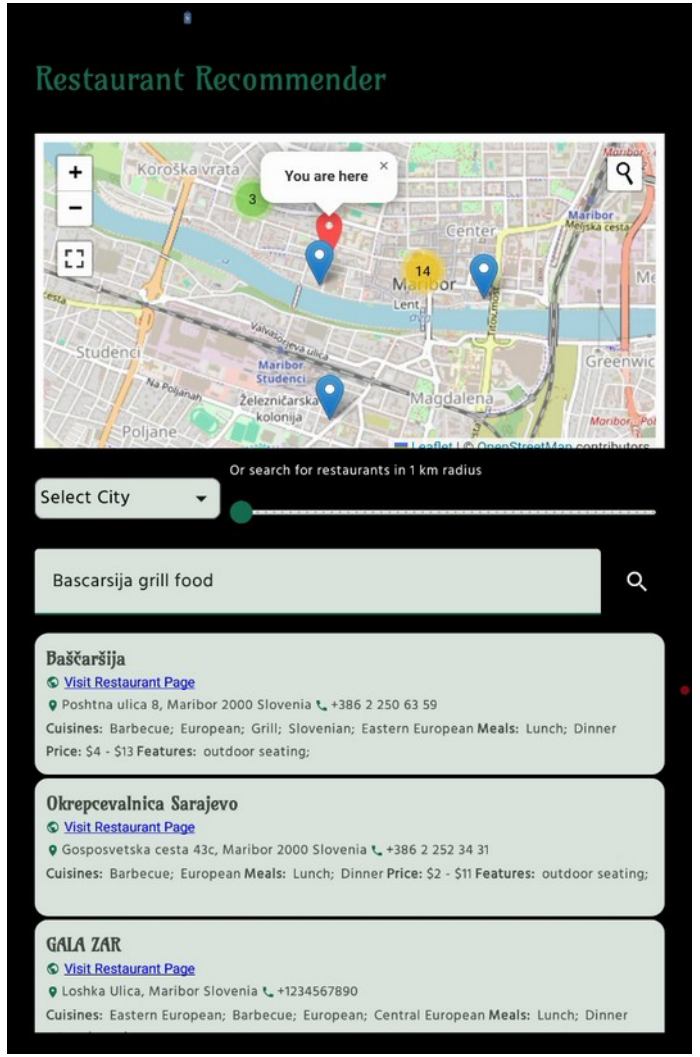


- Development of a content based recommender system
- Definition and deployment of suitable model to enable location- and feature-based search for restaurants
  => e.g. city name, cooking style, price range, rating

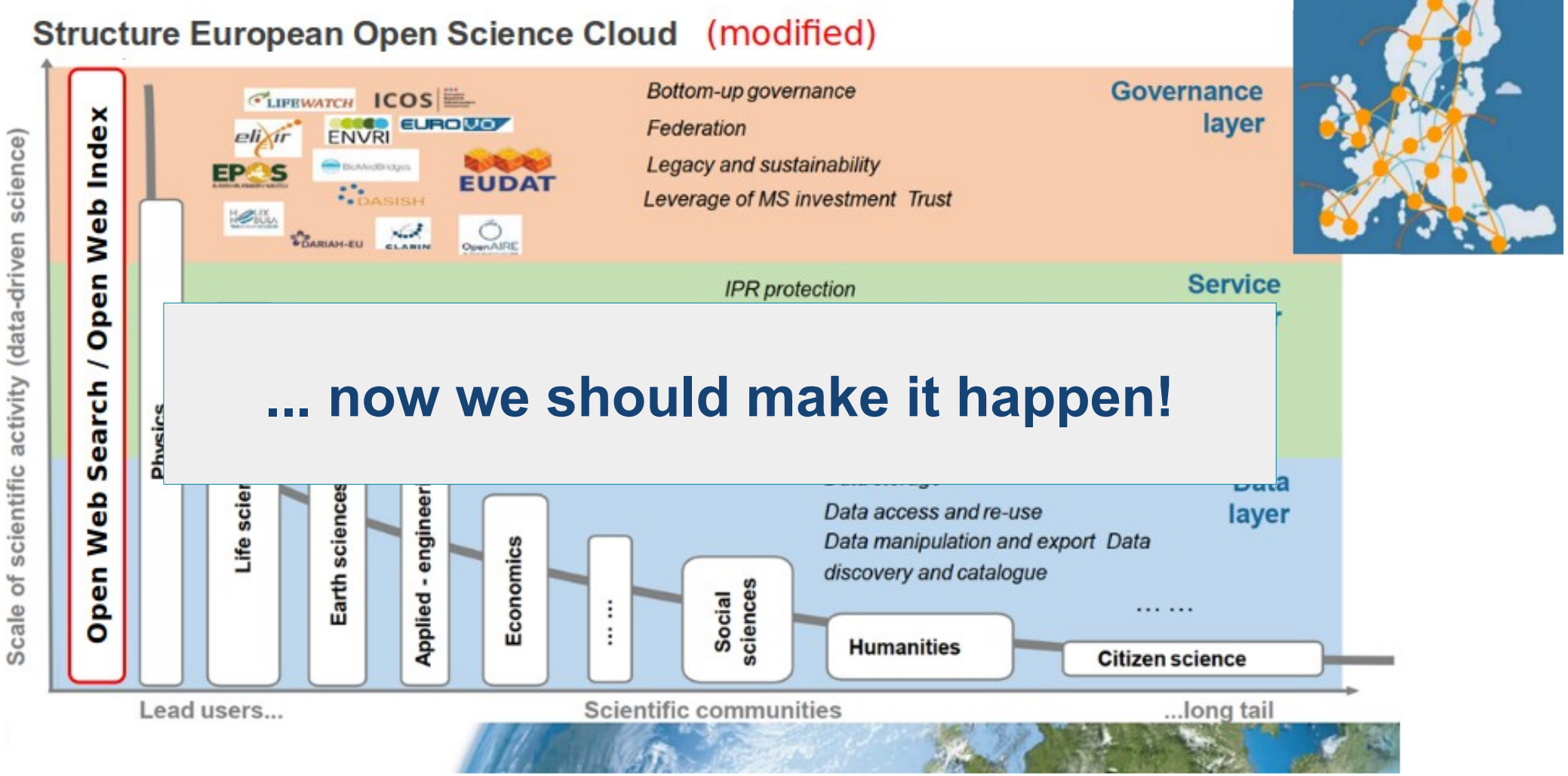
- Integration of Open Web Index as the data source
- Usage of data sets created from Tripadvisor web pages

Some applications built on the Open Web Index
Mobile privacy-preserving, personalised recommendation of geo-entities by A1

- OWS metadata is used to build the dataset
- Meta-Llama-3-8B LLM is used to extract relevant information from “plain_text” column of the OWS metadata
  - Query Imputation pipeline
  - Initial query sent to server
  - Meta-Llama-3-8B detects entities, which are sent back to the app
  - If any entity is missing, query is complemented by user preferences inside phone
- Ranking system
  - We use BM25, a ranking system based on query terms and “bag-of-words” (a combination of all features) of the restaurants
  - Android app using Kotlin, uses OpenStreetMap.



Back in 2018 we already thought how Open Web Search could be part of EOSC ...



... now we should make it happen!

How can Open Web Search and EOSC join forces?

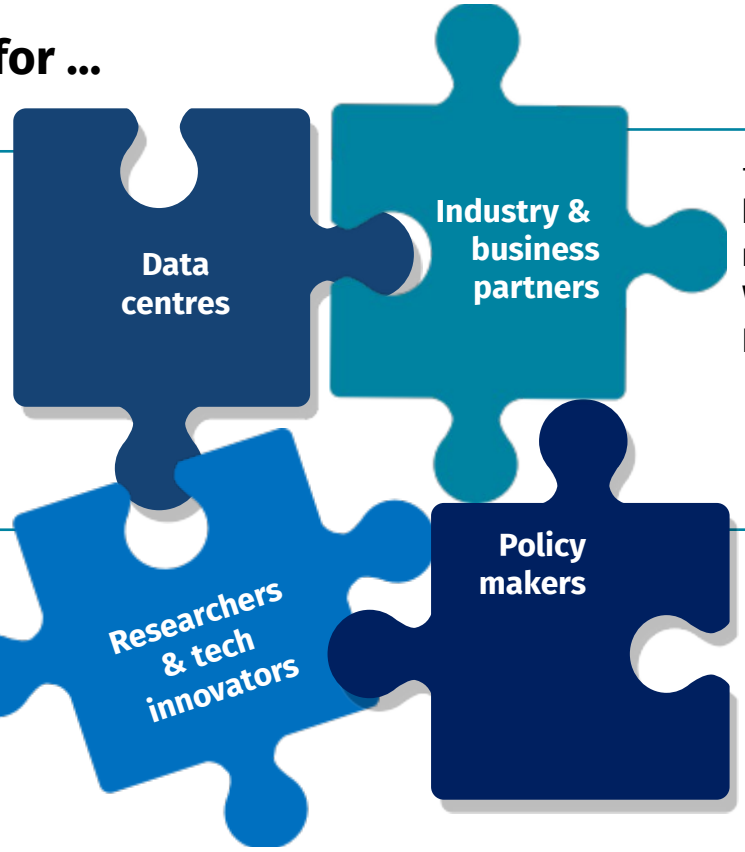
- **Explore synergies** between the two initiatives (technically and organisationally)
- **Exchange on how to set-up and govern** a data-intensive federated infrastructures across Europe
- **Share capacities** (e.g. storage and compute) and **make Open Web Search part of the EOSC services**
- Join forces to **complement scientific data (-spaces) with specific and general web-text corpora** - for enabling search, foresight, training of scientific language models, RAG etc.
- **Exchange on search and analytics strategies** in large large scale distributed data and meta-data repositories (science search, scientific data search, etc.)
- **And many more**...

Europe needs to cooperate across domains and communities to regain sovereignty in using and accessing the Web at scale!

Stefan Voigt
Open Search Foundation
Chairman
Germany

We are looking for ...

→ **Partners to host the distributed Open Web Index**



→ **extend the business and service models of the Open Web Data Infrastructure**

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

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

Contact: ows@openwebsearch.eu & sv@opensearchfoundation.org

Upcoming Events!

Open Web Index - Official Kick-off
June 6, 10:00 - 11:30
Online

https://cscfi.zoom.us/meeting/register/eATIpDQ5TZidh4Jzkim6FQ#/registration

OpenWebSearch.eu session at NGI Forum
June 20, 9:00 - 13:00
Brussels and online

https://ngi.eu/ngi-forum25/