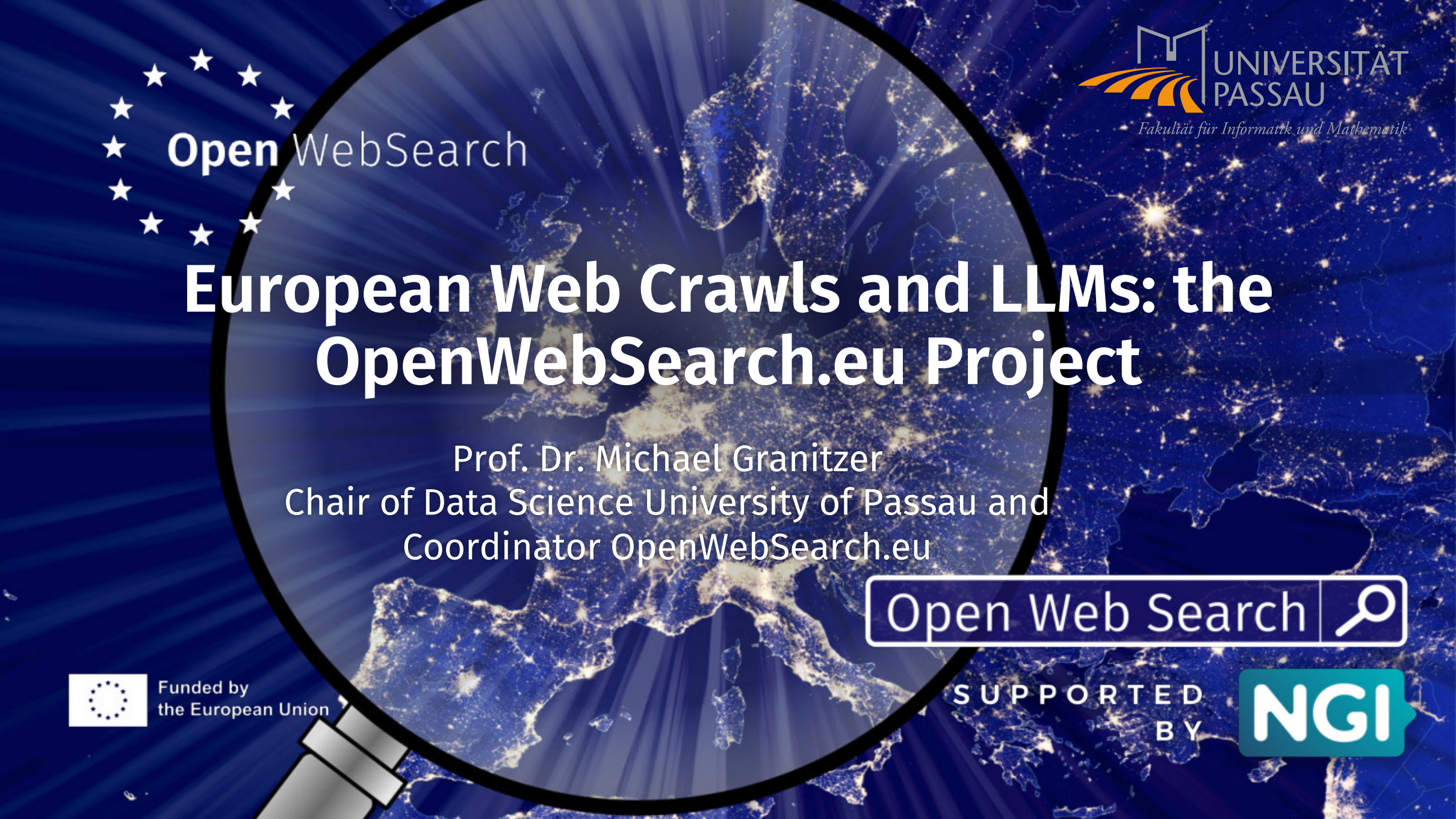

Open WebSearch
UNIVERSITÄT PASSAU
Fakultät für Informatik und Mathematik
European Web Crawls and LLMs: the OpenWebSearch.eu Project
Prof. Dr. Michael Granitzer
Chair of Data Science University of Passau and
Coordinator OpenWebSearch.eu
Open Web Search
Funded by the European Union
SUPPORTED BY
NGI

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

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

Research

lrz

CERN

VSB TECHNICAL UNIVERSITY OF OSTRAVA | IT4INNOVATIONS NATIONAL SUPERCOMPUTING CENTER

CSC ICT Solutions for Brilliant Minds

Infrastructure

osf open search foundation

nlnet FOUNDATION

suma-ev

NGOs

A1

Businesses

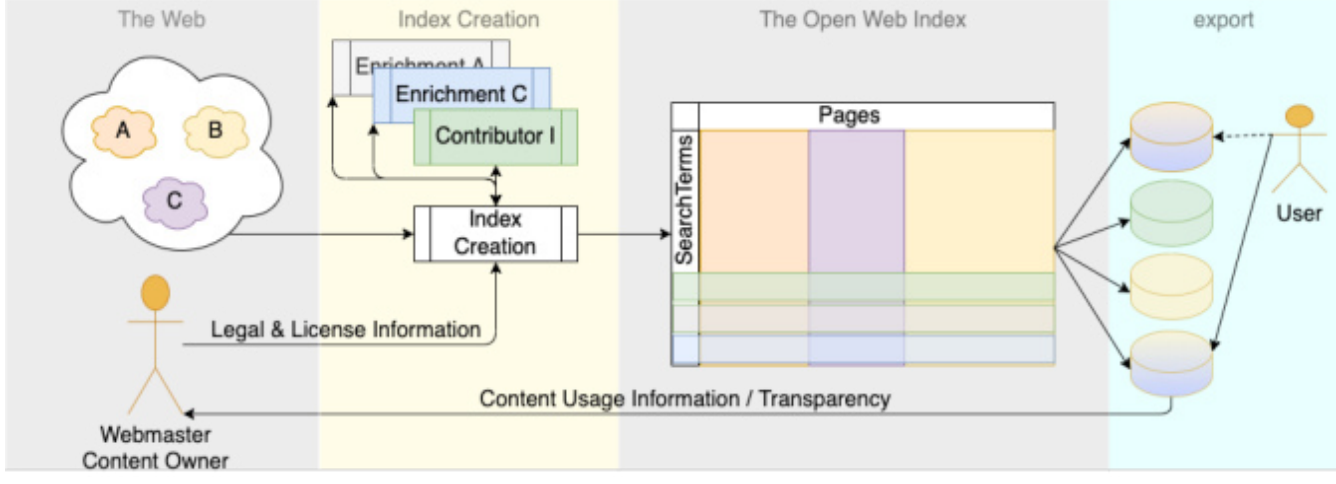
First Goal
Build an Open Web Index Collaboratively

Open WebSearch .eu

- ➜ Build an Open Web Index, i.e. a data structure for searching the Web
- ➜ Empower users, researchers & innovators to build on top of the Index

Principles for an Open Web Index

- Open Data for maximum re-use
- Open Source / Open Configuration for transparency
- Open Resources for using your own compute
- Open to contributions from third parties - empower innovators



- Collaborative management of a Web Index - connect existing infrastructure organisations
- Content control - respect legal, societal and ethical frameworks

Second Goal: Creating an Open Infrastructure

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

Third Goal: Grow an Open Web Search Ecosystem

Open WebSearch .eu

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

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

Data Contributions
(e.g. provision of crawls, content push)

Content Curation
(e.g. science, education, languages)

Technology Contributions
(e.g. enrichment,)

Standards and ELSA Clearance
(e.g. license, metadata formats)

New Search Paradigms
(e.g. Argument search, conversational search)

Vertical Search Engines
(e.g. Open Science / Mobile Location Search)

Language Models
(e.g. language specific, search specific)

Benchmarking
(e.g. search engines, language models)

Web Analytics
(e.g. Content distribution, social media)

Compute

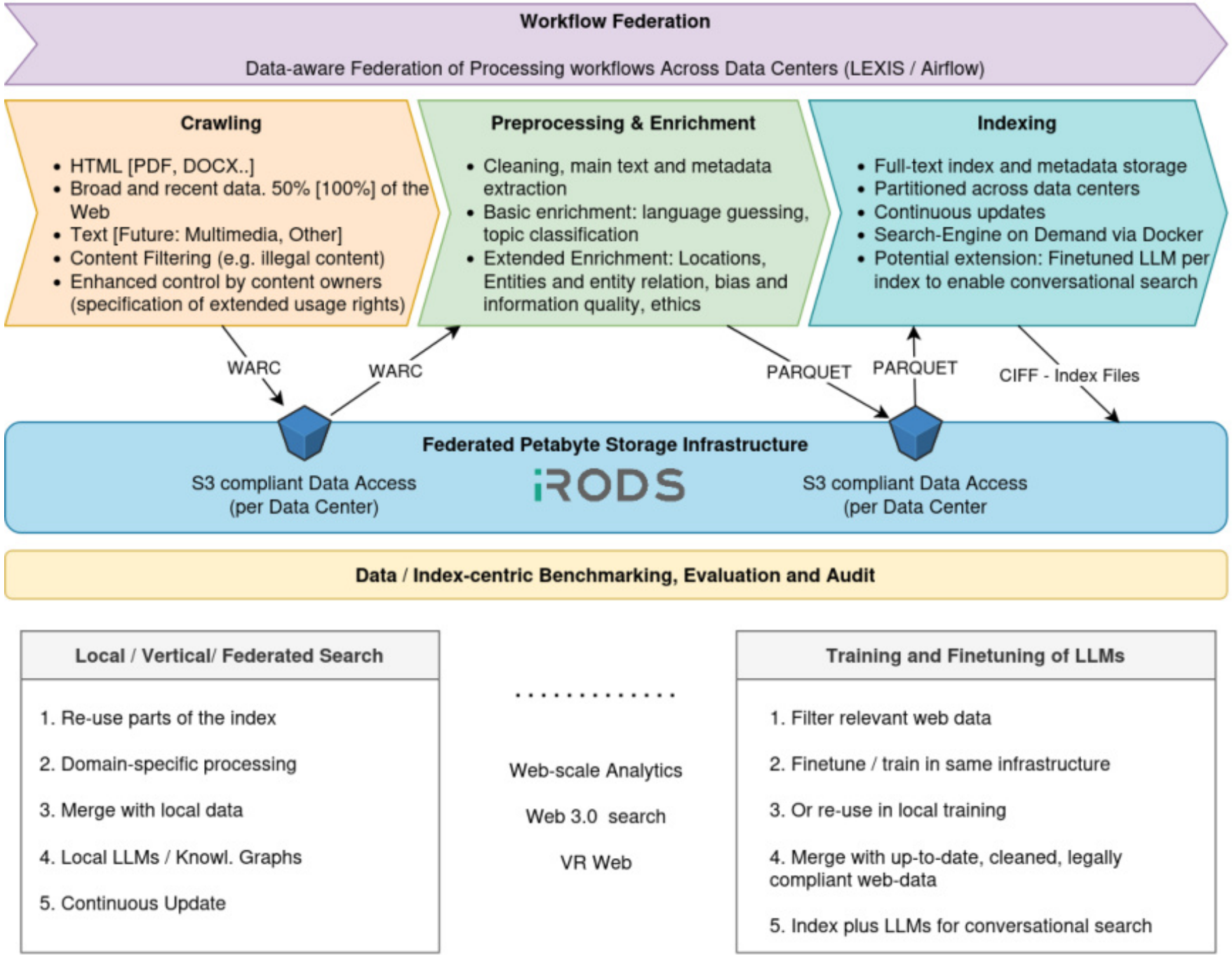
Storage

Search as a Service

Web-scale Platform for heavy-lifting

Applications and Innovations as Multiplicator

Distributed Infrastructure as Enabler



Data Preparation

Storage

Application

Open WebSearch .eu

More technical details on processes and infrastructure

The OWler – an Open Cooperative Web Crawler

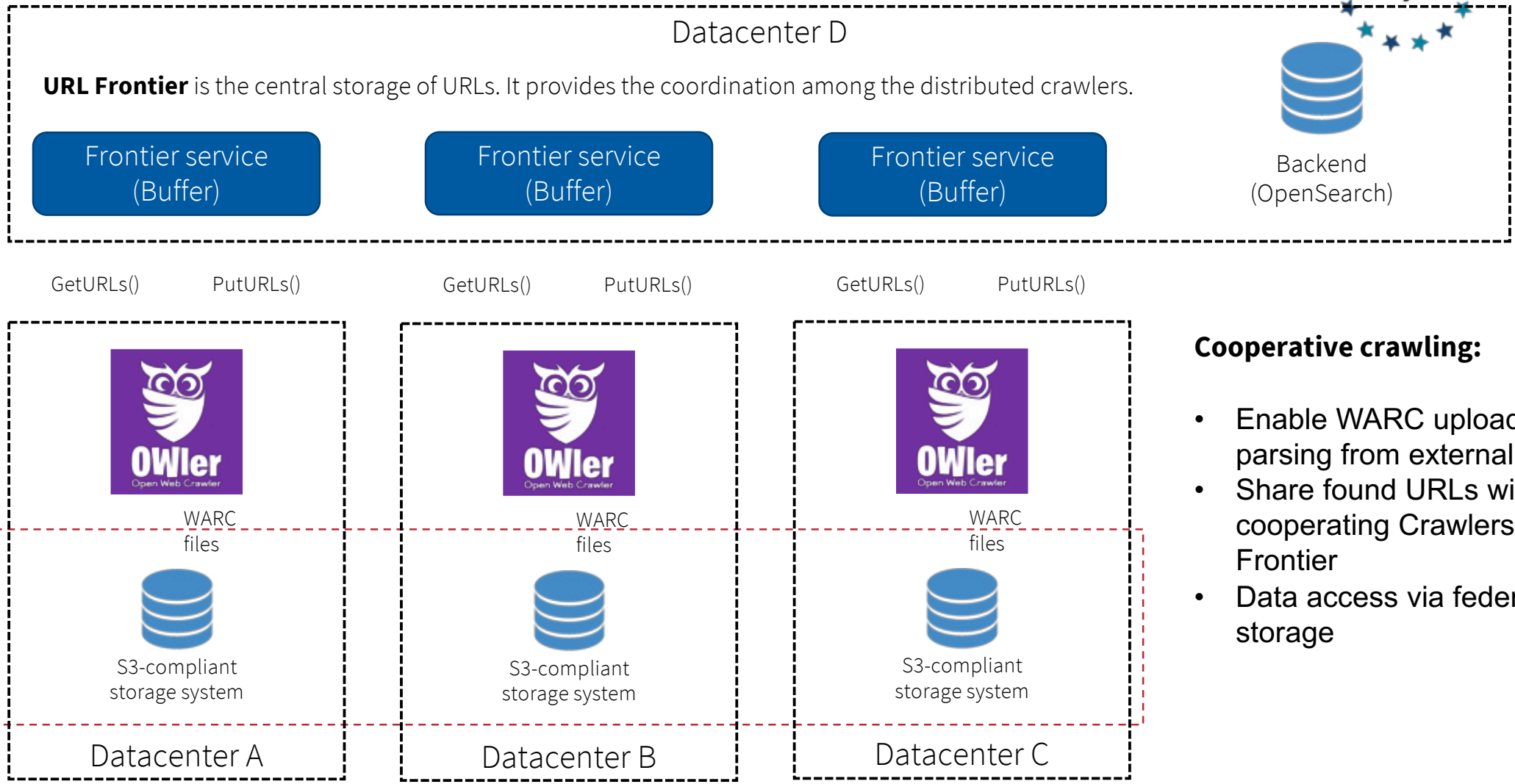
OWler is an incremental and focused web crawler that extends the StormCrawler. It collects web pages that satisfy some specific criteria, e.g., URLs that belong to a given domain or that contain a user-specified topic, and documents its fetch activities in the WARC file format.

Code: https://openwebsearcheu.pages.it4i.eu/wp1/owseu-crawler/owler/

StormCrawler is a popular and mature open source web crawler. It is written in Java and is both lightweight and scalable, thanks to the distribution layer based on Apache Storm.

Apache Storm is a distributed system for processing streams of data. The work is delegated to different types of components that are each responsible for a simple specific processing task.

Cooperative Crawling



Cooperative crawling:

- Enable WARC upload and parsing from external Crawlers
- Share found URLs with cooperating Crawlers via the URL Frontier
- Data access via federated storage

Timeline

- Pipeline implementation and deployment until September 23
- Continuous Crawling from September onwards
- **Results**
  - Partitioned Index for download (Top Level Domain / topic based organisation)
  - Additional: Metadata, cleaned main text
  - Resiliparse as Preprocessing and Cleaning Library
- **Future Challenges**: legal compliance in crawling (e.g. IPR, GDPR), data quality, integration of LLMs, support training for LLMs, scaling-up, legal and ethical clearance

Advantages of an Open Web Search Infrastructure

- → A strategic contribution to **digital sovereignty**
- → Restoring an open, **human-centric search engine market** in Europe - for diverse and unbiased access to knowledge and information.
- → **Lowering entry barriers for tapping the Web as a resource** at scale: for researchers, innovators, and businesses
- → Providing **web services for other digital infrastructures** and data spaces, e.g. the European Open Science Cloud, GAIA-X, EDICs/MCPs, Copernicus, and many more.
- → **Increasing transparency** of web content usage
- → **Increase of control** for content owners and users
- → Enable Europe to develop its own **large-scale language models and generative AI**

Conclusion

- Web and Web Search critical for Europe's digital sovereignty
- Reducing entry barriers for researchers, innovators and business to increase competitiveness with Big Tech
  - Opening the core element in Web Search: an Open Web Index
  - Open software stack and open pipelines
  - A distributed, data-centric and Web-scale compute and storage analytics infrastructure
- Build an ecosystems around data, software and infrastructure
- Transparency and control for considering legal, ethical and societal values

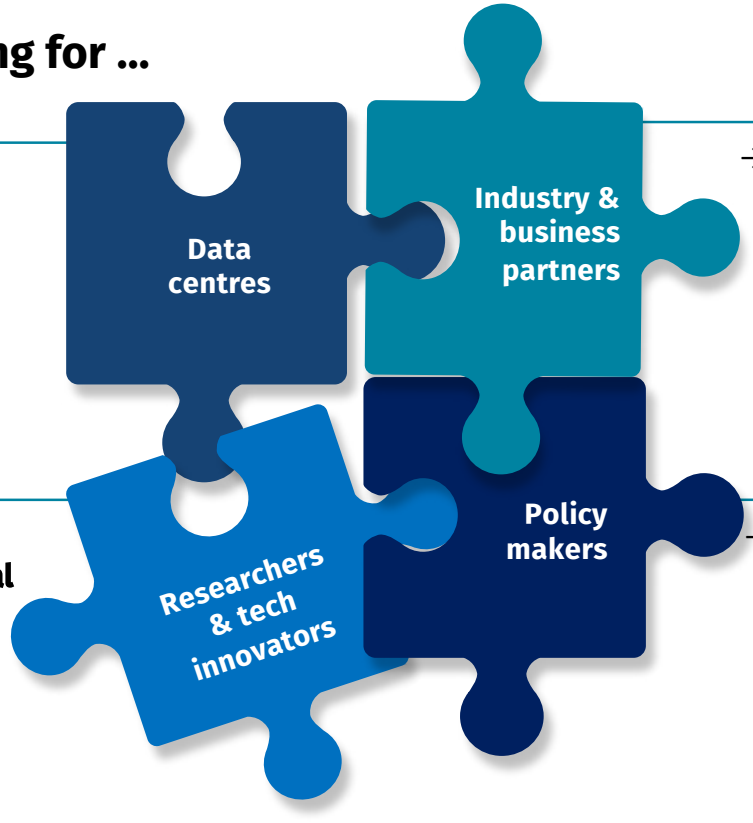
New Search Paradigms
(e.g. Argument search, conversational search)

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

→ discover the business models of an Open Web Index

→ develop new search & retrieval paradigms and content analysis algorithms

→ help shaping the governance of an open search ecosystem

→ We will also offer small grants for potential contributors.