Creating Search Applications Using the Open Web Index

Alexander Nussbaumer, Roxanne ElBaff, Json Theodoropoulos,
Noor Afshan Fathima,Izidor Mlaker, Ines Zelch

Webinar
7 April 2025

Open Web Search

Funded by
the European Union

SUPPORTED BY NGI

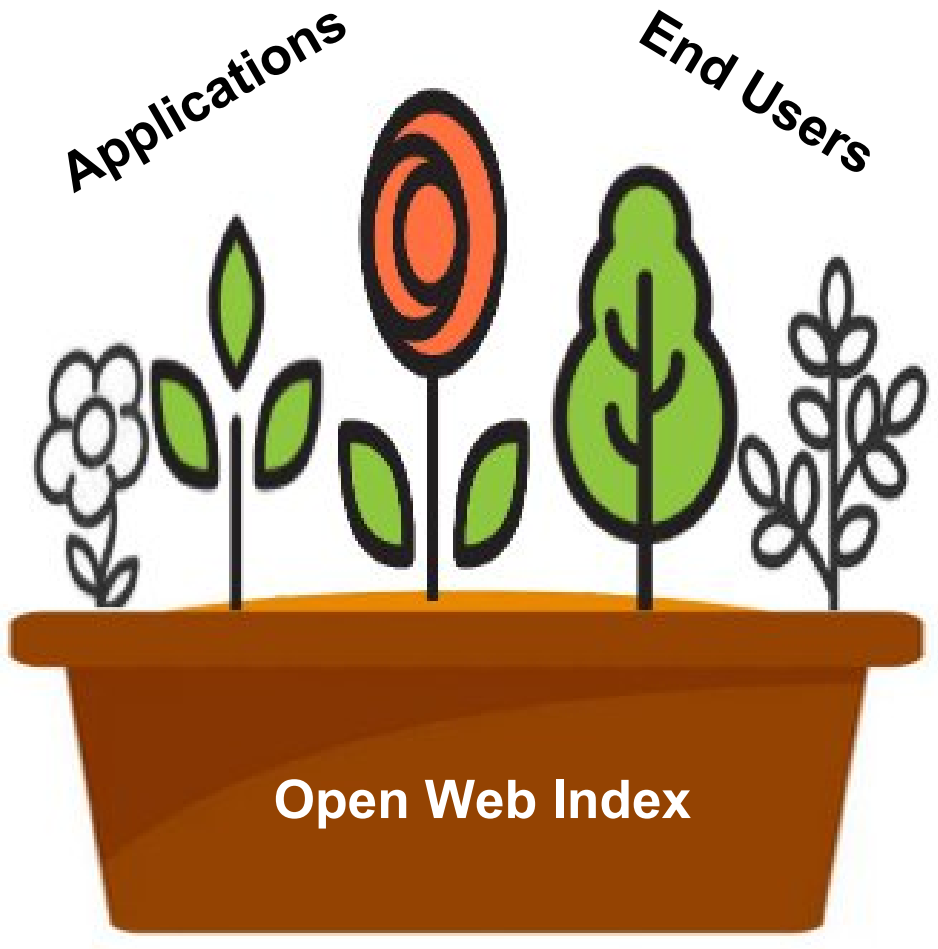
Overview and Agenda

Part 1: Concept of search applications

- Introduction and overview
- Creating search applications

Part 2: Examples of search applications

- Science Search: DLR Use Case
- Science Search: CSC Use Case
- Science Search: CERN Use Case
- Location-based Search: A1 Use Case
- Argumentation Search



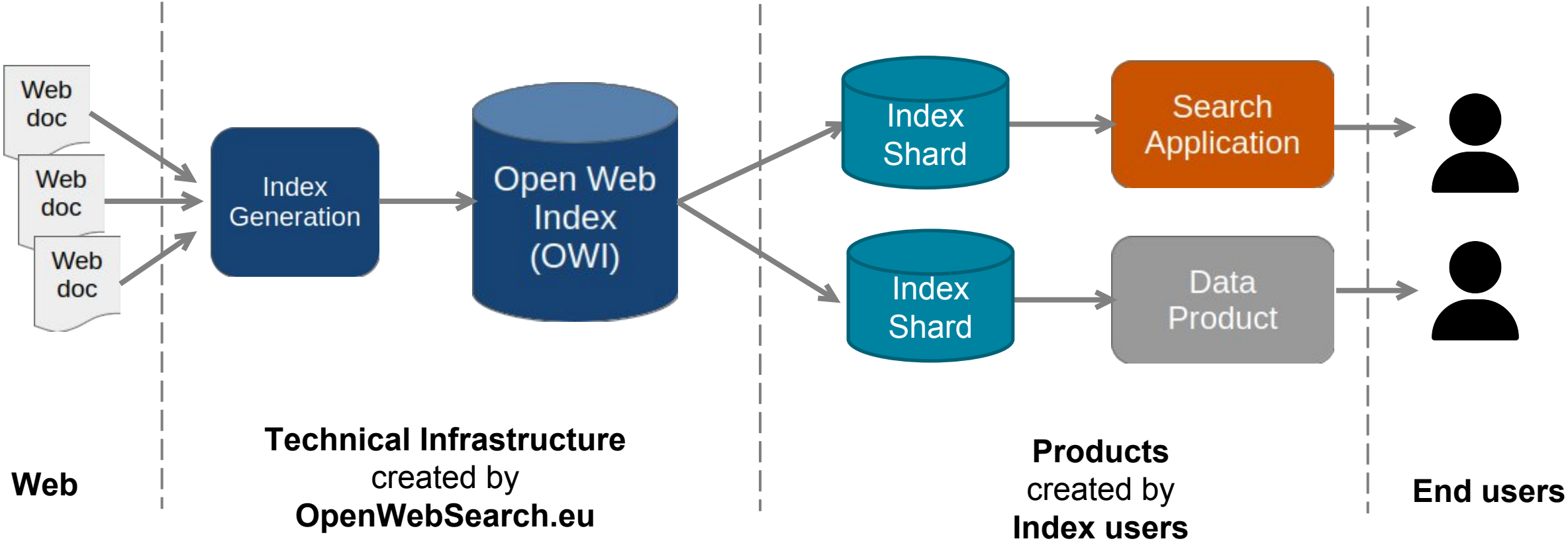
Part 1:
Concept of Search Applications

Terminology

- **General Search**: Search over whole Open Web Index, similar to large search engine available on the market
- **Vertical search engine**: Special search engine related to a specific domain or purpose, such as product search
- **Search application**: Application with more complex functionality than traditional search consisting of search queries and results
- **Conversational search**: Retrieving information through conversation in natural language, such as ChatGPT
- **Retrieval-Augmented Generation:** Supporting conversational search through classical information retrieval

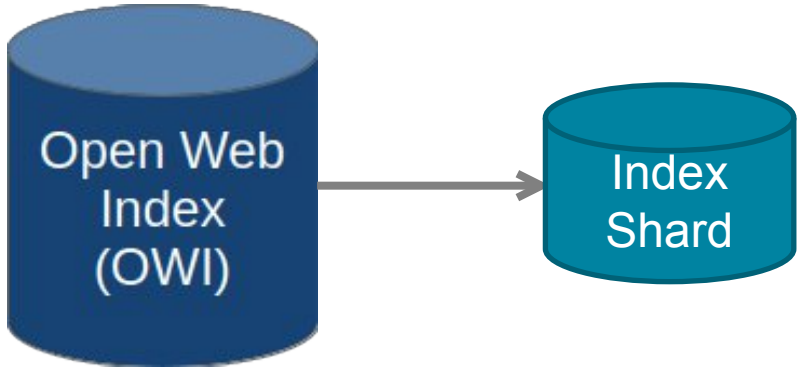
Search Applications in OpenWebSearch.eu

- The overall process of creating and using the Open Web Index (OWI)



Retrieving Index Partitions with Owilix

- Owilix allows to download index shards from the Open Web Index (OWI)
- Command line tool with flexible command structure:
  - Filter by data center, creation date, metadata (e.g. language, topic)
- Registration, End User License, Ethical Self-Assessment



More information:

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

Structure of the Index (OWI and Index Shards)

- Index consists of CIFF and Parquet files

CIFF File: Inverted Index
- Reference to all web documents

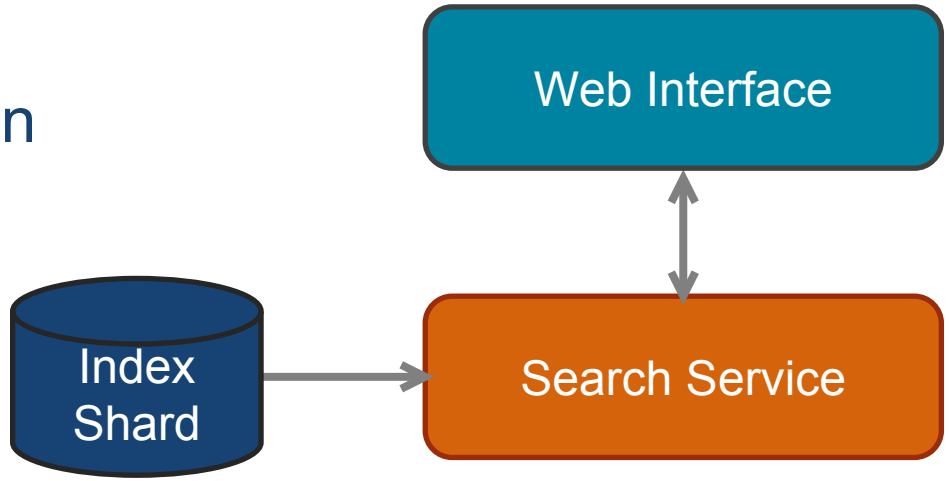
Term	List of (Document, Frequency)
Term 1	(docID A, 1) (docID B, 3) (docID C, 1)
Term 2	(docID D,2)
Term 3	(docID E, 4) (docID F, 2)
Term 4	(docID H, 2) (docID I, 1)

Document	Lang	Full text	WARC date	Location	<Curlie topic>
docID A	eng	... text ...	2023-07-01		
docID B	deu	... text ...	2023-07-01		
docID C	deu	... text ...	2023-07-01		
docID D	eng	... text ...	2023-07-01		

Parquet File: Metadata
- full text
- language
- geo-information
- micro-data
- topic (curlie label)
- links
- embeddings
- ...

MOSAIC

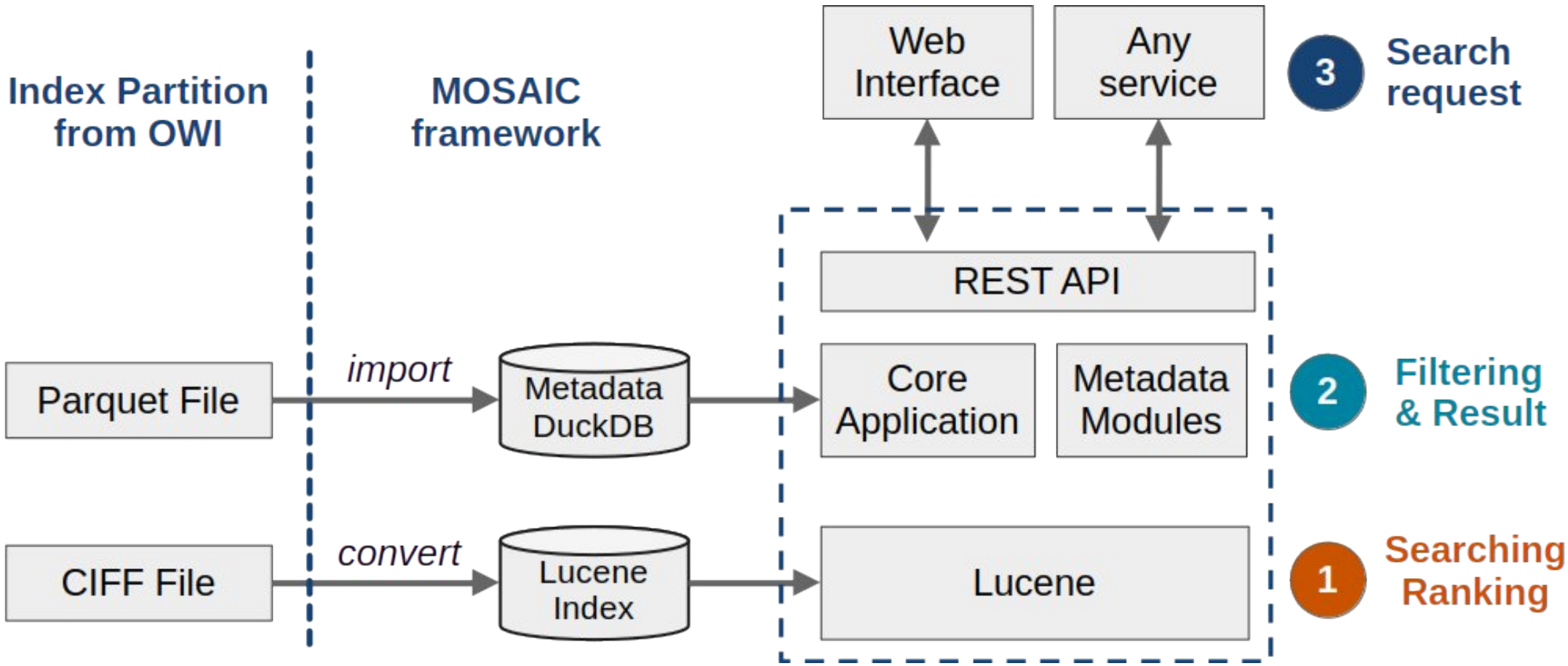
- **Mo**dular **S**earch **A**pplication based on **I**ndex Fra**c**tions
- Generic implementation of an OWS.eu vertical search engine
  - Demonstration of the concept of an OWS.eu vertical engine
  - Out-of-the-box search engine
  - Toolbox for an own search application
- Uses index shards from the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



More information:

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

MOSAIC Concept



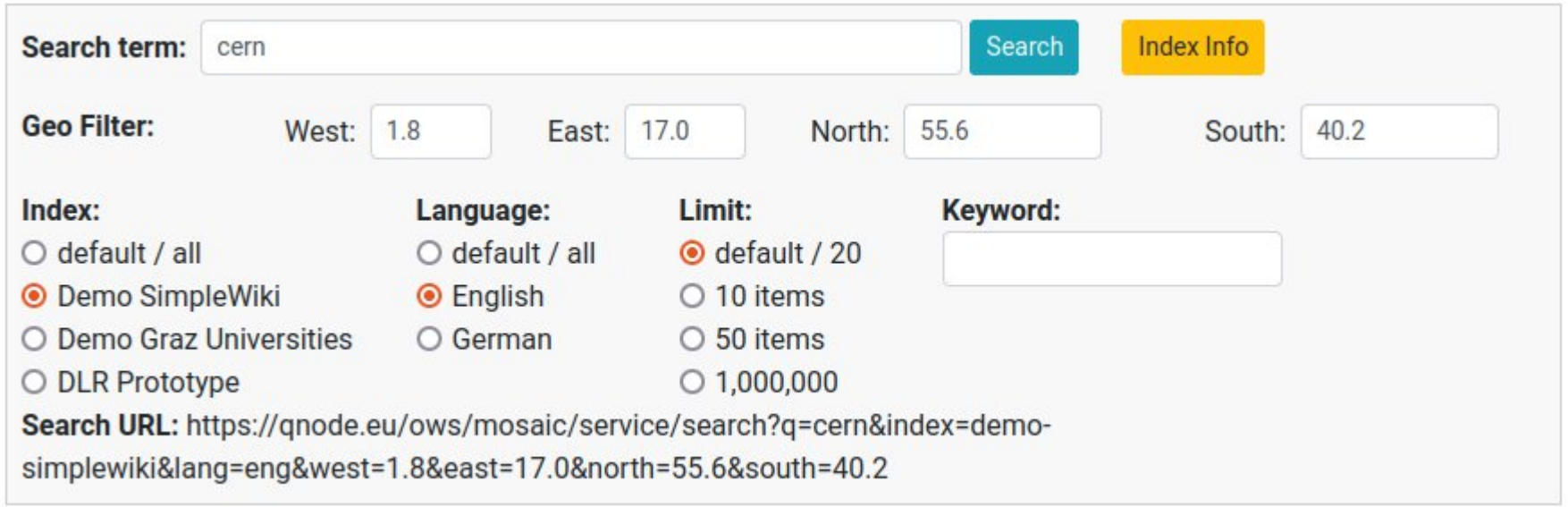
MOSAIC Front-end (for Developers)

Search term

Location filter

Language filter

Index selection



Text snippet

Metadata

Wikipedia: World Wide Web

The World Wide Web ("WWW" or "The Web") is the part of the Internet that contains websites and webpages. It was invented in 1989 by Tim Berners-Lee at CERN, Geneva, Switzerland.

Metadata: *language:eng, word count:36, index date:NaN-NaN-NaN NaN:NaN*

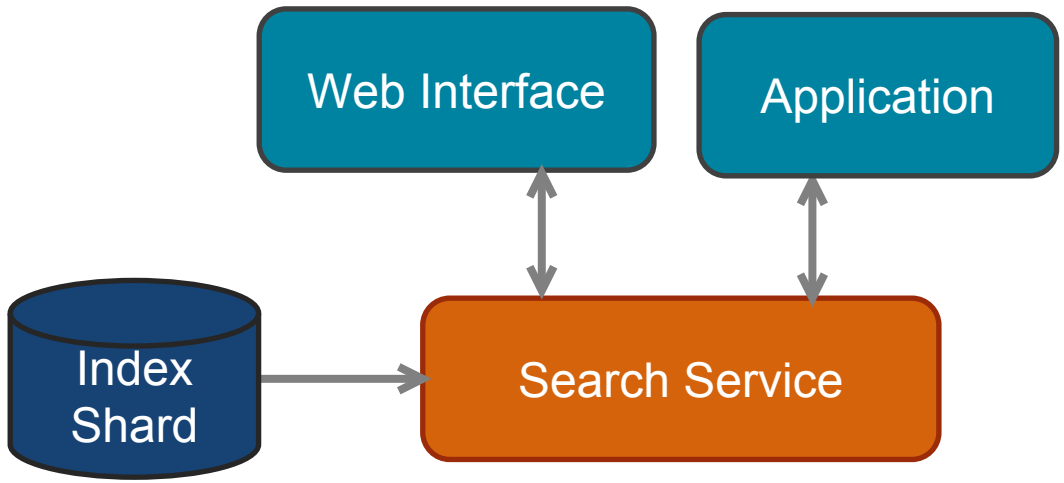
Locations: Geneva • Switzerland •

Keywords:

https://simple.wikipedia.org/wiki/World_Wide_Web

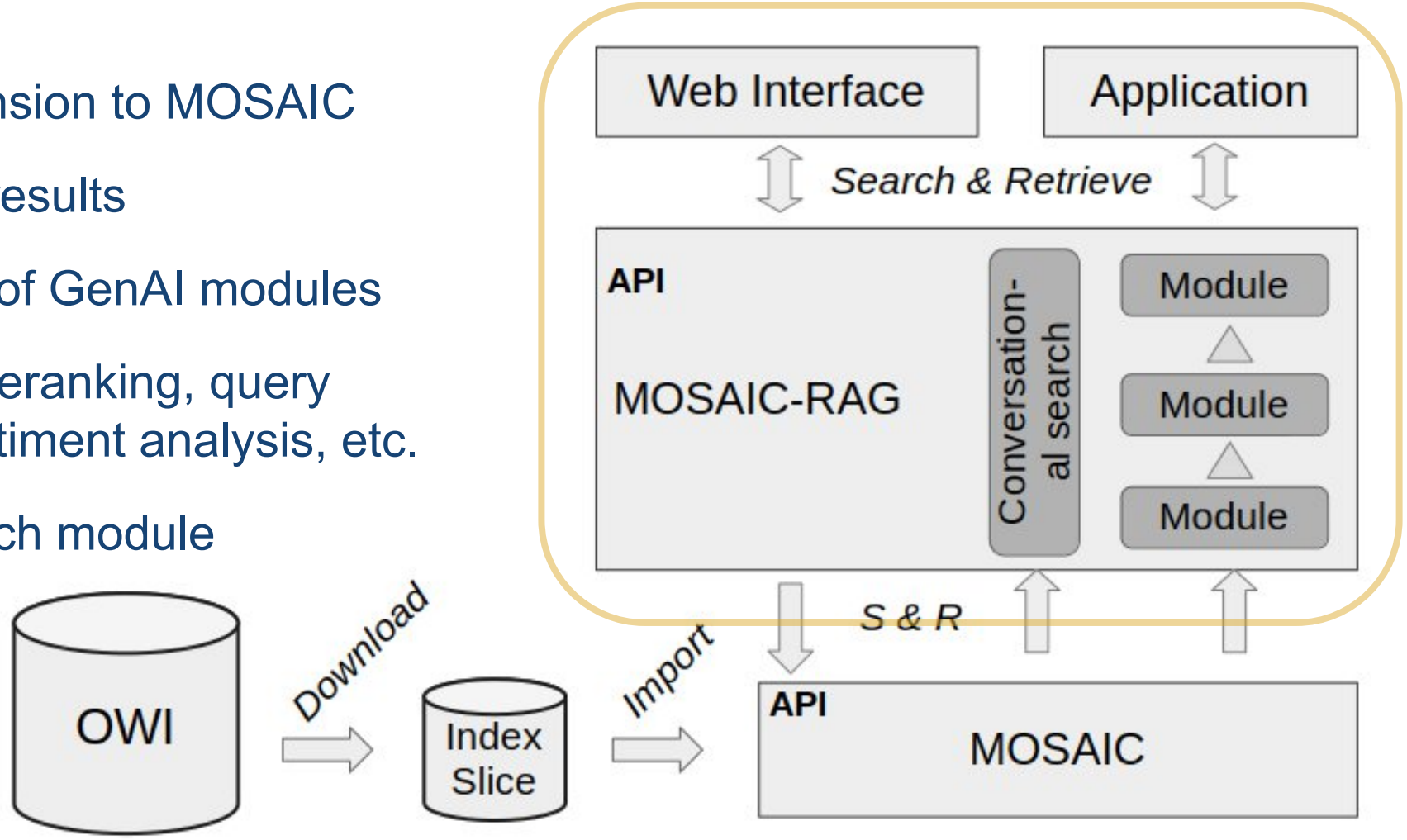
Creation of a Search Application

- Search Domain
  - Define the search domain and download/create the index shard
- Search Service
  - Use MOSAIC out-of-the-box
  - Update MOSAIC (add metadata module)
  - Create your own Service/Application
- Front-end
  - Create your own Web Interface
  - Create an Application



MOSAIC-RAG

- RAG approach as extension to MOSAIC
  - Based on MOSAIC results
  - Processing pipeline of GenAI modules
    - Summarisation, reranking, query optimisation, sentiment analysis, etc.
  - Conversational search module



Part 2:
Examples of Search Applications

Open Science Search: DLR Use Case
Motivation

Challenges in Science Search

Information Overload

- Overwhelming amount of scientific information

Current Limitations

- Fragmented data sources (publications vs. blogs)
- Lack of contextual, geolocation-based search

Objective

A system that

- ✓ **bridges multiple text genres**
- ✓ enables **vertical search** (focusing on one domain)
- ✓ enables geolocation-based contextual search

to help researchers navigate the vast scientific literature for a scientific domain.

Domain-focus: Earth Observation Science (EO) used by different fields such as oceanography and environmental science,

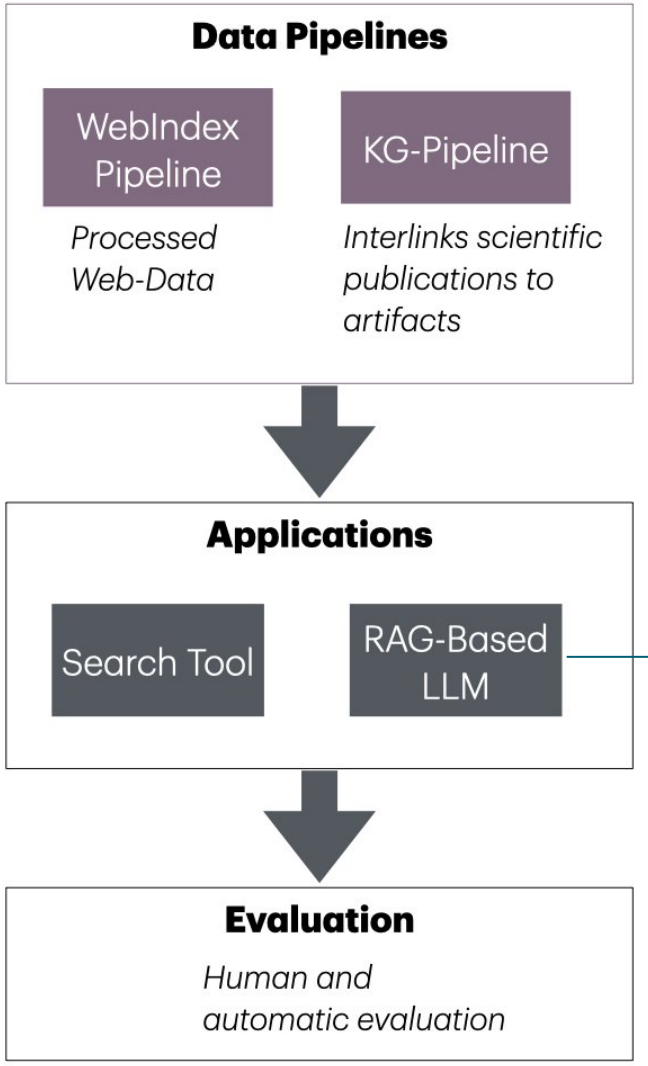
Open Science Search: DLR Use Case
A Hybrid Approach

A 3-Stage Approach

1. **Data Pipelines**
   - Creating a WebIndex from web-data (using owilix).
   - Creating a Knowledge graph (KG) to interlink different scientific text genres.

2. **Applications**
   - Using the data pipelines for: a web search tool and a RAG-Based LLM.

3. **Evaluation**
   - *Search Tool*: evaluating the web application by domain experts, by pre-defining use cases.
   - *RAG-Based LLM:* Evaluating the importance of each data pipeline using ablation.



RAG combines retrieval and generation to improve response accuracy by using external knowledge: it reduces errors in domain-specific tasks.

Open Science Search: DLR Use Case
Data Pipelines (1/2)

Domain: Earth Observation Science

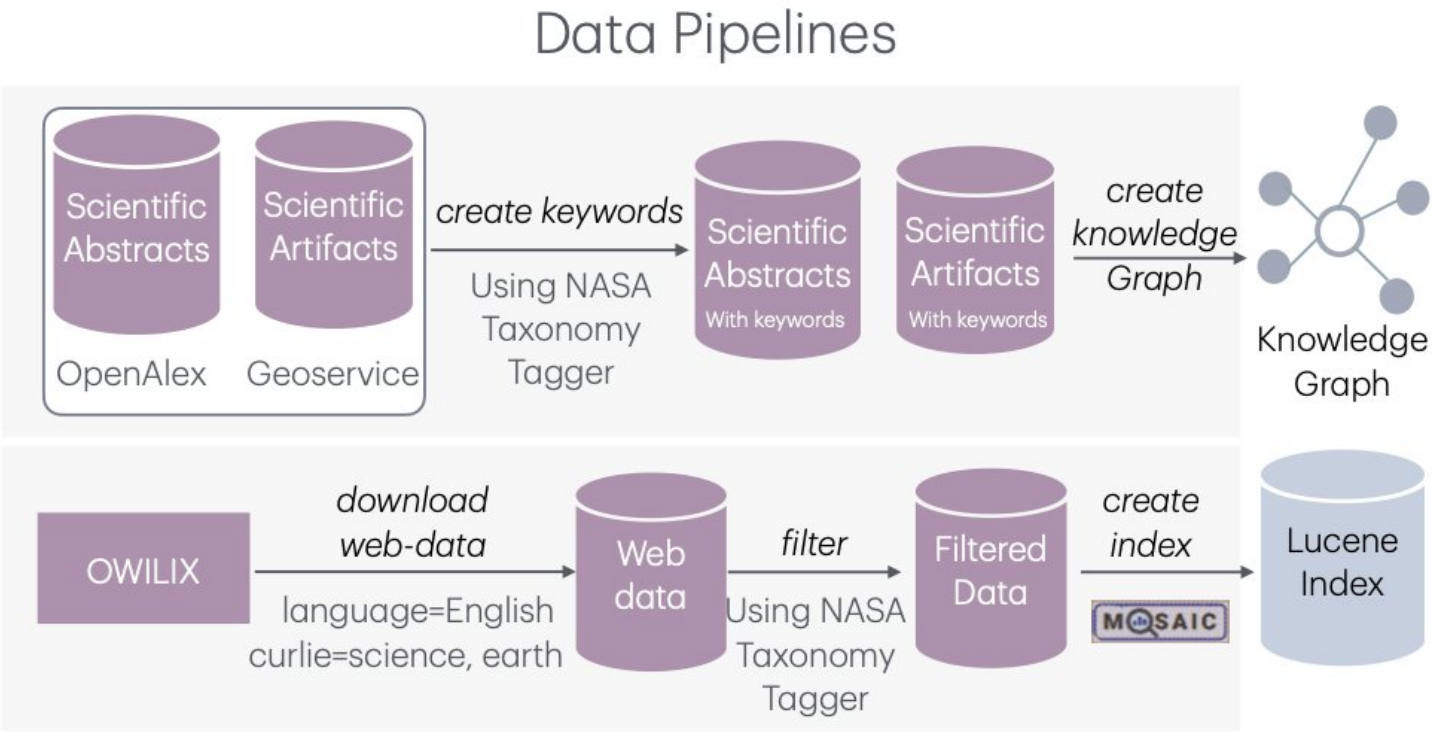
INDEX-Pipeline:

- Uses **OWILIX** to download English web-data shard.
  - *OWILIX allows filtering based on language*
- Gets vast web data relevant to *science* and *earth* domains
  - Tagged with curlie labels: ['geo', 'science', 'earth', 'bio']
    - *A curlie label is a category used to classify websites.*
  - We get a CIFF file and a parquet file (containing metadata such as curlie and geolocation info)
- Filters out web pages with no EO keywords, using TaxoTagger.
- Indexes data using **MOSAIC**
  - Takes CIFF+PARQUET and creates a LUCENE INDEX
    - *A Lucene index is a data structure that allows fast and efficient searching of large volumes*.

KG-Pipeline

- Downloads scientific publications and EO artifacts
- Uses the TaxoTagger for tagging each text using the NASA EO taxonomy *(given a text, the tagger returns the top n keywords with a score)*
- Builds a knowledge graph connecting abstracts and artifacts via *keywords*.

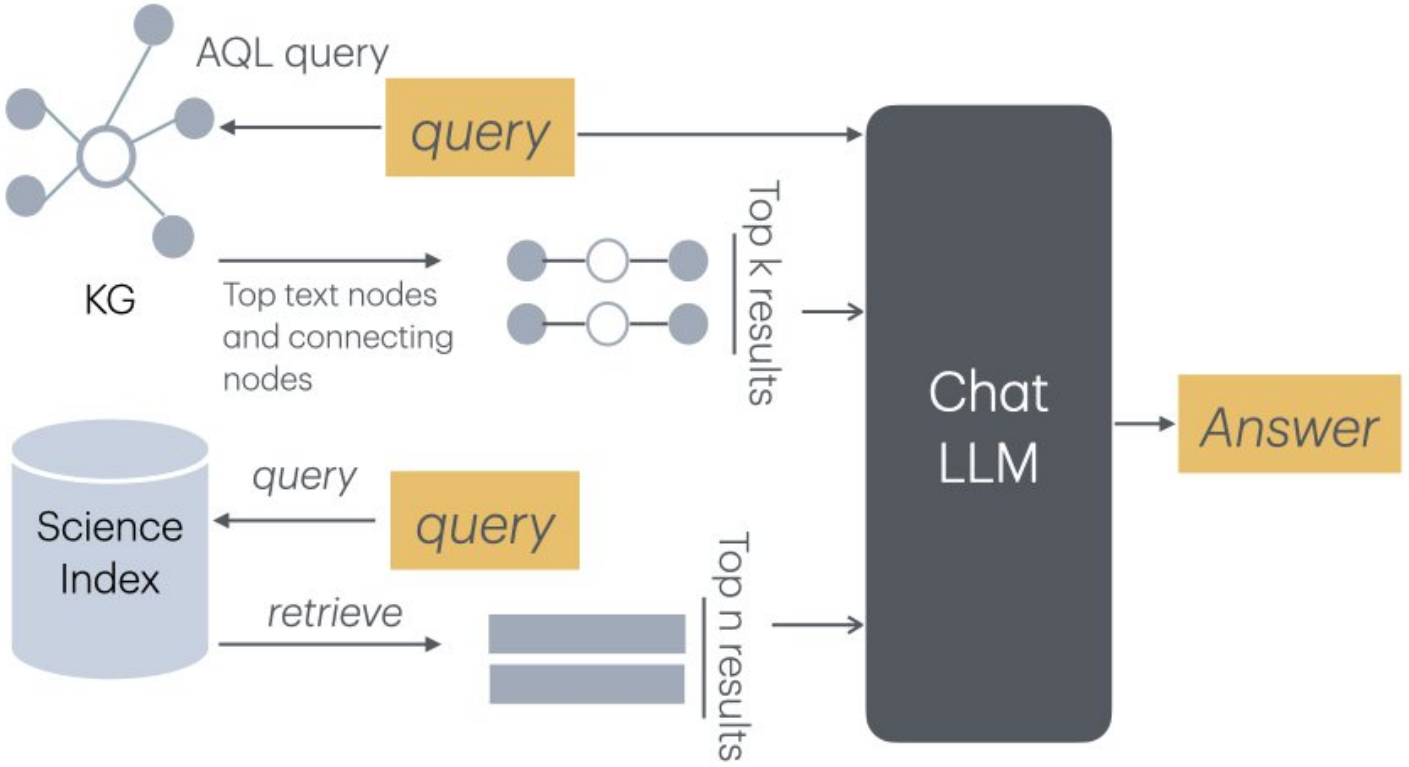
Open Science Search: DLR Use Case
Data Pipelines (2/2)



Open Science Search: DLR Use Case
2. Applications (1/2) – RAG-BASED

A Hybrid Retrieval

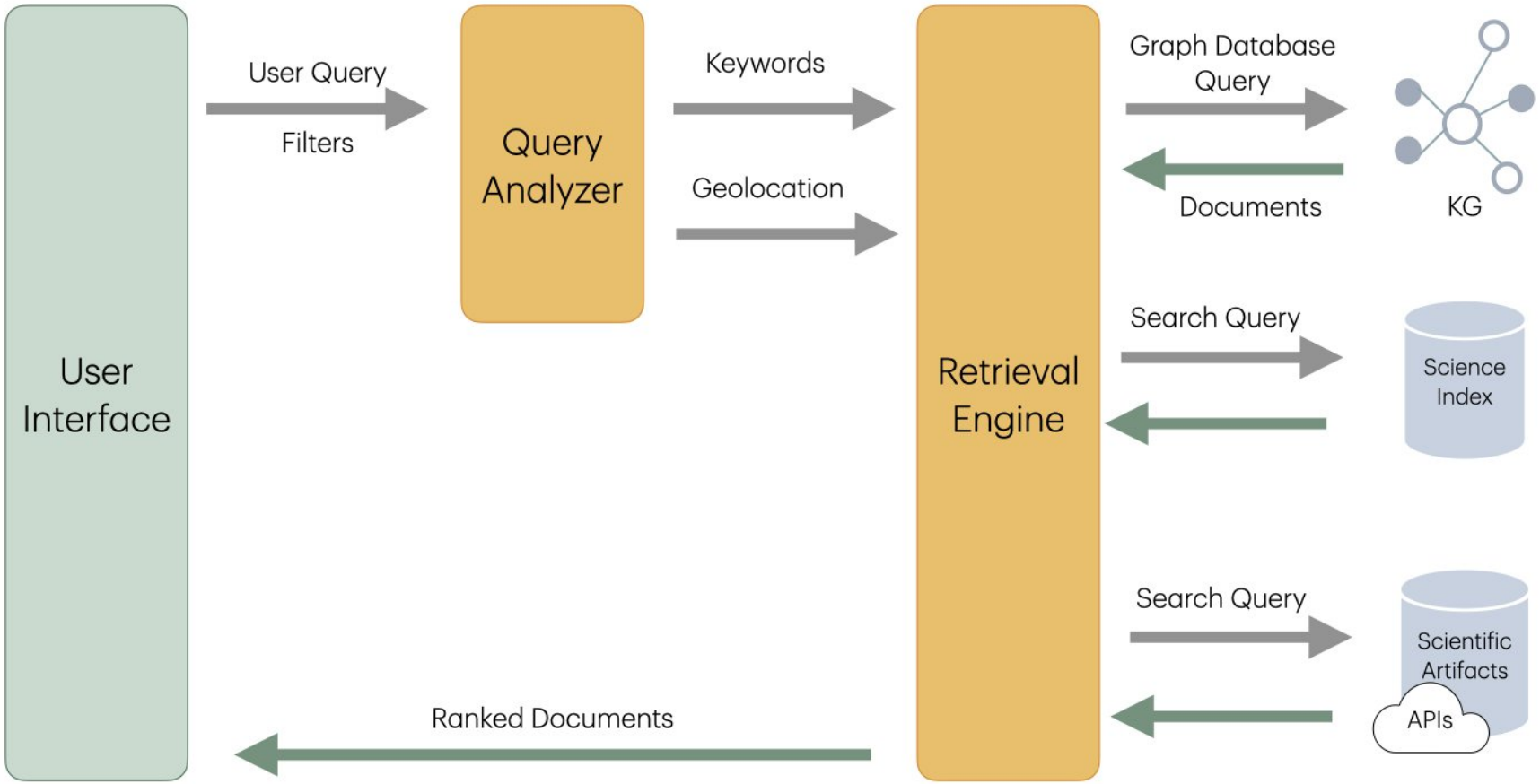
- Combines indexed web data and knowledge graph and
- Query results from both pipelines are merged into
  - LLM prompts within the RAG-based System



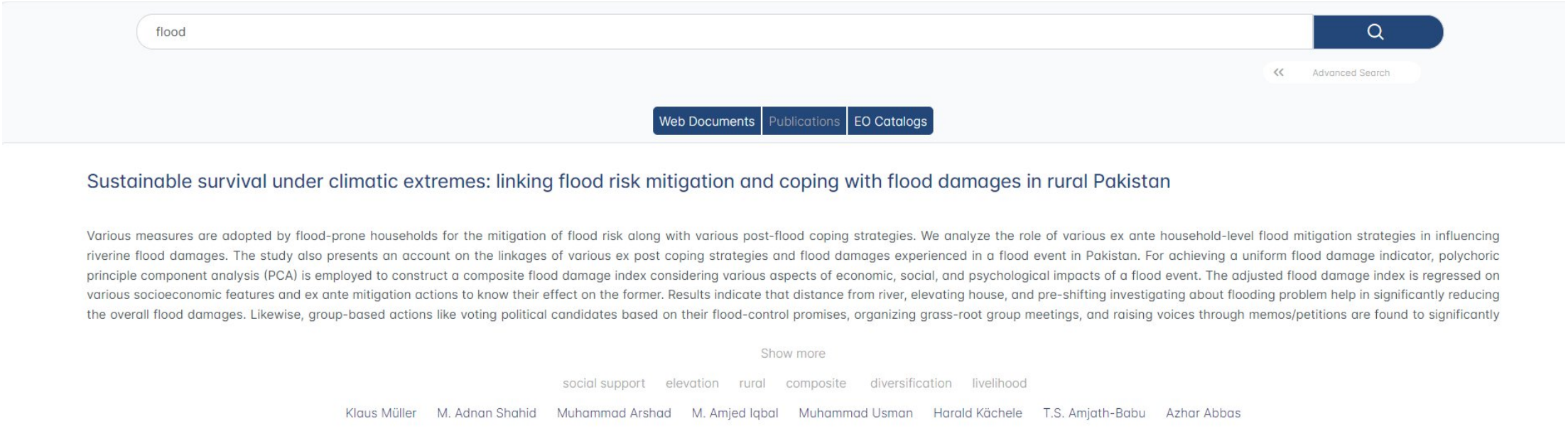
Open Science Search: DLR Use Case
2. Applications (1/2) – Science Search

A Hybrid Retrieval

- Combines indexed web data and knowledge graph and
- Query results from both pipelines are merged into
  - The user interface Within the Web Search tool



Open Science Search: DLR Use Case



Paper Honeder, J., El Baff, R., Hecking, T., Nussbaumer, A., & Guetl, C. (2025). **A Geo-Contextualized Multi-Genre Scientific Search Engine: A Novel Conceptual Design and Prototype Evaluation.** ICGDA'25 - Springer.

Science Search Prototype

Open Science Search: CSC Use Case

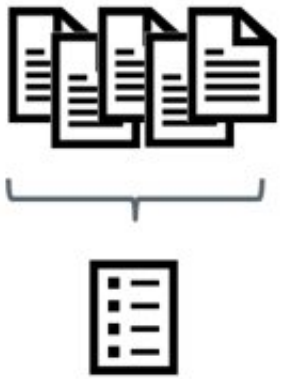
- Research.fi is a Finnish portal showcasing national research related outputs provided by the Ministry of Education and Culture and developed by CSC – IT Center for Science
  - It the external service of the Research Information Hub, which acts as a national aggregator of research-related data in Finland
- The service contains information on research conducted in Finland including publications, grants, organizations and infrastructures.
- **We have been studying how to utilize the Open Web Index to create our own multilingual science index and how this index could be useful for developing features to Research.fi**

Pipeline in coarse detail

Filter suitable websites

site	content	metadata
berkeley.edu	...	...
~~randomsite234.net~~	...	...
nasa.gov	...	...
~~marvel.com~~	...	...

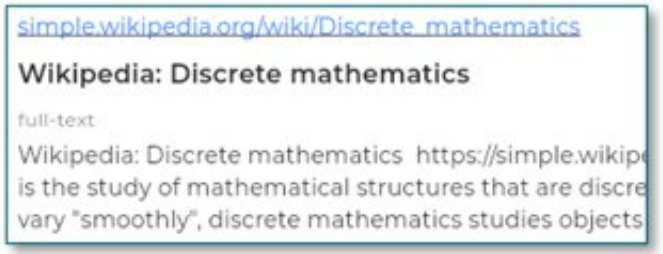
Build new index



Filtered parquet files

New .ciff index file

Search with MOSAIC



Mock example of search results

- Very simply put:
  - Suitable data is filtered from the daily crawls using tools developed in the project
  - Filtered data is used to build a new index suitable for use cases
  - The new index is queried with MOSAIC
- In the future, the filtering could be done already using the Open Web Index, without using the parquet files

A plan for a proof of concept

Downloading processed data

- Data crawled to the OWI is downloaded with the tool owilix

Filtering pages related to science

- Pages related to science are filtered by their Curlie labels

Preprocessing the text

- Unwanted text such as special symbols and metadata are removed

Filtering out low quality pages

- Pages are labeled by the quality of the text with a language model to remove e.g. message board threads

Enriching the data with scientific pages

- The index could be enriched with focused crawling on scientific pages

Building the "science index"

- The filtered pages are used to build a new index

Analysing search results from the new index

- The new index is being used for search using the tool MOSAIC and results are being analysed manually and with LLM summarizations

LLM

The provided URL leads to an IAEA (International Atomic Energy Agency) Data Portal where users can access a dataset focused on the "Assessment of Contamination in Agriculture."...

Open Science Search: CSC Use Case Next Steps

- Our scientific index does require fine tuning to improve its quality and reliability, but we have been planning on how to utilize it with Research.fi with plans including:
  - Expanding the search results of Research.fi with sites from the OWI, thus containing information of science and research conducted outside of Finland.
  - Creating a field of related content for entries present in Research.fi; e.g. using the OWI to identify research done using a specific grant.
  - Utilizing the OWI to measure and provide insights of the impact of research output in metrics other than citations.
  - Using a knowledge graph and topic modeling based on data from Research.fi to enhance queries made to the OWI.

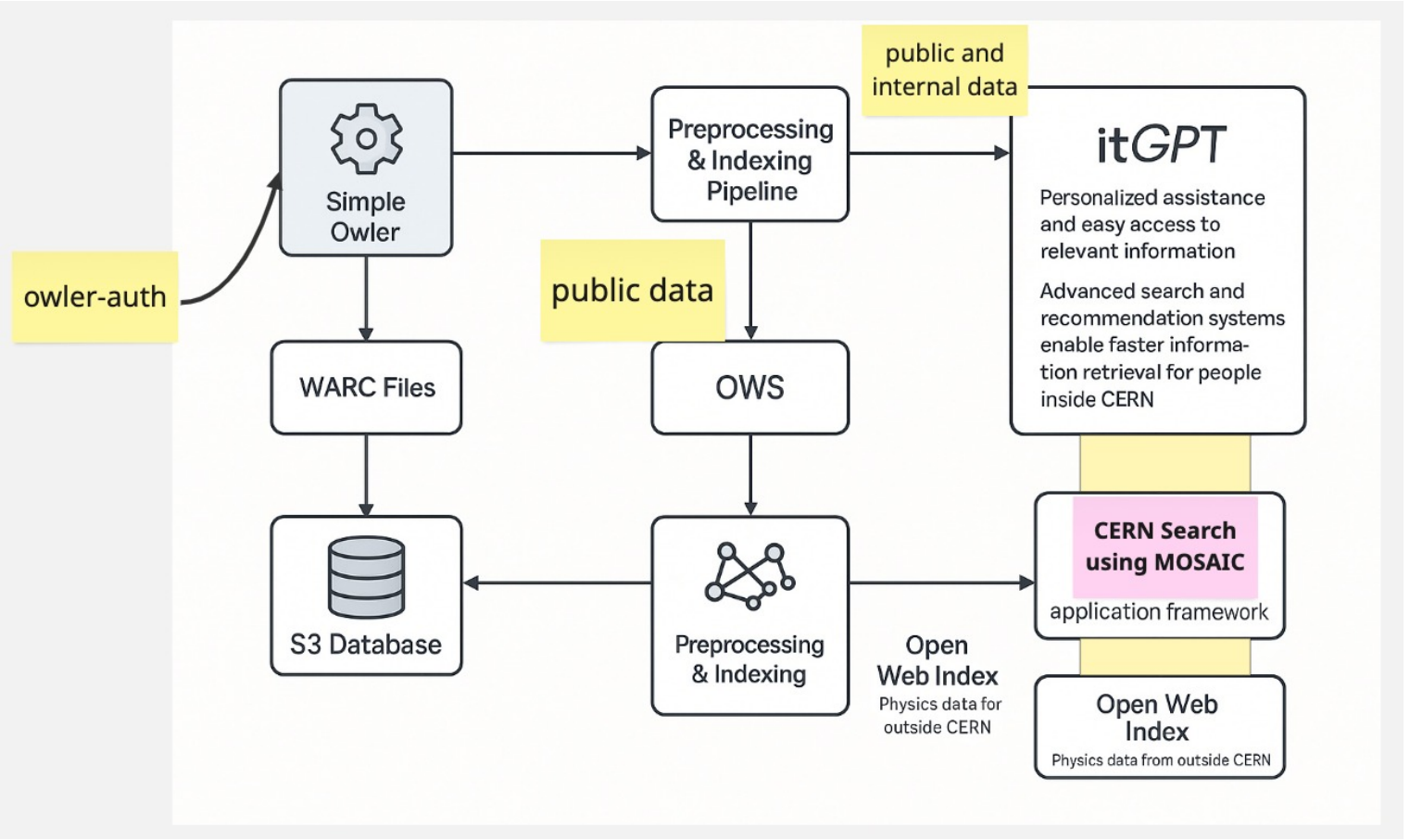
Open Science Search: CERN Use Case

Current goals:

- To be able to internally crawl some of CERN IT data and index it for internal use
- To be able to use MOSAIC as the UI of this data
- To be able to also provide this crawled data to ItGPT and CERN Search
- Dedicated nodes set-up for crawling activities and for deploying UI

Open Science Search: CERN Use Case

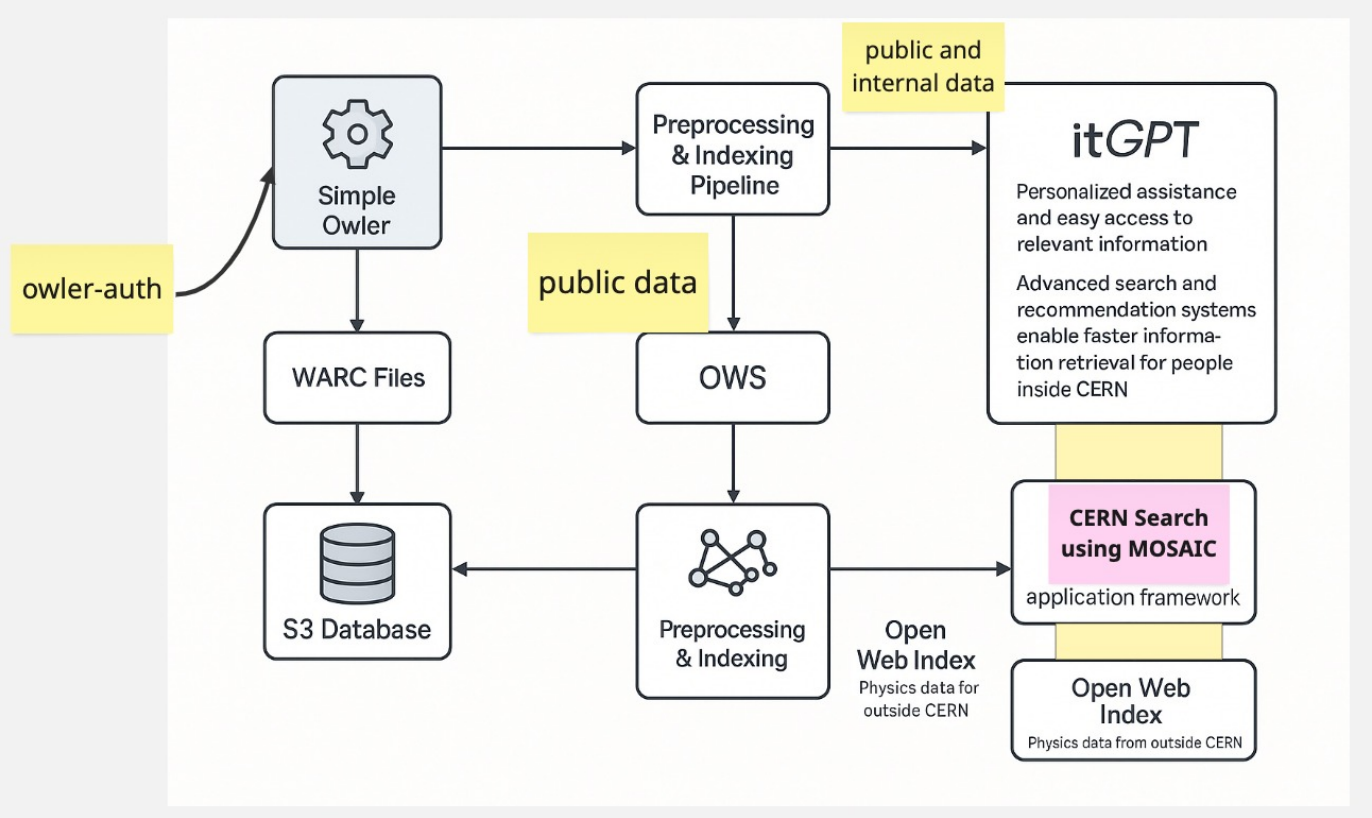
- Crawling and Storage
  - **Simple Owler** (One-time crawler using a seed URL list)
    - Outputs → **WARC files**
  - **WARC Files** sent to:
    - S3 Database
- Usage of WARC Files
  - Public WARC Files
    - → Preprocessing & Indexing Pipeline (OWS)
    - → Indexed Data sent **back to CERN**
    - → Used in **Mosaic** (search engine app framework)



Open Science Search: CERN Use Case

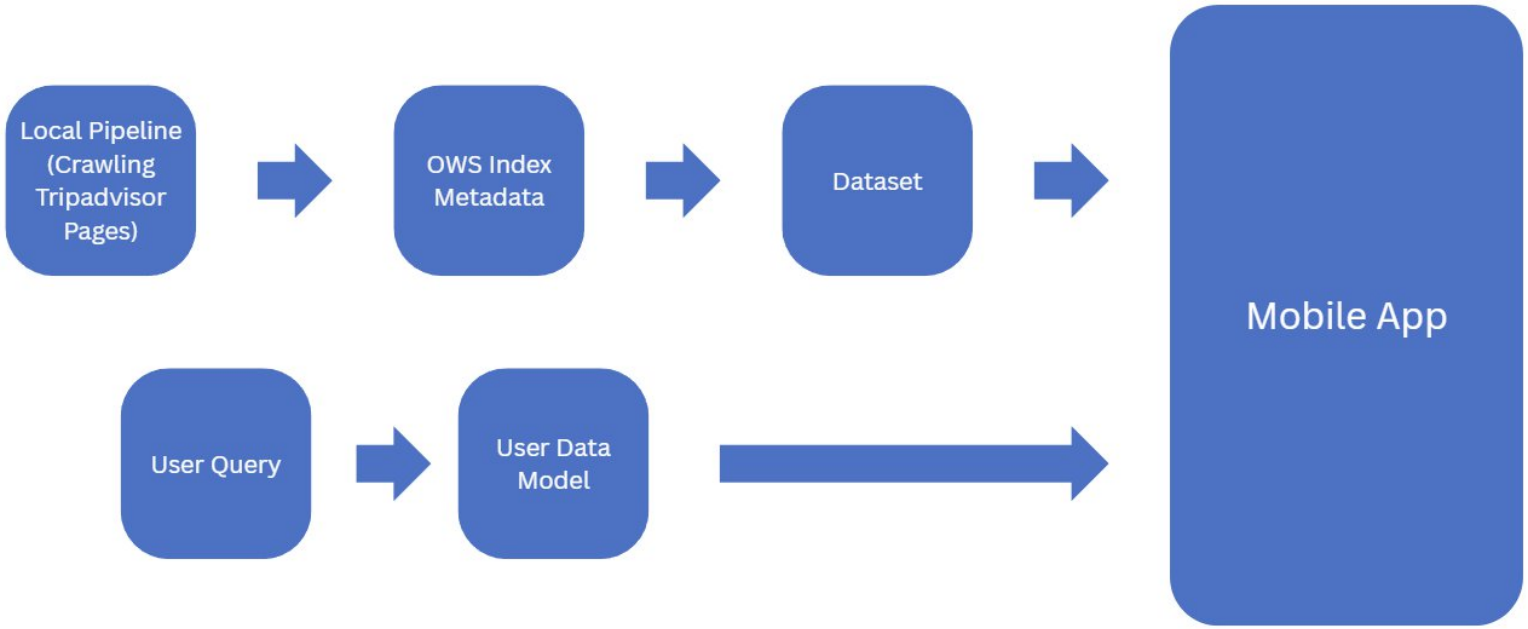
- **Private WARC Files** (Behind authentication)
  - → Stored at **CERN**
  - → (To be decided): Custom Preprocessing & Indexing
- All WARC Files
  - → **itGPT :** Provides personalized assistance and smart search within CERN

- MOSAIC
  - Displays: Indexed CERN data (from public WARC → OWS pipeline)
  - External physics data via **Open Web Index**



Mobile privacy-preserving, personalized recommendation of geo-entities: Pipeline Overview

- Privacy-preserving, content-based, user-centric recommendation system
- Suggests restaurants based on user queries, contextual information, and location.
- Three main components: OWS Index, backend server, and Android mobile app.
- Search string and Location is the only information that is collected from the user.

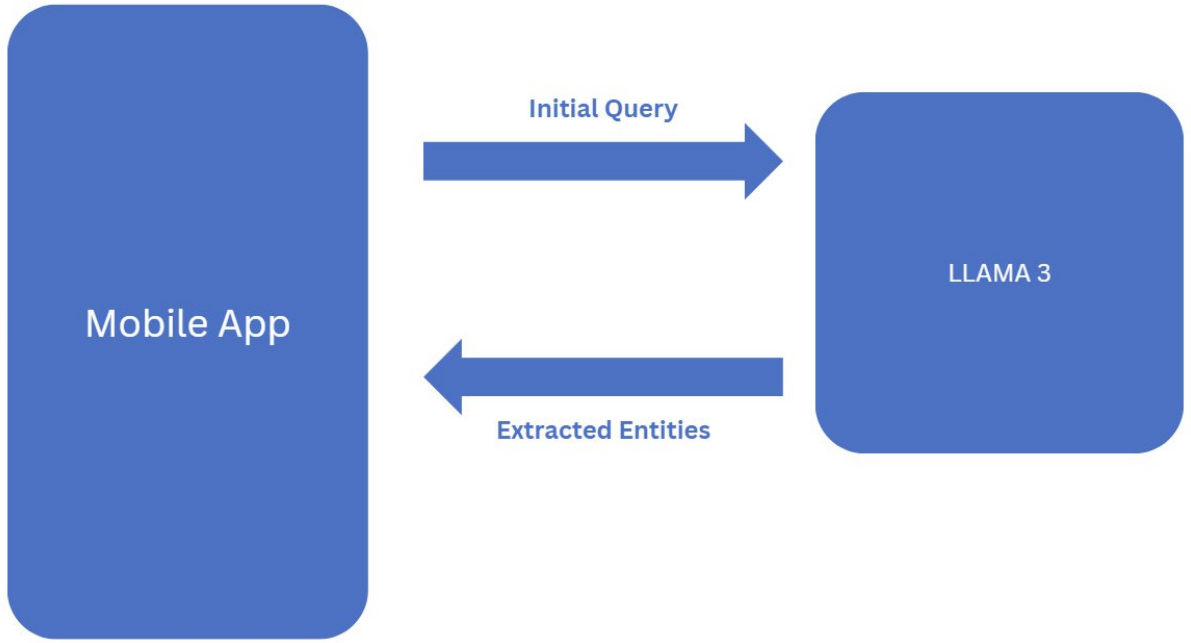


Mobile privacy-preserving, personalized recommendation of geo-entities: Index Building

- OWS Index metadata is used to build the dataset.
- Meta-Llama-3-8B LLM is used to extract & structure relevant information from “plain_text” of the pages.
- After post-processing dataset, saved in a parquet file, contains relevant information: name, address, url, cuisine, price range, meals, features, reviews etc.
- Bag of words column is created as a combination of all columns.

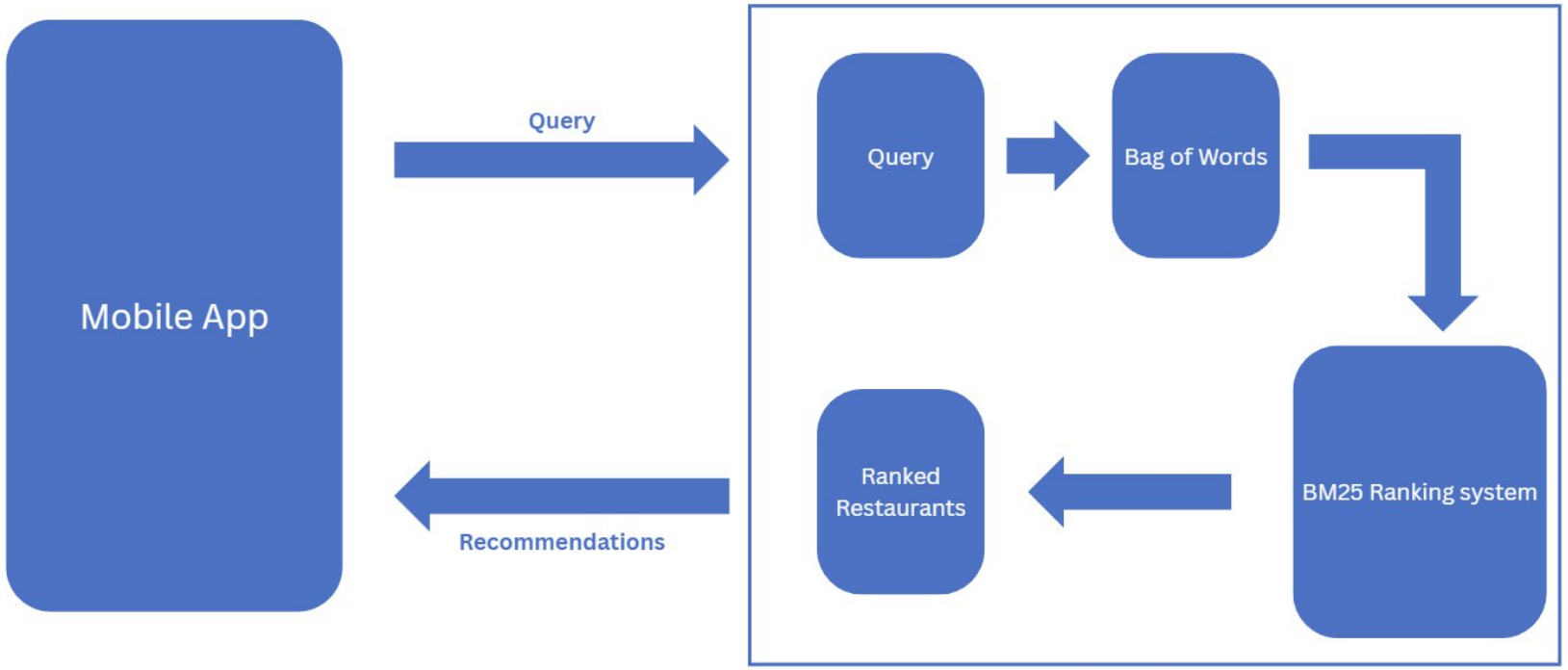
Mobile privacy-preserving, personalized recommendation of geo-entities: User Contextual Model

- User preferences are extracted from user queries and user interaction with the app (clicking restaurant cards)
- Meta-Llama-3-8B LLM is used to extract entities from initial user query: Cuisine, Meals, Price Range, Other Features
- These entities are saved inside user phone, not leaving the phone at any point
- User Model is build by complementing the initial user query with its preferences that are saved inside the phone



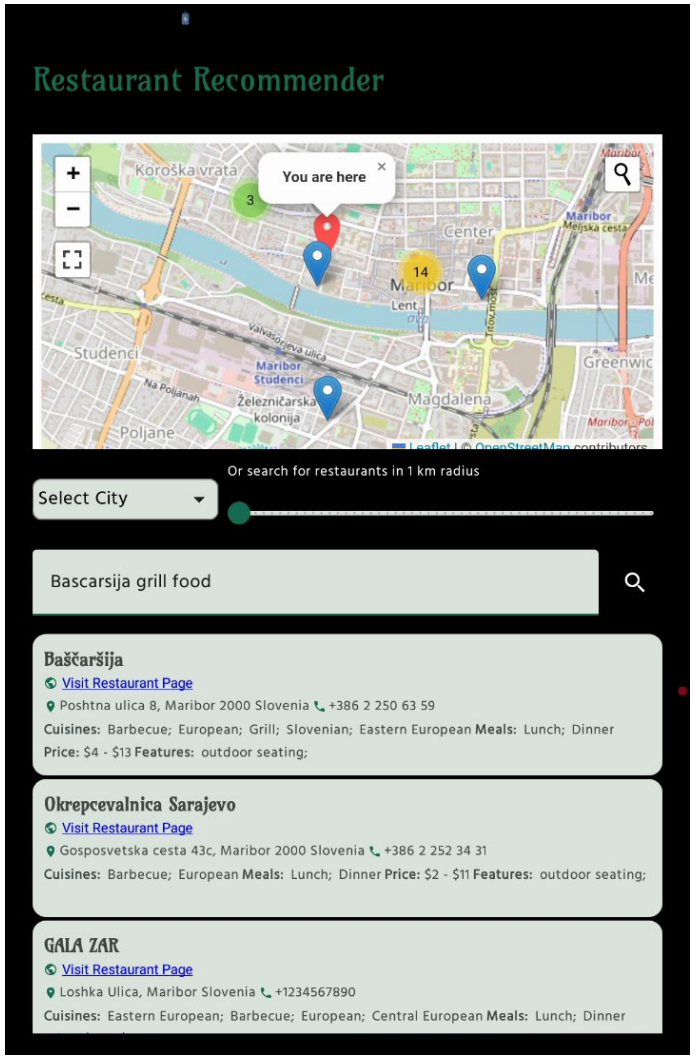
Mobile privacy-preserving, personalized recommendation of geo-entities: Getting Recommendations

- First restaurants are filtered by the location the user set (by city), or around the user's current location
- BM25 is used as a ranking system: ranks a set of documents based on the query terms appearing in each document



Mobile privacy-preserving, personalized recommendation of geo-entities: Mobile App Interface

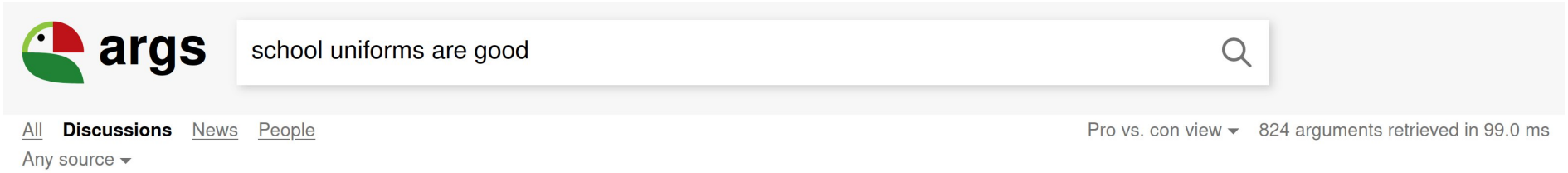
- Android app is developed using Kotlin.
- App uses SharedPreferences to store user data in an XML file inside the phone.
- App has an interactive map section powered by OpenStreetMaps.
- Interactive map is implemented using Leaflet.js



Mobile privacy-preserving, personalized recommendation of geo-entities: Next Steps

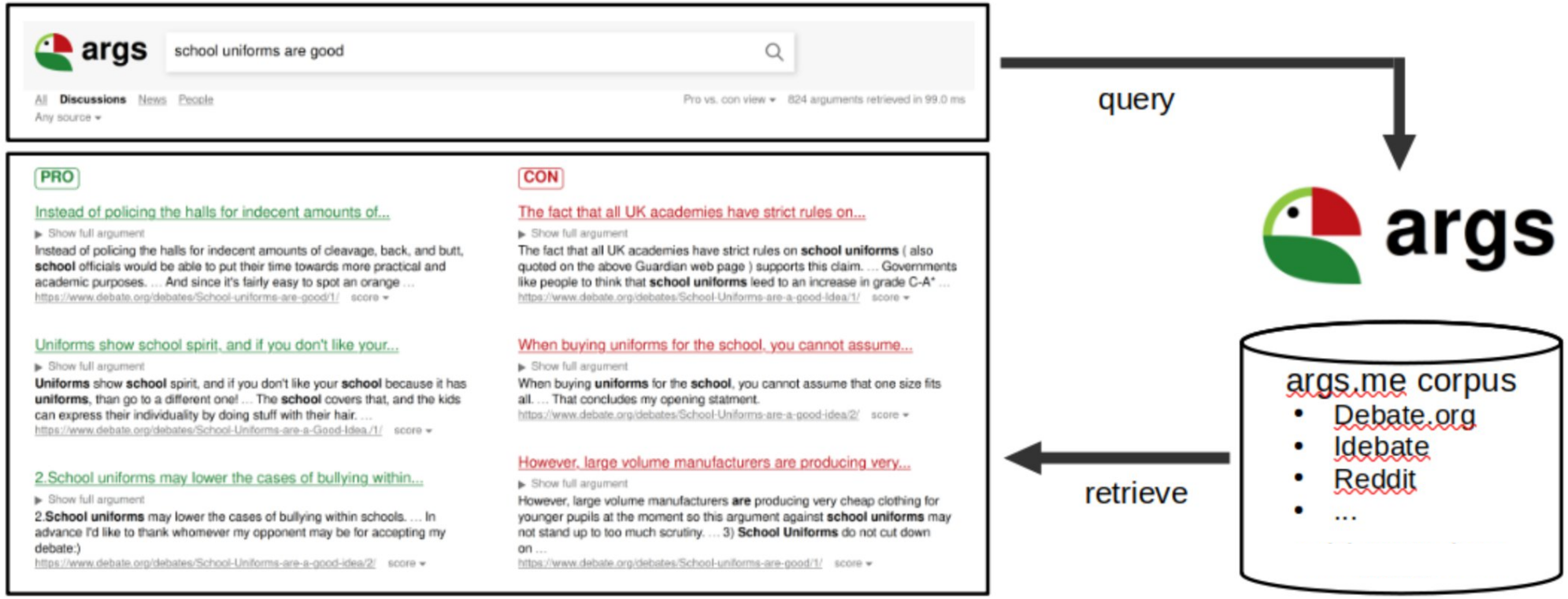
- improve the similarity measures and personalization of responses with rating
- extend the app to be multilingual
- integrate multi-source datasets with a semi-automated pipeline
- experiment with Large Vision Language Models to generate additional (contextual) information about the restaurant and food, and retrain the models

Argumentation Search

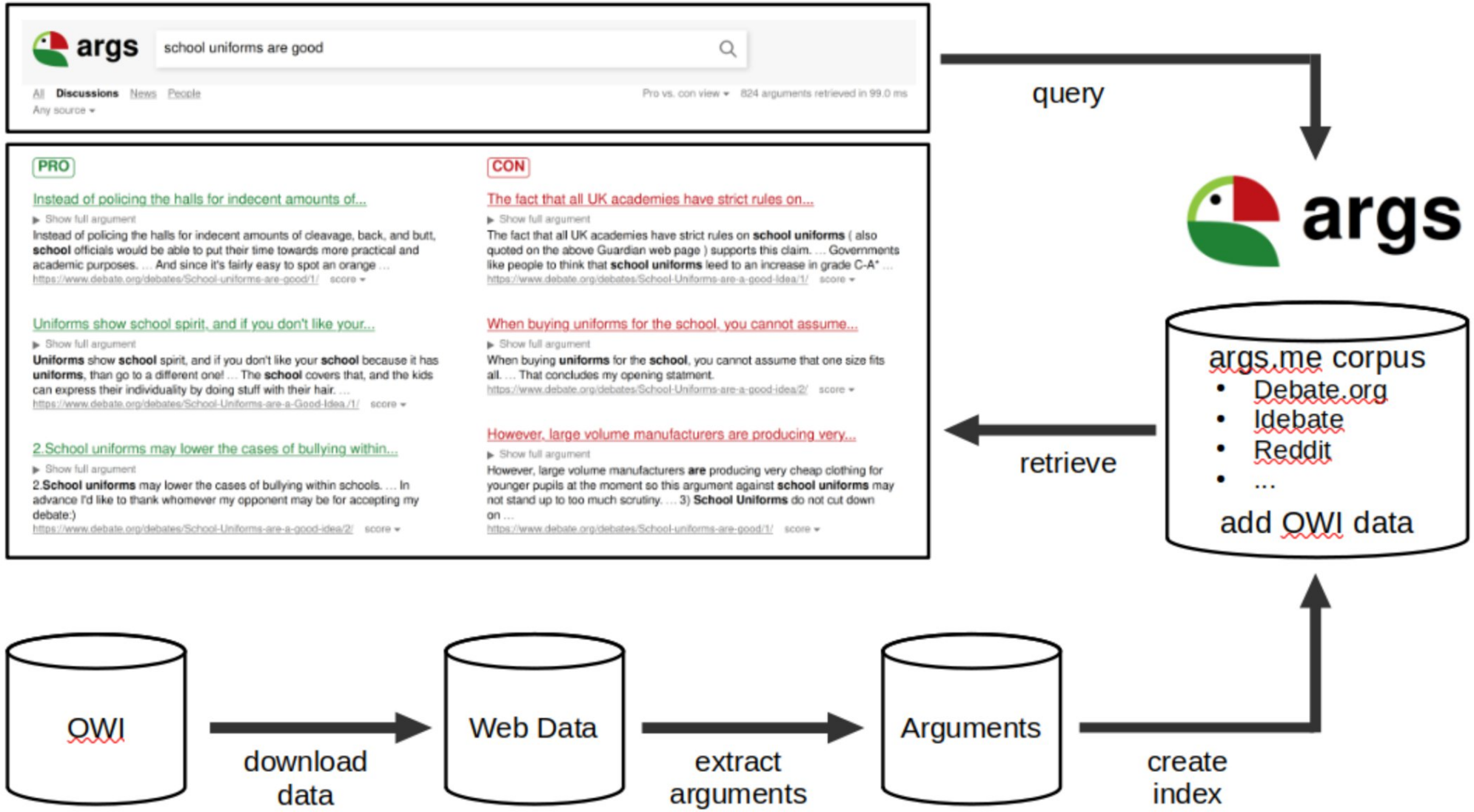


www.args.me

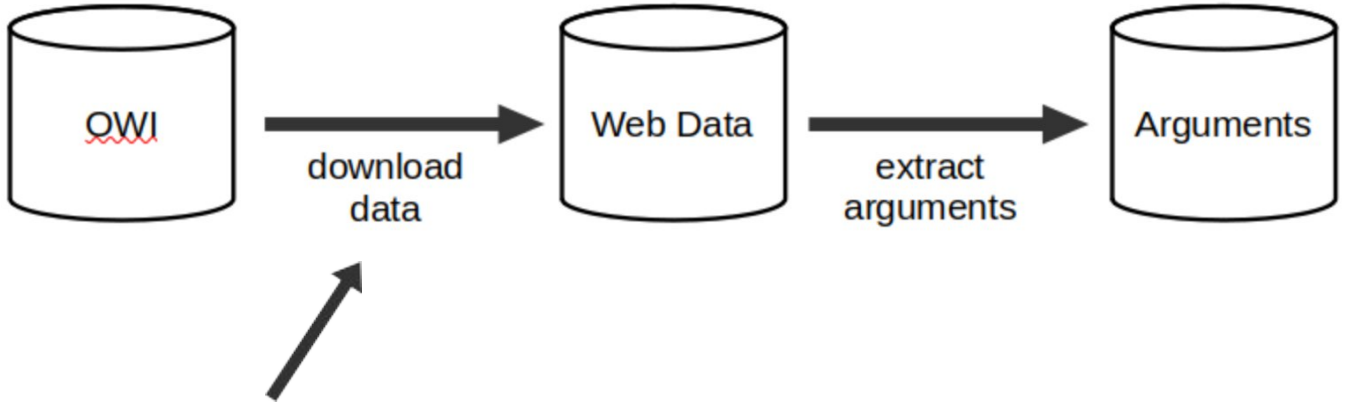
Argumentation Search



Argumentation Search

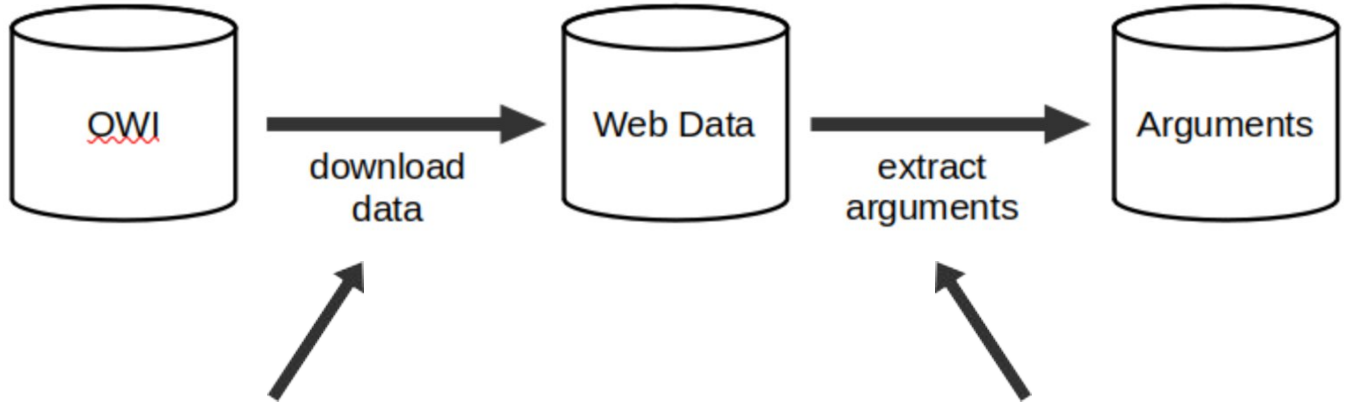


Argumentation Search



- Filter for forums / blogs
- Filter by curlielabels, e.g.:
  - News / Analysis and Opinion
  - Society / Lifestyle Choices / Homeschooling / Vegetarianism
  - Society / Issues / Abortion / Animal Welfare

Argumentation Search



- Filter for forums / blogs
- Filter by curlielabels, e.g.:
  - News / Analysis and Opinion
  - Society / Lifestyle Choices / Homeschooling / Vegetarianism
  - Society / Issues / Abortion / Animal Welfare

- Create text segments
  - paragraphs
  - sentences
- Filter non-argumentative
  - argument classification model
- Cluster similar arguments

Questions and Comments

Contact

- **Alexander Nussbaumer** <alexander.nussbaumer@tugraz.at>
- **Roxanne El Baff** <Roxanne.ElBaff@dlr.de>
- **Jason Theodoropoulos** <jason.theodoropoulos@csc.fi>
- **Noor Afshan Fathima** <noor.afshan.fathima@cern.ch>
- **Izidor Mlakar** <izidor.mlakar@amis.net>
- **Ines Zelch** <ines.zelch@uni-jena.de>