



Creating Search Applications Using the Open Web Index

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

Webinar
7 April 2025

Open Web Search



Funded by
the European Union

SUPPORTED
BY

NGI

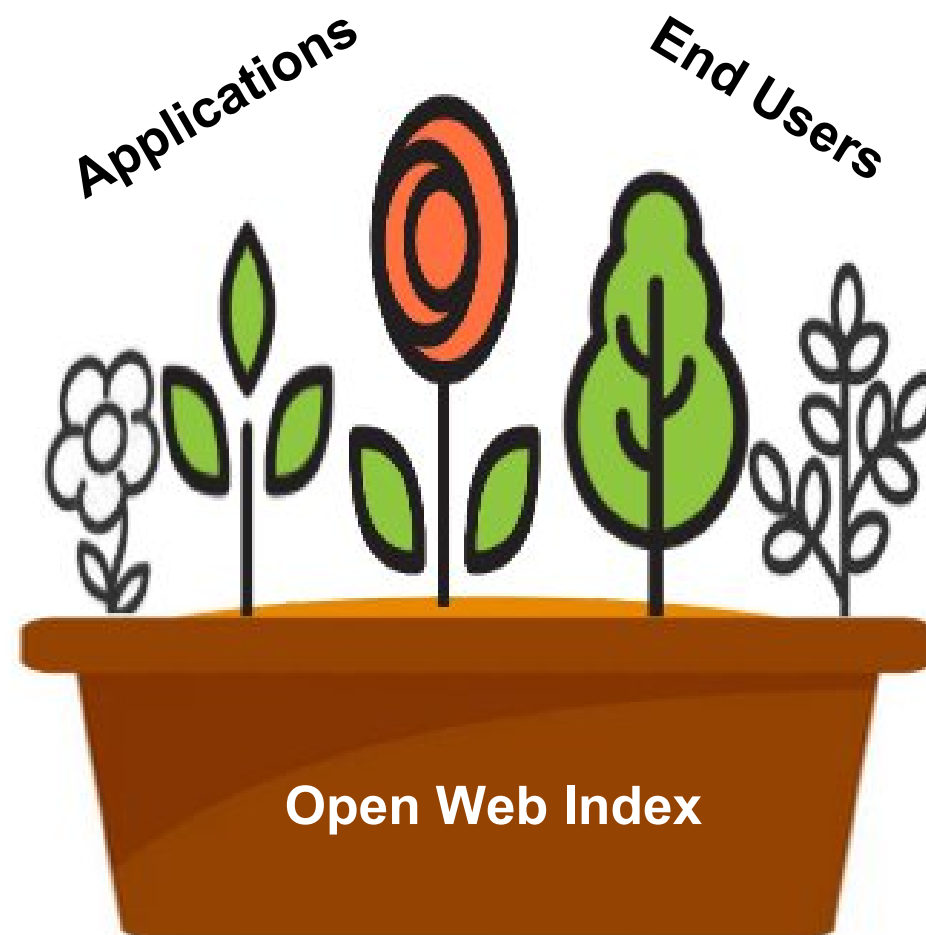
Overview and Agenda

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:

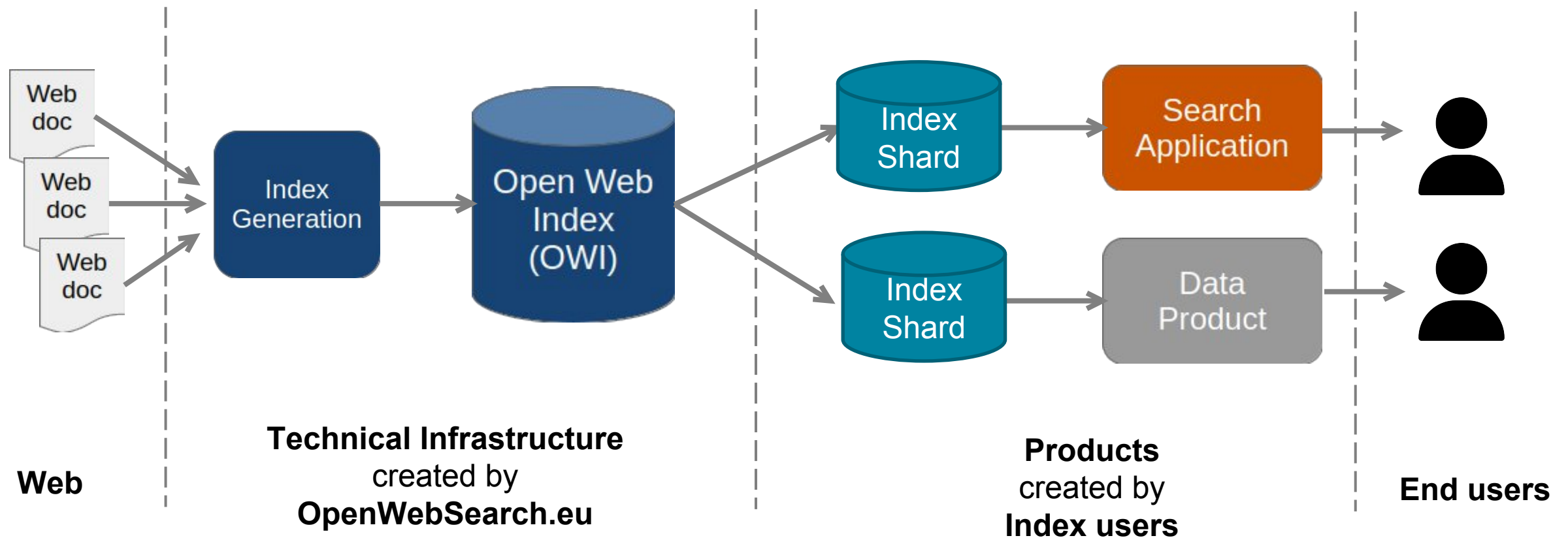
Concept of Search Applications

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

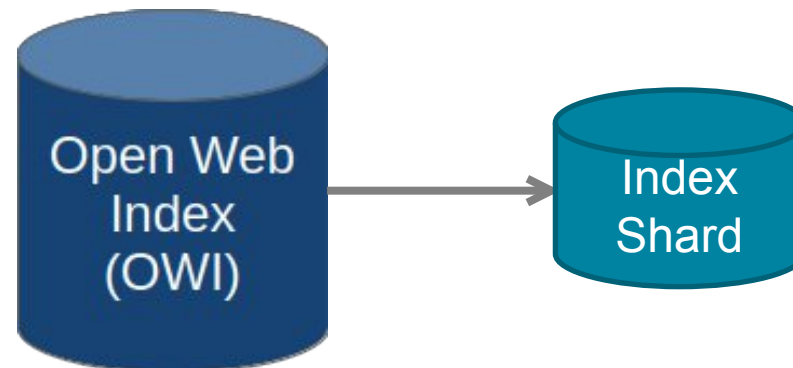
Search Applications in OpenWebSearch.eu

- 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)

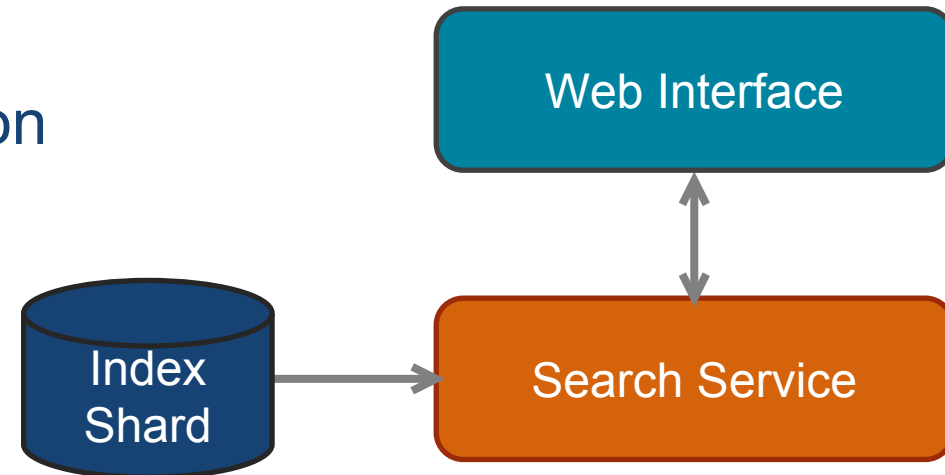
Document	Lang	Full text	WARC date	Location	<Curly topic>
docID A	eng	... text ...	2023-07-01		
docID B	deu	... text ...	2023-07-01		
docID C	deu	... text ...	2023-07-01		
docID D	eng	... text ...	2023-07-01		

Parquet File: Metadata

- full text
- language
- geo-information
- micro-data
- topic (curly label)
- links
- embeddings
- ...

MOSAIC

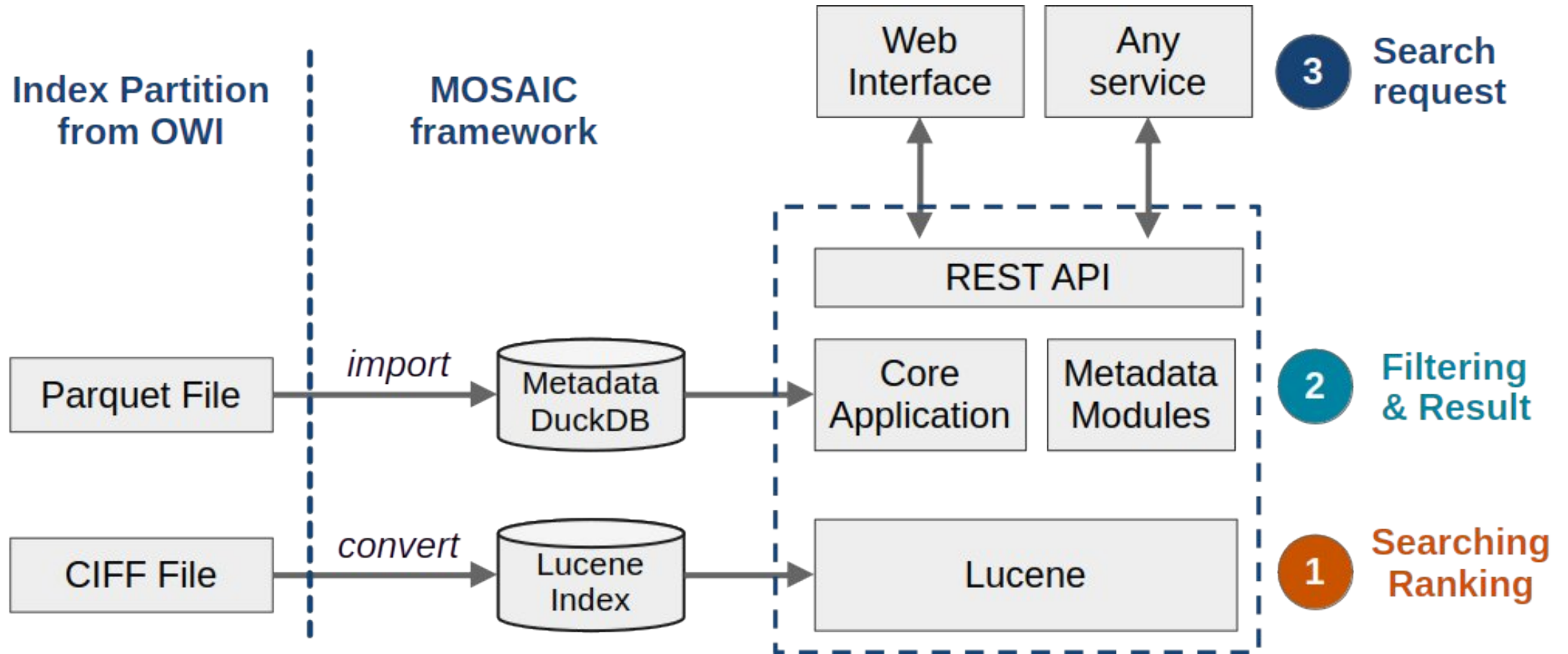
- **M**odular **S**earch **A**pplication based on **I**ndex **F**ract**I**ons
- Generic implementation of an OWS.eu vertical search engine
 - Demonstration of the concept of an OWS.eu vertical engine
 - Out-of-the-box search engine
 - Toolbox for an own search application
- Uses index shards from the 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

Search term:

Location filter

Geo Filter: West: East: North: South:

Language filter

Index:	Language:	Limit:	Keyword:
<input type="radio"/> default / all	<input type="radio"/> default / all	<input checked="" type="radio"/> default / 20	<input type="text"/>
<input checked="" type="radio"/> Demo SimpleWiki	<input checked="" type="radio"/> English	<input type="radio"/> 10 items	
<input type="radio"/> Demo Graz Universities	<input type="radio"/> German	<input type="radio"/> 50 items	
<input type="radio"/> DLR Prototype		<input type="radio"/> 1,000,000	

Index selection

Search URL: <https://qnode.eu/ows/mosaic/service/search?q=cern&index=demo-simplewiki&lang=eng&west=1.8&east=17.0&north=55.6&south=40.2>

Text snippet

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

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

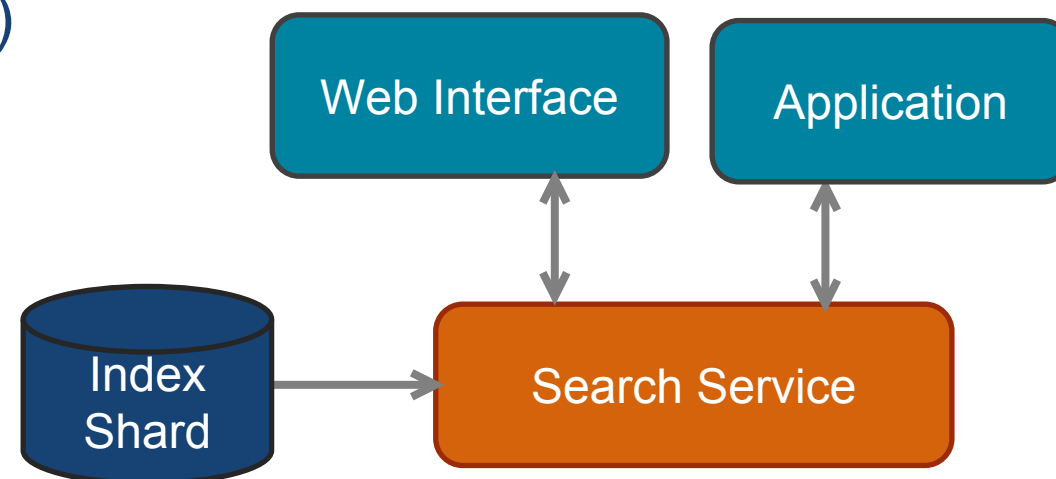
Locations: *Geneva • Switzerland •*

Keywords:

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

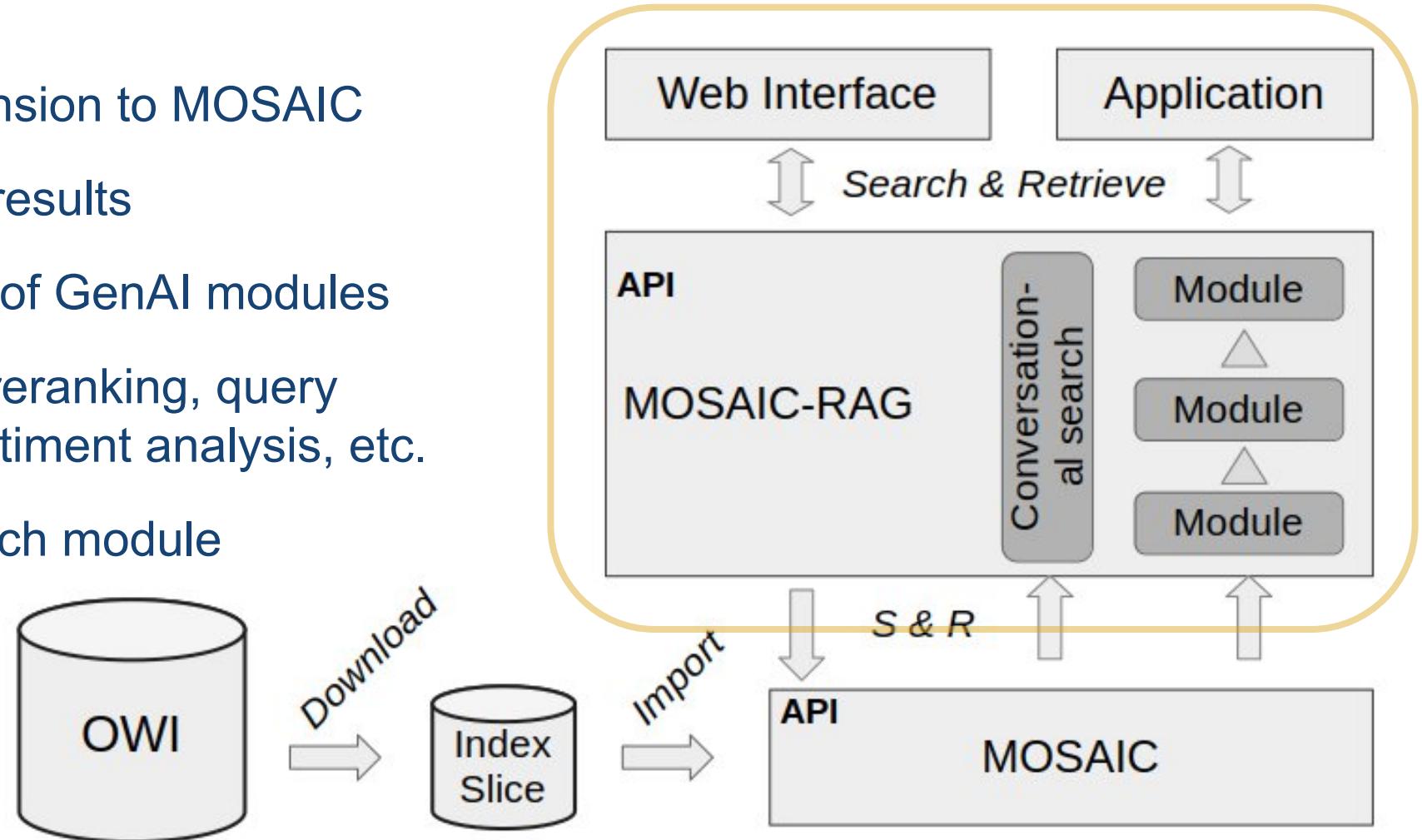
Creation of a Search Application

- 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

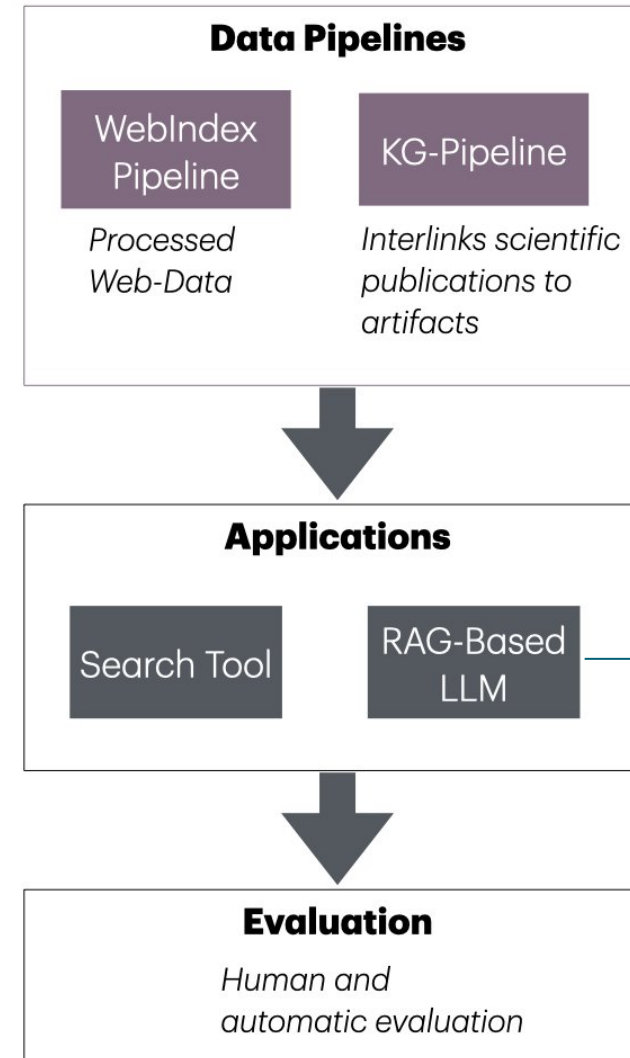
- Creating a WebIndex from web-data (using [owilix](#)).
- Creating a Knowledge graph (KG) to interlink different scientific text genres.

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.*

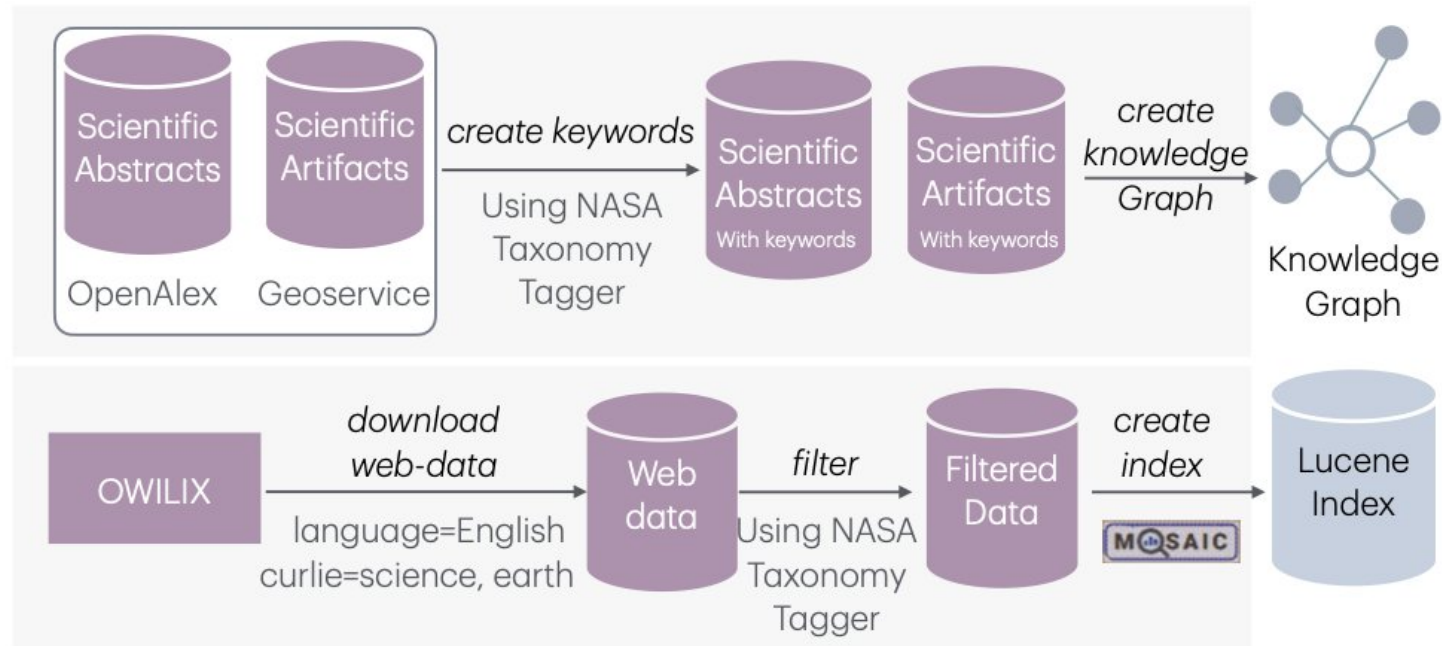
KG-Pipeline

- 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)

Data Pipelines

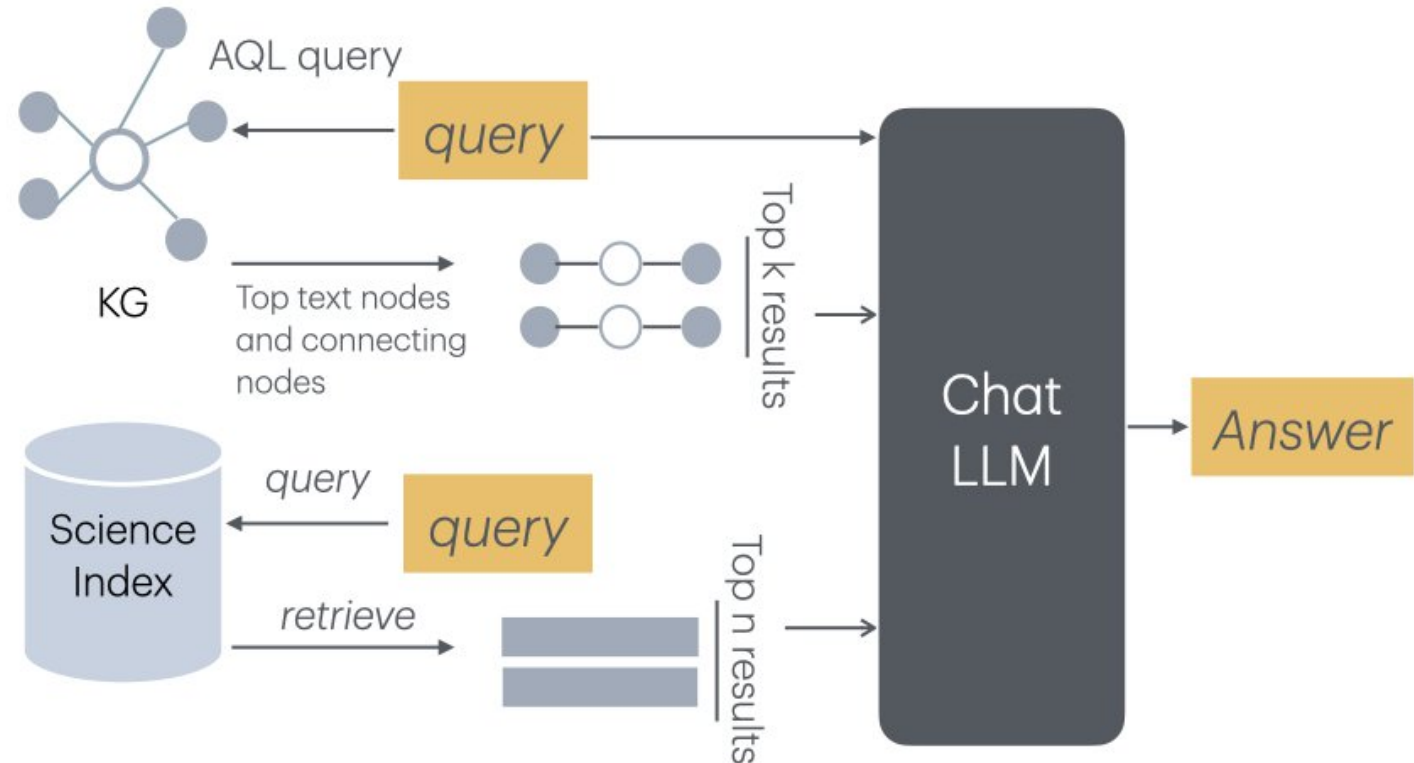


Open Science Search: DLR Use Case

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

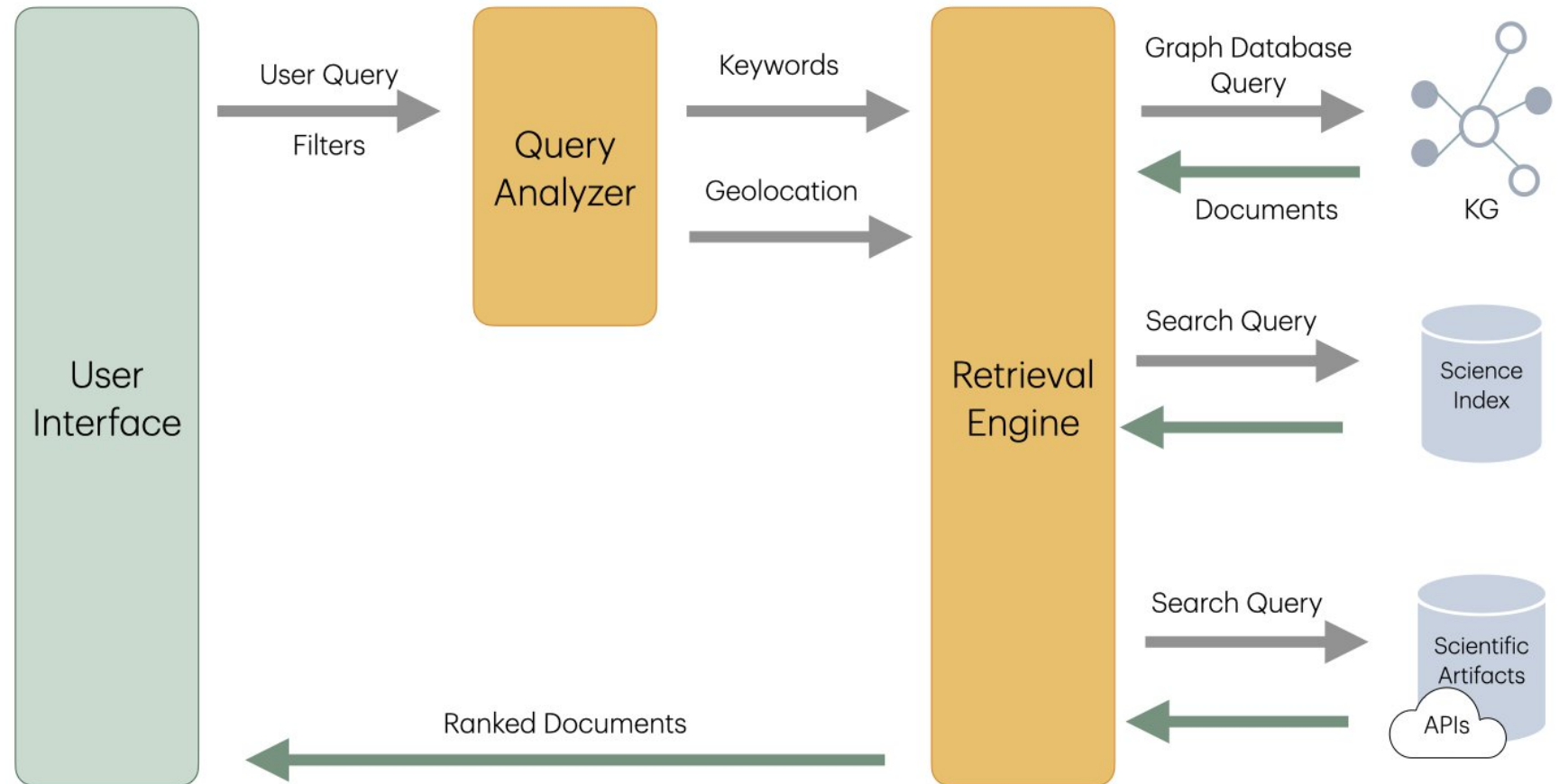


Open Science Search: DLR Use Case

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

flood



<< Advanced Search

Web Documents | Publications | EO Catalogs

Sustainable survival under climatic extremes: linking flood risk mitigation and coping with flood damages in rural Pakistan

Various measures are adopted by flood-prone households for the mitigation of flood risk along with various post-flood coping strategies. We analyze the role of various ex ante household-level flood mitigation strategies in influencing riverine flood damages. The study also presents an account on the linkages of various ex post coping strategies and flood damages experienced in a flood event in Pakistan. For achieving a uniform flood damage indicator, polychoric principle component analysis (PCA) is employed to construct a composite flood damage index considering various aspects of economic, social, and psychological impacts of a flood event. The adjusted flood damage index is regressed on various socioeconomic features and ex ante mitigation actions to know their effect on the former. Results indicate that distance from river, elevating house, and pre-shifting investigating about flooding problem help in significantly reducing the overall flood damages. Likewise, group-based actions like voting political candidates based on their flood-control promises, organizing grass-root group meetings, and raising voices through memos/petitions are found to significantly

Show more

social support elevation rural composite diversification livelihood

Klaus Müller M. Adnan Shahid Muhammad Arshad M. Amjed Iqbal Muhammad Usman Harald Kächele T.S. Amjath-Babu Azhar Abbas

Flood Mapping and Flood Dynamics of the Mekong Delta: ENVISAT-ASAR-WSM Based Time Series Analyses

Satellite remote sensing is a valuable tool for monitoring flooding. Microwave sensors are especially appropriate instruments, as they allow the differentiation of inundated from non-inundated areas, regardless of levels of solar illumination or frequency of cloud cover in regions experiencing substantial rainy seasons. In the current study we present the longest synthetic aperture radar-based time series of flood and inundation information derived for the Mekong Delta that has been analyzed for this region so far. We employed overall 60 Envisat ASAR Wide Swath Mode data sets at a spatial resolution of 150 meters acquired during the years 2007–2011 to facilitate a thorough understanding of the flood regime in the Mekong Delta. The Mekong Delta in southern Vietnam comprises 13 provinces and is home to 18 million inhabitants. Extreme dry seasons from late December to May and wet seasons from June to December characterize people's rural life. In this study, we show which areas of the delta are frequently affected by floods and which regions remain dry all year round. Furthermore, we present which areas are flooded at which frequency and elucidate the

Show more

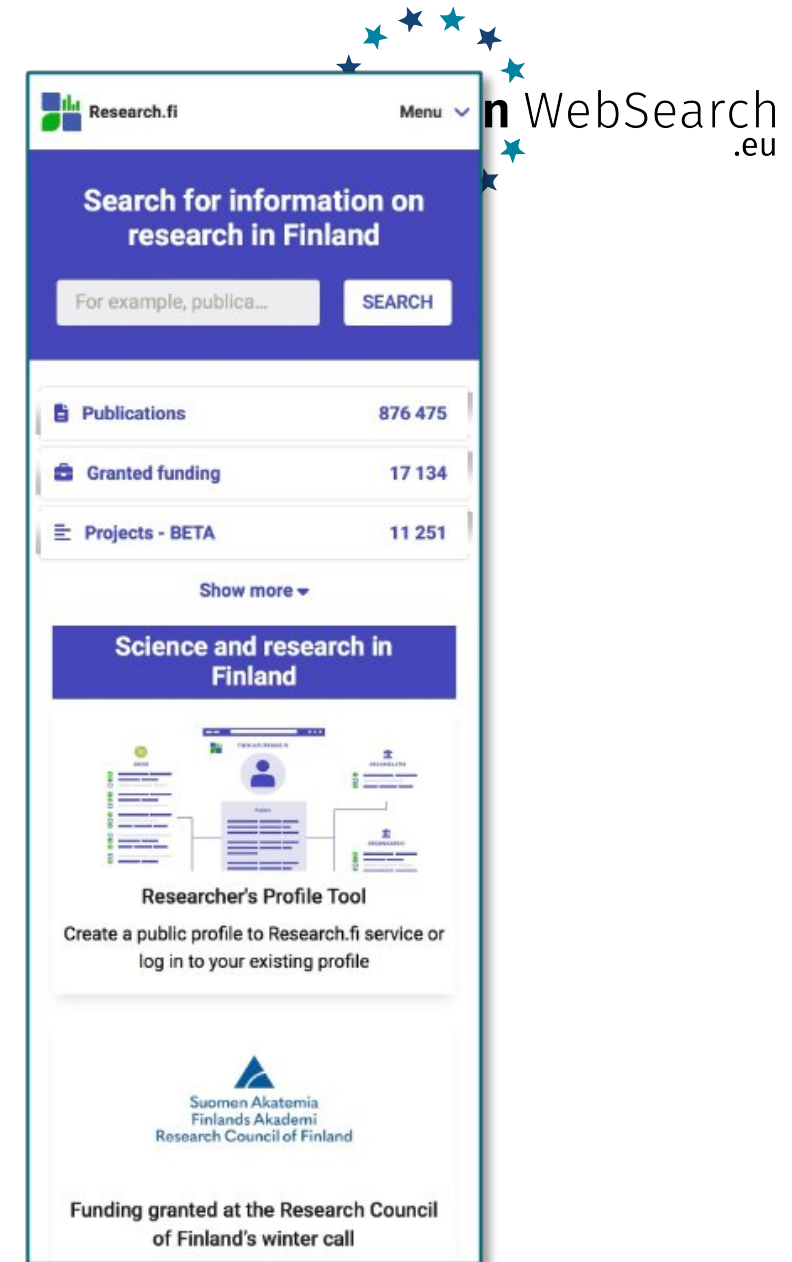
vietnam mekong delta time series feature extraction wsm asar envisat radar inundation water detection flood progression flood dynamics flood

Stefan Dech Xinwu Li Patrick Leinenkugel Juliane Huth Huadong Guo Claudia Künzer

Normalized Difference Flood Index for rapid flood mapping: taking advantage of EO big data

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 is 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**



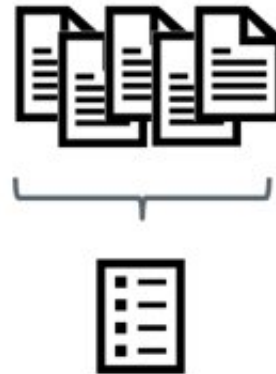
Pipeline in coarse detail

Filter suitable websites

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



Build new index



Filtered
parquet files

New .ciff
index file



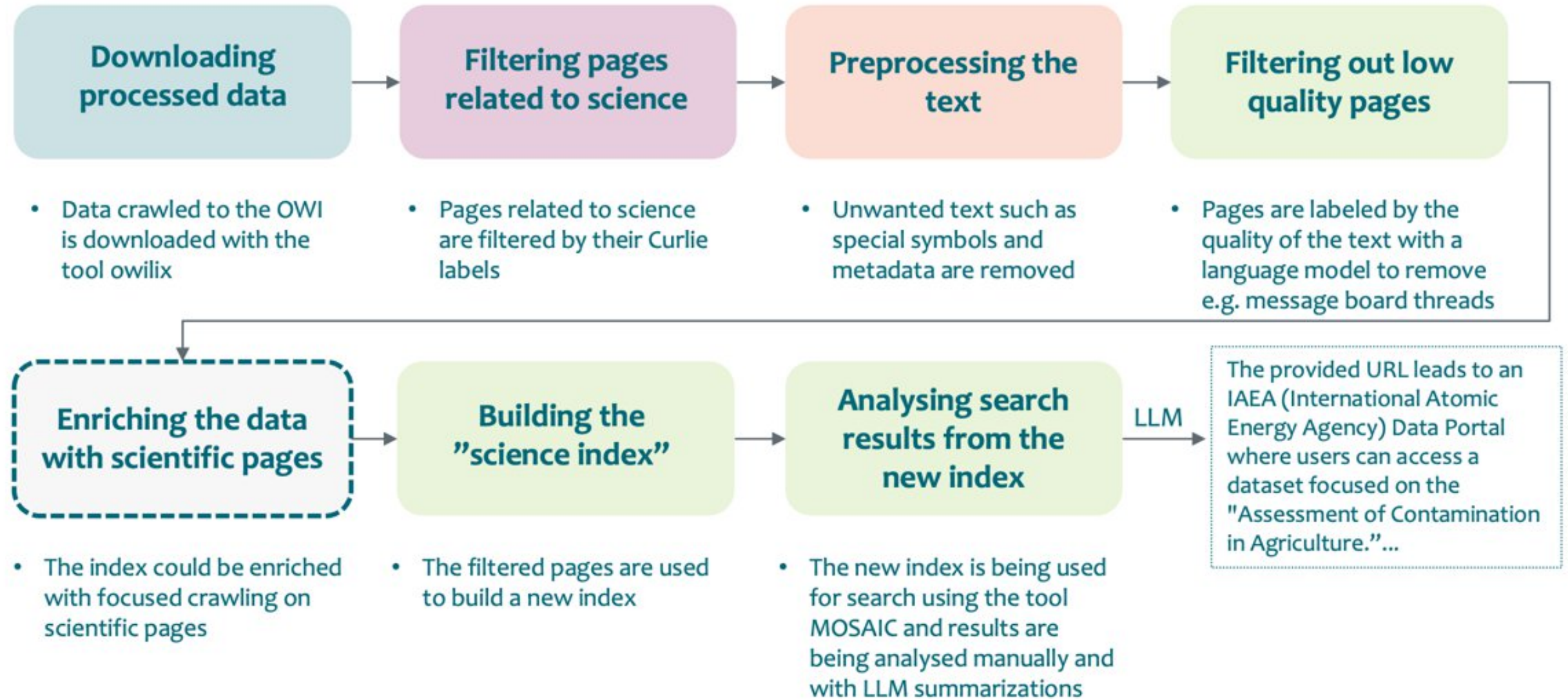
Search with MOSAIC

simple.wikipedia.org/wiki/Discrete_mathematics
Wikipedia: Discrete mathematics
 full-text
 Wikipedia: Discrete mathematics https://simple.wikiped
 is the study of mathematical structures that are discre
 vary "smoothly", discrete mathematics studies objects

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



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:

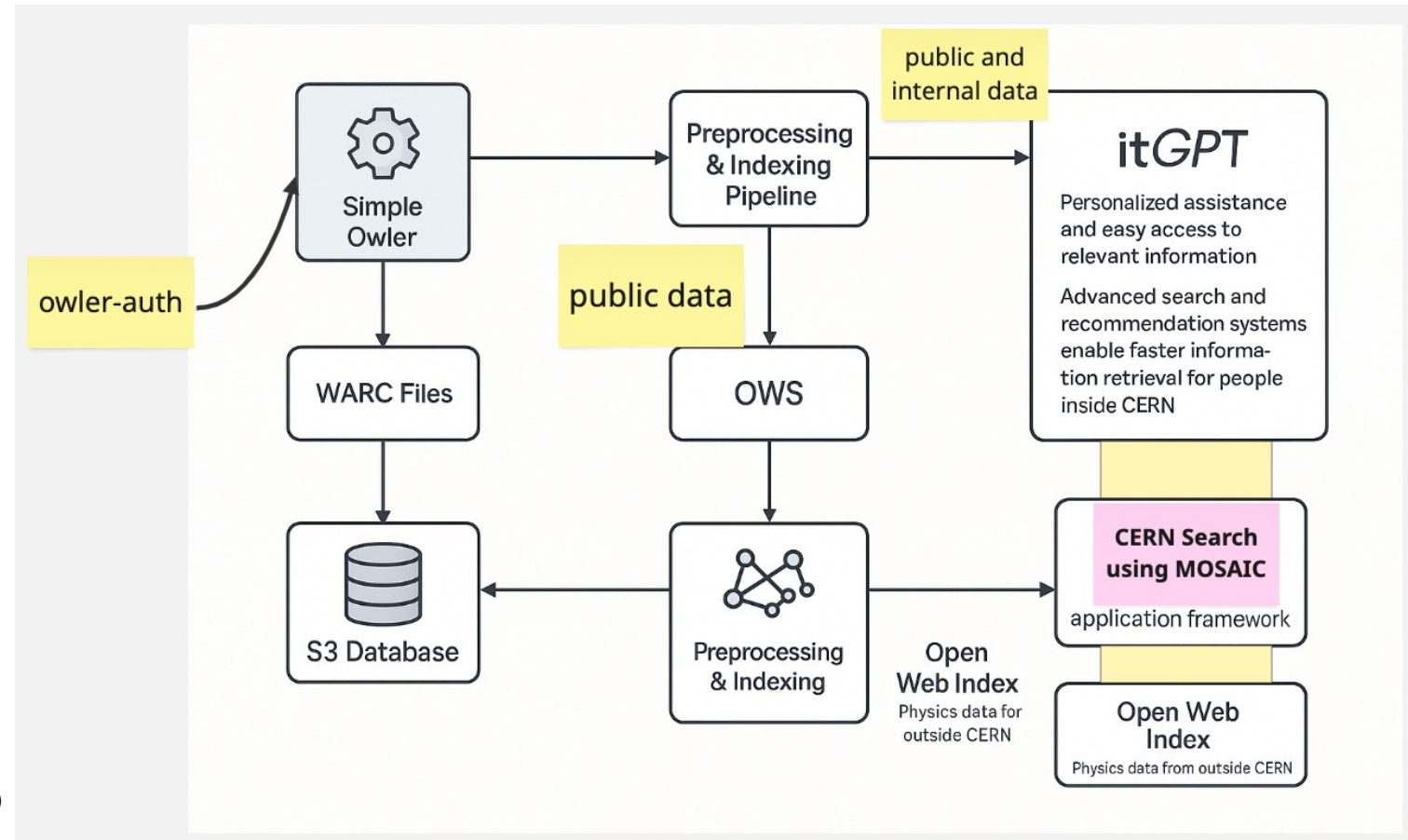
- To be able to internally crawl some of CERN IT data and index it for internal use
- To be able to use MOSAIC as the UI of this data
- To be able to also provide this crawled data to ItGPT and CERN Search
- Dedicated nodes set-up for crawling activities and for deploying UI

Open Science Search: CERN Use Case



• Crawling and Storage

- **Simple Oowler** (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)

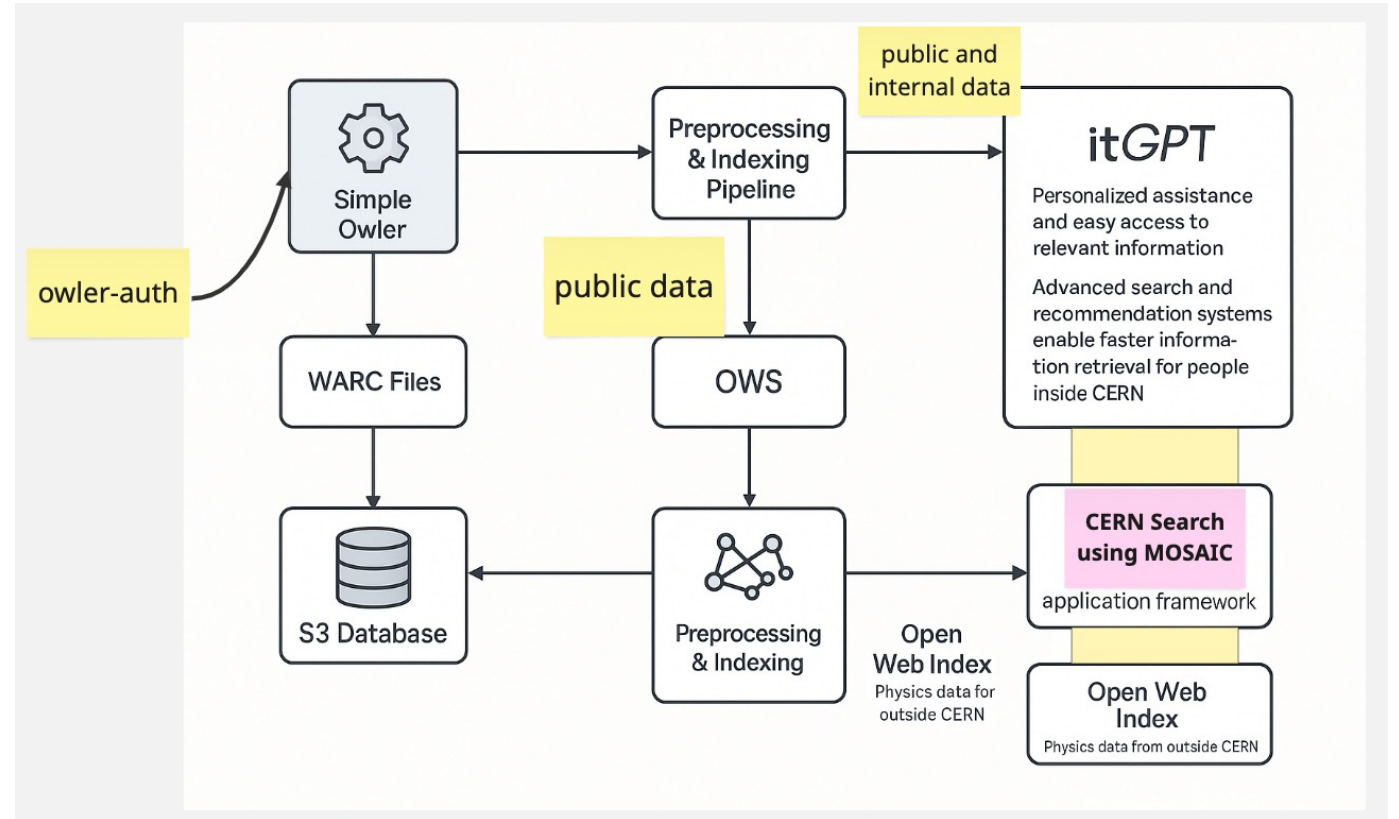


Open Science Search: CERN Use Case

- **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

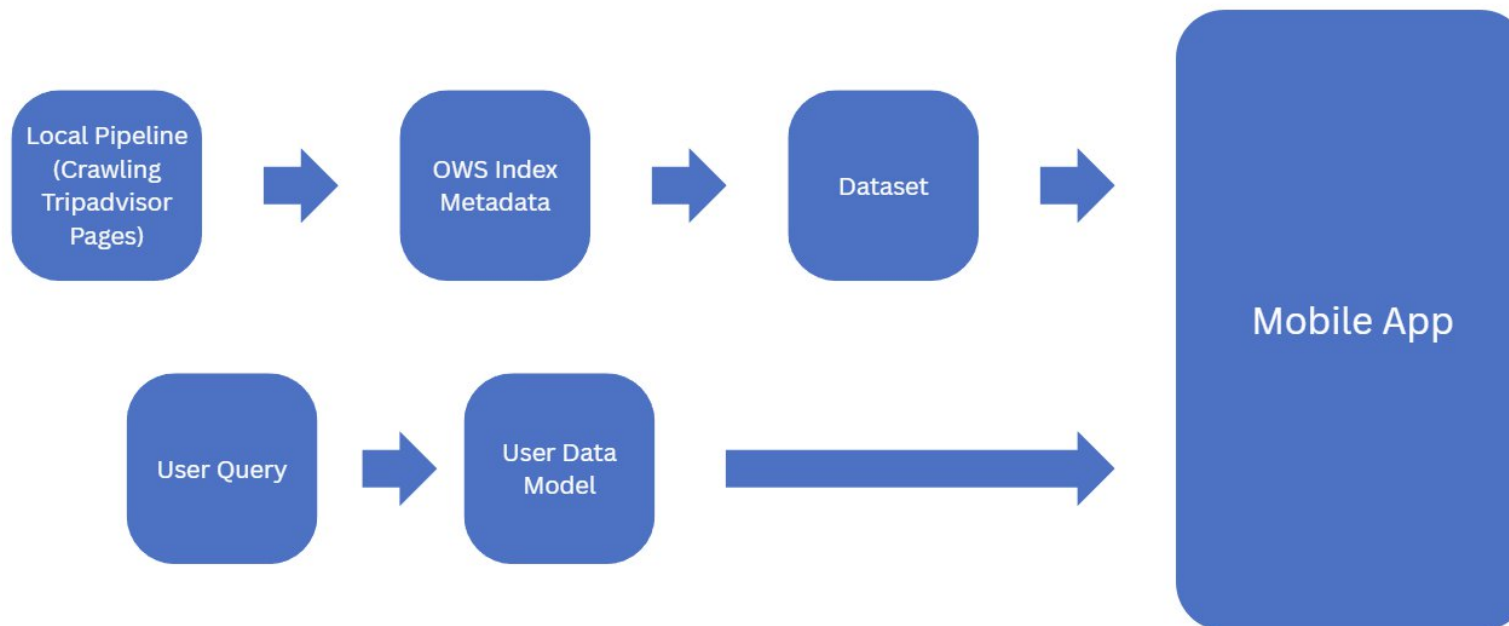
• MOSAIC

- 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.



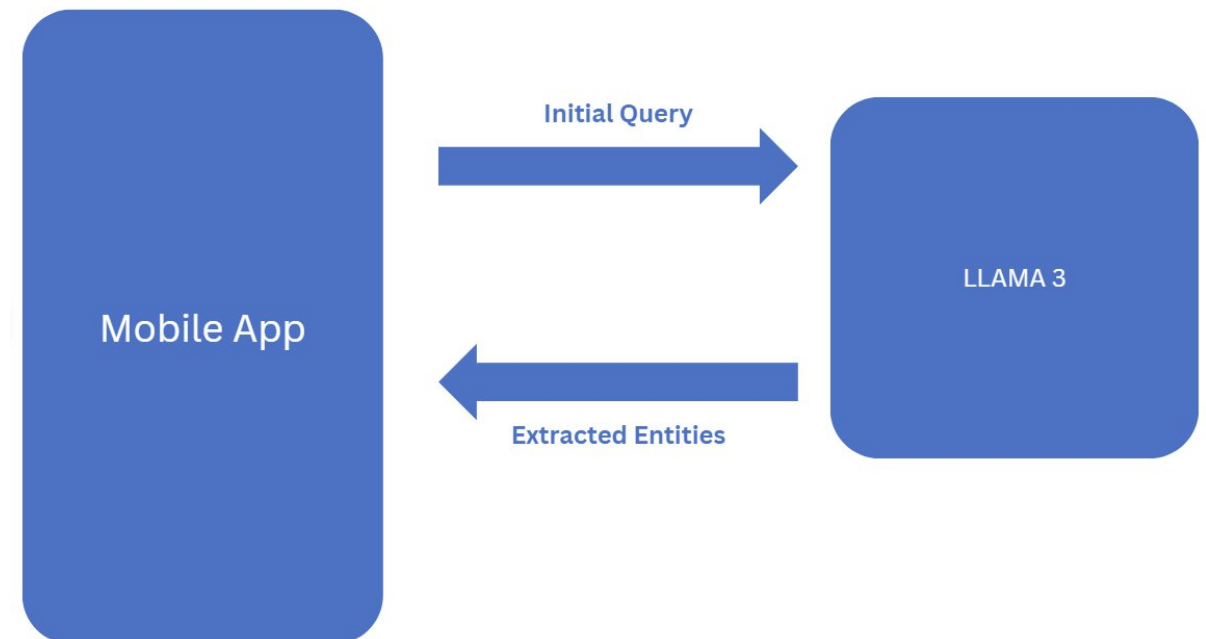
Mobile privacy-preserving, personalized recommendation of geo-entities: Index Building

- 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.



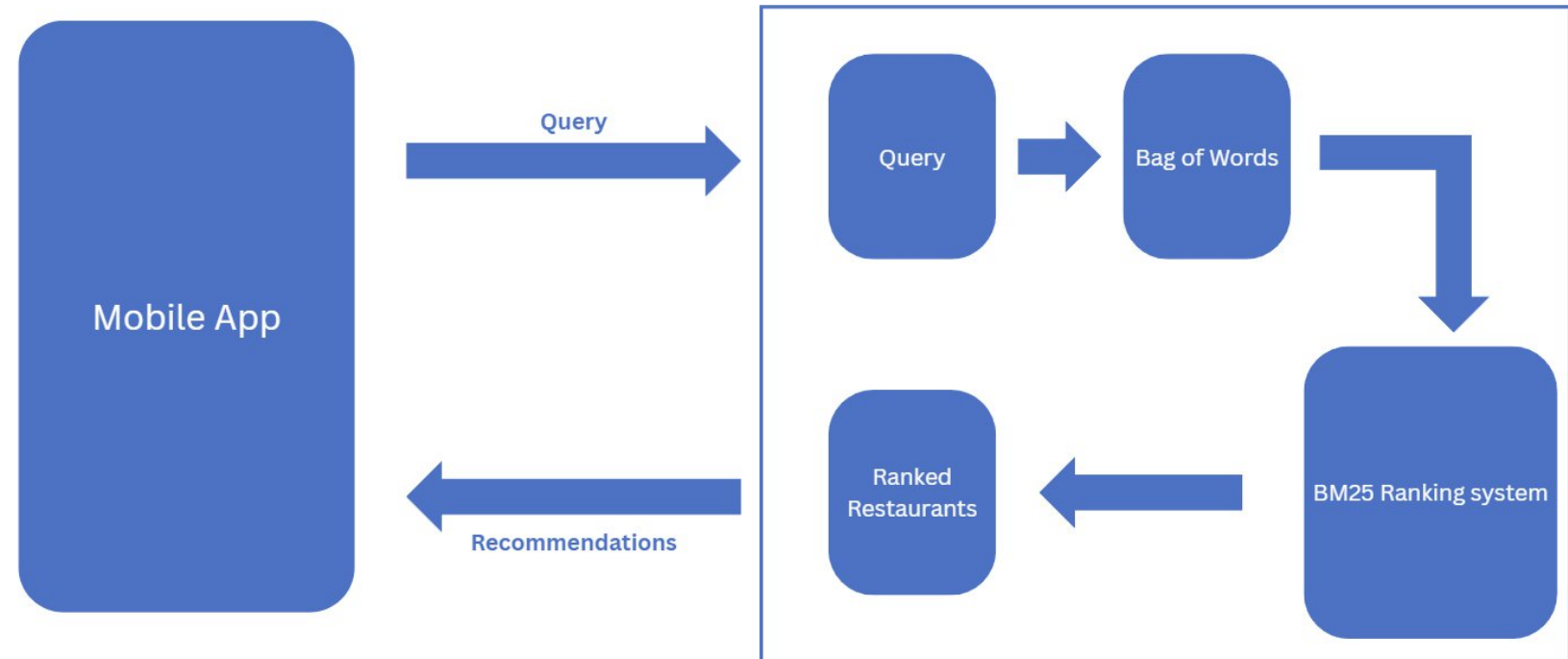
Mobile privacy-preserving, personalized recommendation of geo-entities: User Contextual Model

- 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



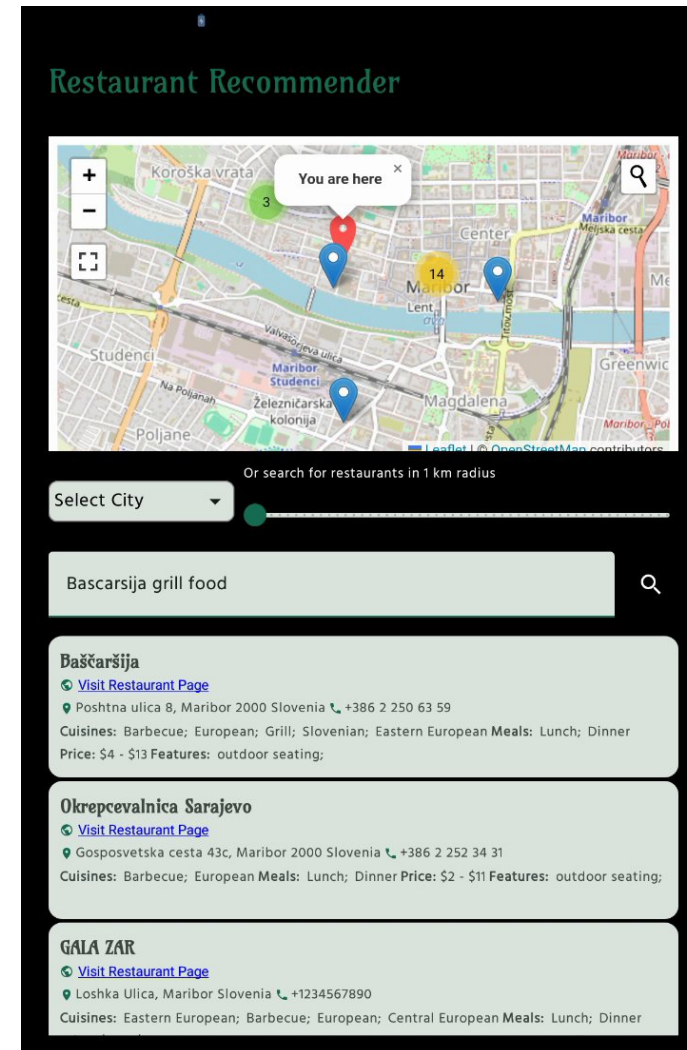
Mobile privacy-preserving, personalized recommendation of geo-entities: Getting Recommendations

- 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



school uniforms are good



All **Discussions** [News](#) [People](#)

Any source ▾

Pro vs. con view ▾ 824 arguments retrieved in 99.0 ms

PRO

[Instead of policing the halls for indecent amounts of...](#)

► Show full argument

Instead of policing the halls for indecent amounts of cleavage, back, and butt, **school** officials would be able to put their time towards more practical and academic purposes. ... And since it's fairly easy to spot an orange ...

<https://www.debate.org/debates/School-uniforms-are-good/1/> score ▾

[Uniforms show school spirit, and if you don't like your...](#)

► Show full argument

Uniforms show **school** spirit, and if you don't like your **school** because it has **uniforms**, than go to a different one! ... The **school** covers that, and the kids can express their individuality by doing stuff with their hair. ...

<https://www.debate.org/debates/School-Uniforms-are-a-Good-Idea./1/> score ▾

[2.School uniforms may lower the cases of bullying within...](#)

► Show full argument

2.**School uniforms** may lower the cases of bullying within schools. ... In advance I'd like to thank whomever my opponent may be for accepting my

CON

[The fact that all UK academies have strict rules on...](#)

► Show full argument

The fact that all UK academies have strict rules on **school uniforms** (also quoted on the above Guardian web page) supports this claim. ... Governments like people to think that **school uniforms** lead to an increase in grade C-A* ...

<https://www.debate.org/debates/School-Uniforms-are-a-good-Idea/1/> score ▾

[When buying uniforms for the school, you cannot assume...](#)

► Show full argument

When buying **uniforms** for the **school**, you cannot assume that one size fits all. ... That concludes my opening statment.

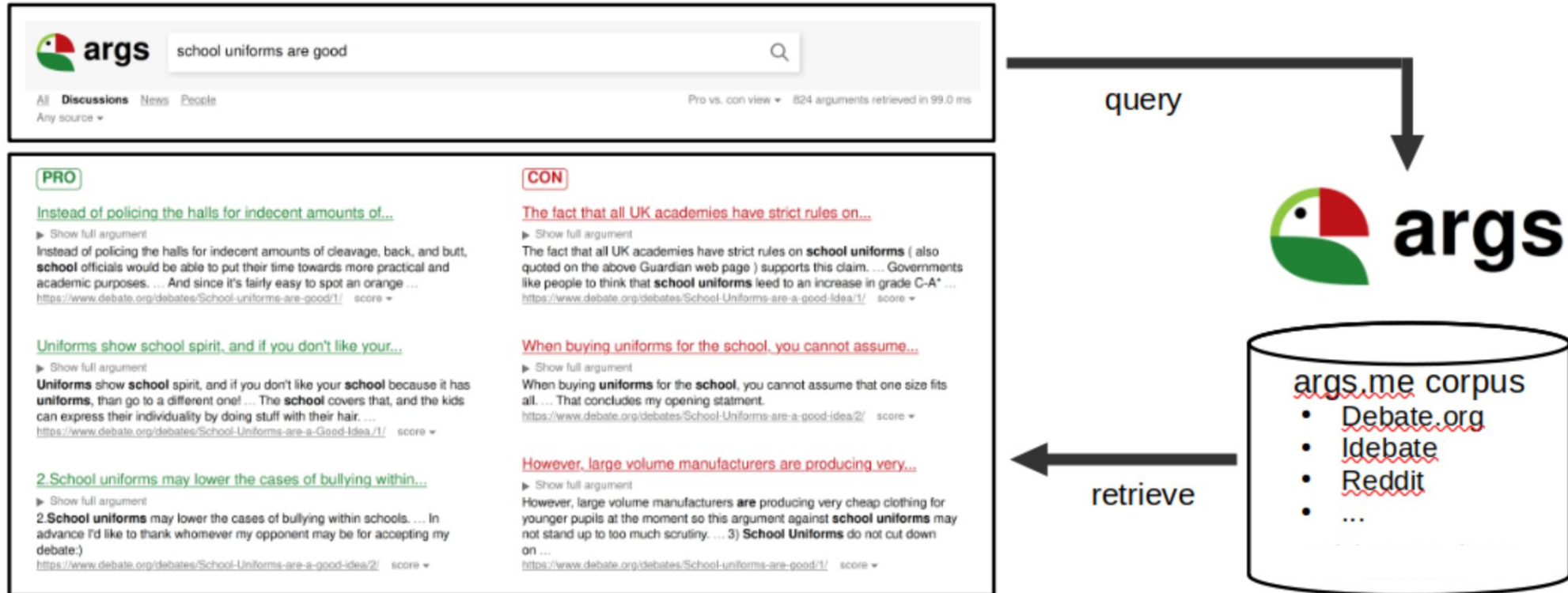
<https://www.debate.org/debates/School-Uniforms-are-a-good-idea/2/> score ▾

[However, large volume manufacturers are producing very...](#)

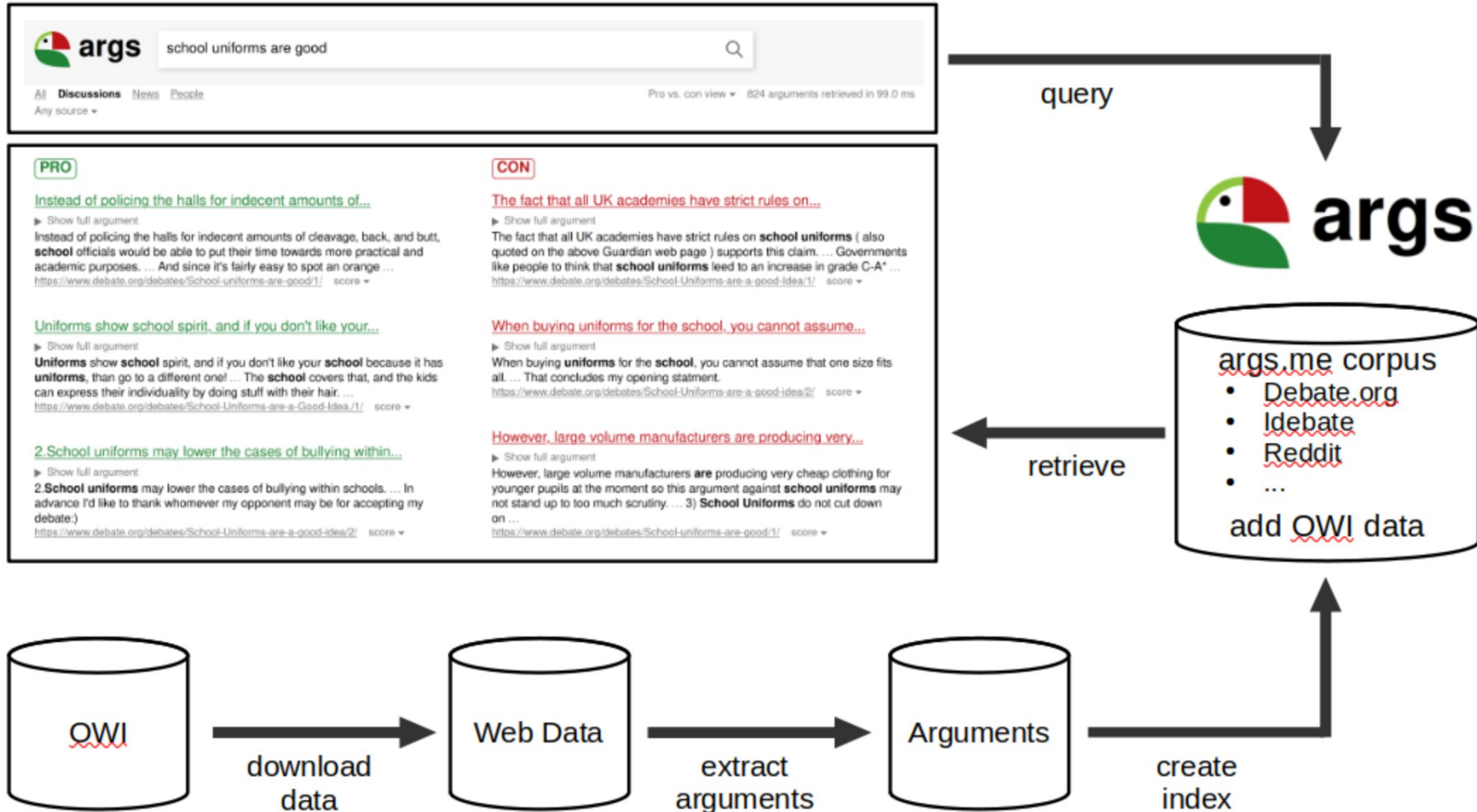
► Show full argument

However, large volume manufacturers **are** producing very cheap clothing for younger pupils at the moment so this argument against **school uniforms** may not stand up to too much scrutiny. ... 3) **School Uniforms** do not cut down

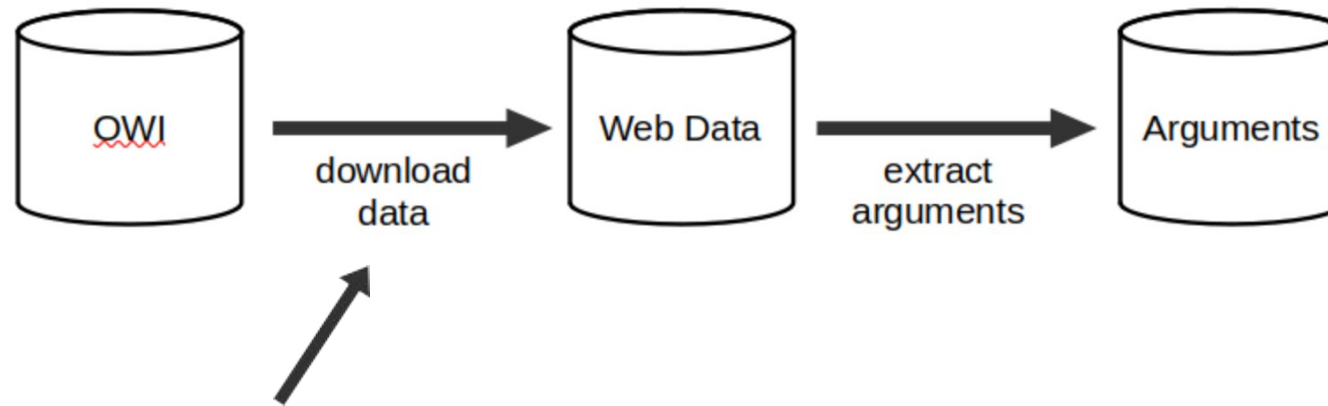
Argumentation Search



Argumentation Search

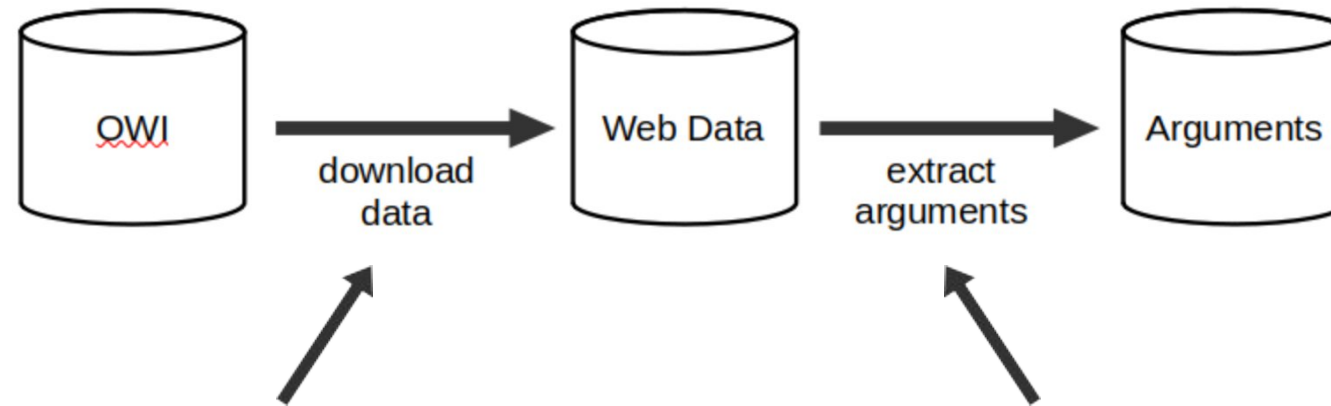


Argumentation Search



- 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>