Faculty of Environmental Sciences
Chair of Geoinformatics

Large Language Models for Conversational Geodata Search

AGILE 2025 Tutorial, Dresden, Germany

Simeon Wetzel, Auriol Degbelo, Stephan Maes

Agenda

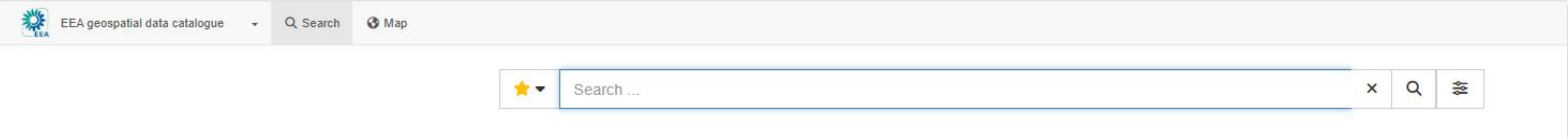
- **Part 1** (45min)**:**
  - Introduction / Motivation / Scenario

- **Part 2** (75 min)**:**
  - LLM Calls
  - Retrieval Augmented Generation (RAG)
  - Geocoding/ Query Interpretation
  - Conversation
- **Part 3** (75 min)**:**
  - Agents
  - SmolAgents Framework

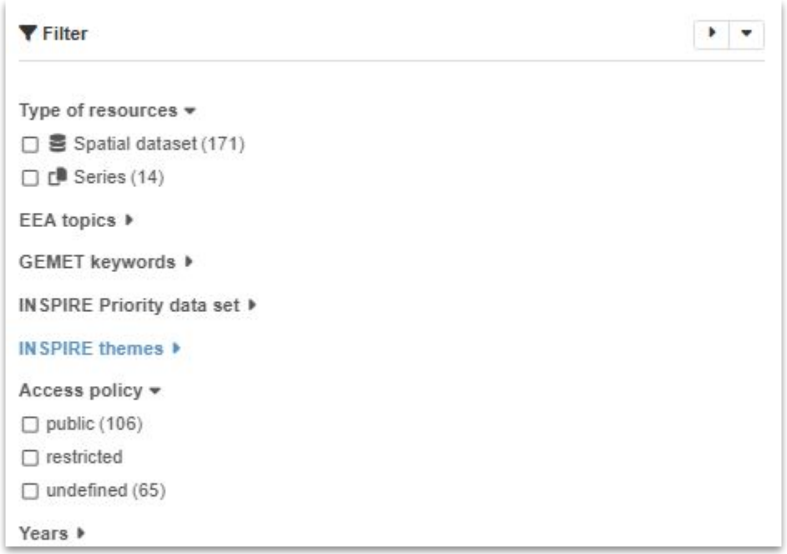
Motivation: LLMs for Conversational Geodata Search

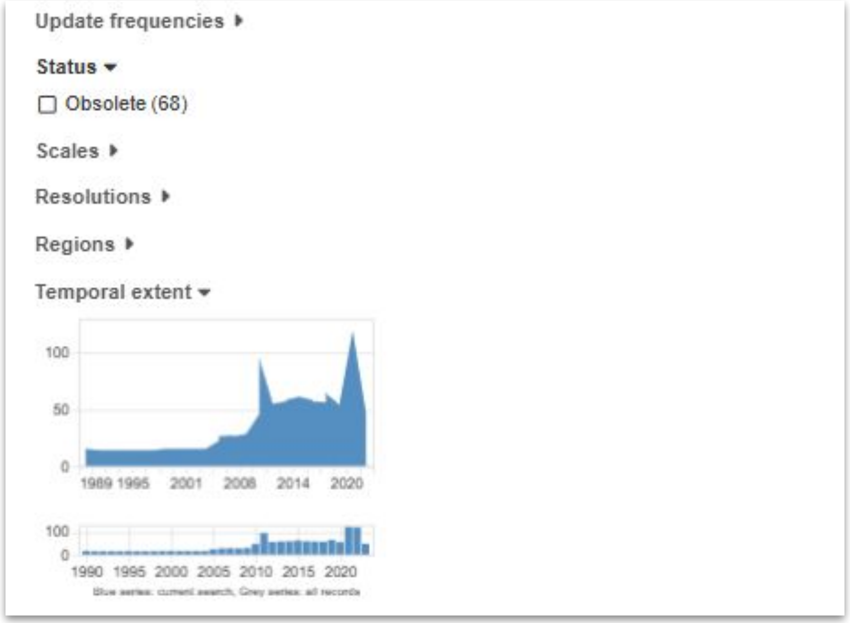
Traditional Search Approach

- **Metadata Catalogues / Geoportals** (e.g. EEA SDI Catalogue)
  - Full-text interface



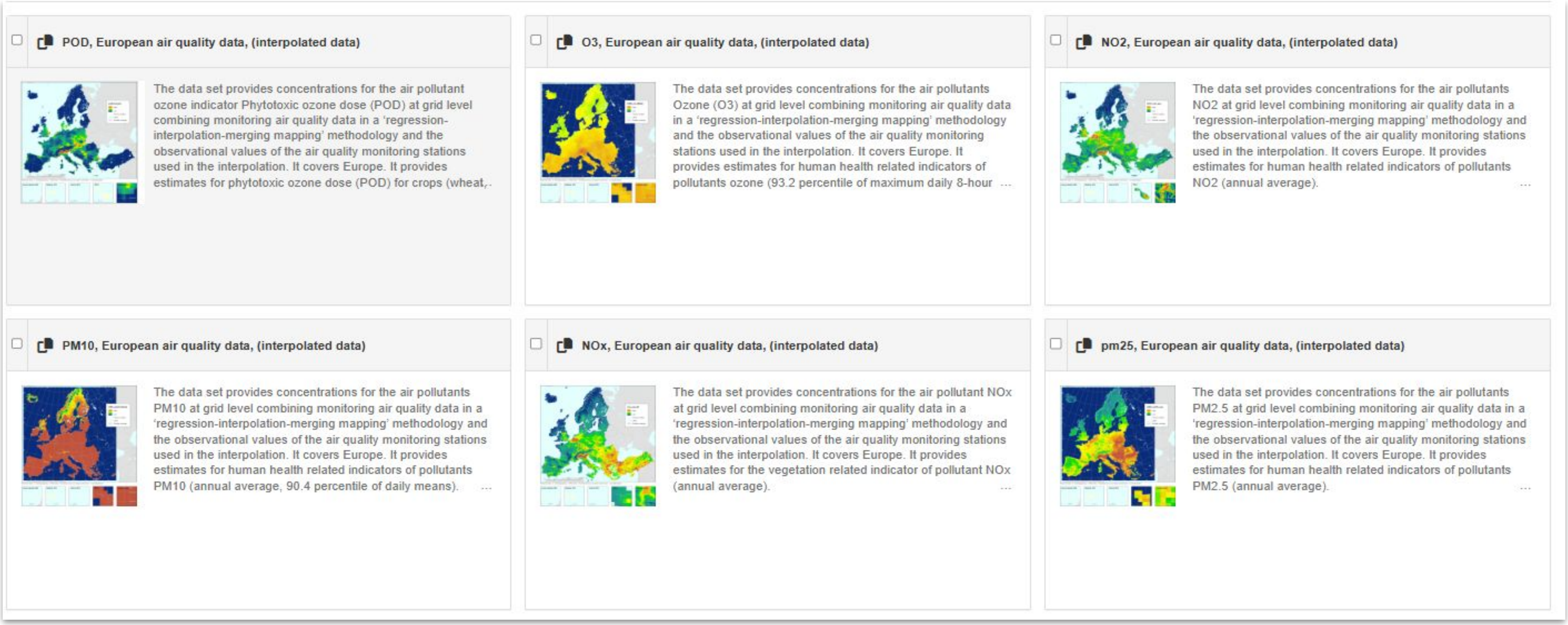
  - Search Filter





Traditional Search Approach

- **Metadata Catalogues / Geoportals**
  - SERP-based result list



□ **POD, European air quality data, (interpolated data)**

The data set provides concentrations for the air pollutant ozone indicator Phytotoxic ozone dose (POD) at grid level combining monitoring air quality data in a 'regression-interpolation-merging mapping' methodology and the observational values of the air quality monitoring stations used in the interpolation. It covers Europe. It provides estimates for phytotoxic ozone dose (POD) for crops (wheat,.

□ **O3, European air quality data, (interpolated data)**

The data set provides concentrations for the air pollutants Ozone (O3) at grid level combining monitoring air quality data in a 'regression-interpolation-merging mapping' methodology and the observational values of the air quality monitoring stations used in the interpolation. It covers Europe. It provides estimates for human health related indicators of pollutants ozone (93.2 percentile of maximum daily 8-hour …

□ **NO2, European air quality data, (interpolated data)**

The data set provides concentrations for the air pollutants NO2 at grid level combining monitoring air quality data in a 'regression-interpolation-merging mapping' methodology and the observational values of the air quality monitoring stations used in the interpolation. It covers Europe. It provides estimates for human health related indicators of pollutants NO2 (annual average). …

□ **PM10, European air quality data, (interpolated data)**

The data set provides concentrations for the air pollutants PM10 at grid level combining monitoring air quality data in a 'regression-interpolation-merging mapping' methodology and the observational values of the air quality monitoring stations used in the interpolation. It covers Europe. It provides estimates for human health related indicators of pollutants PM10 (annual average, 90.4 percentile of daily means). …

□ **NOx, European air quality data, (interpolated data)**

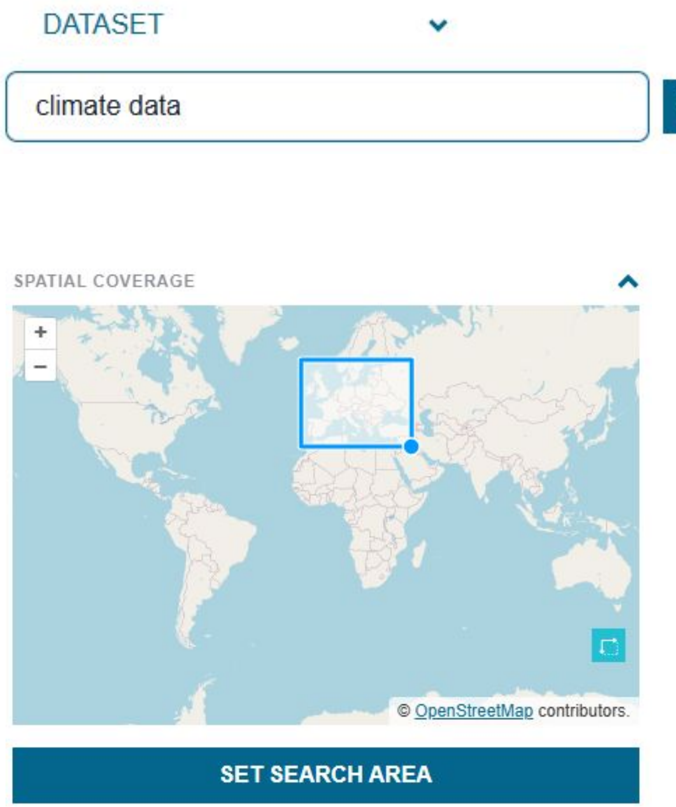
The data set provides concentrations for the air pollutant NOx at grid level combining monitoring air quality data in a 'regression-interpolation-merging mapping' methodology and the observational values of the air quality monitoring stations used in the interpolation. It covers Europe. It provides estimates for the vegetation related indicator of pollutant NOx (annual average). …

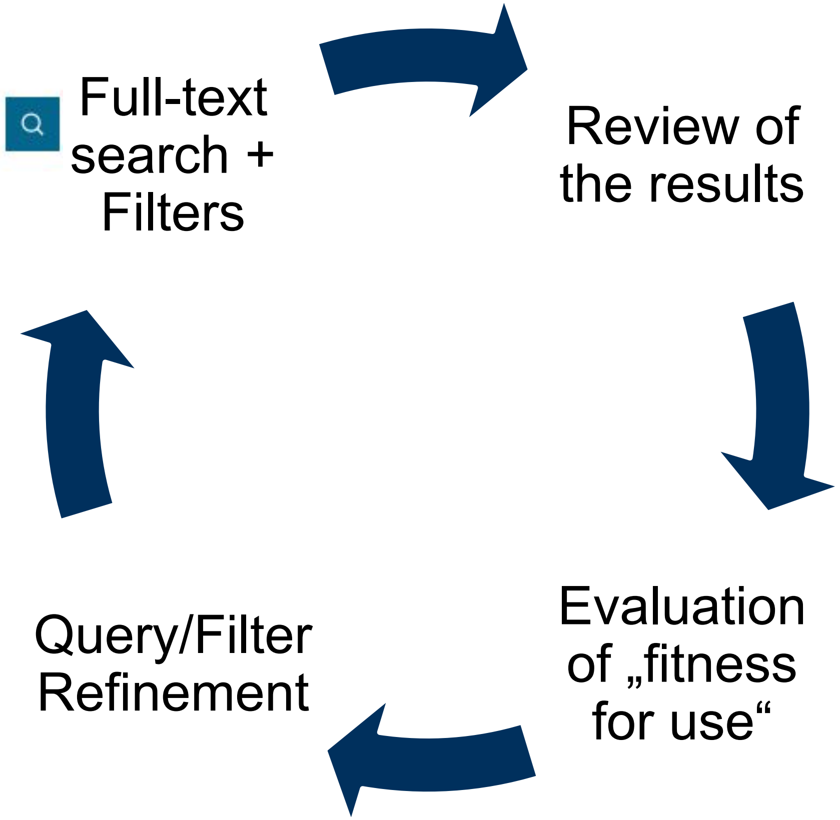
□ **pm25, European air quality data, (interpolated data)**

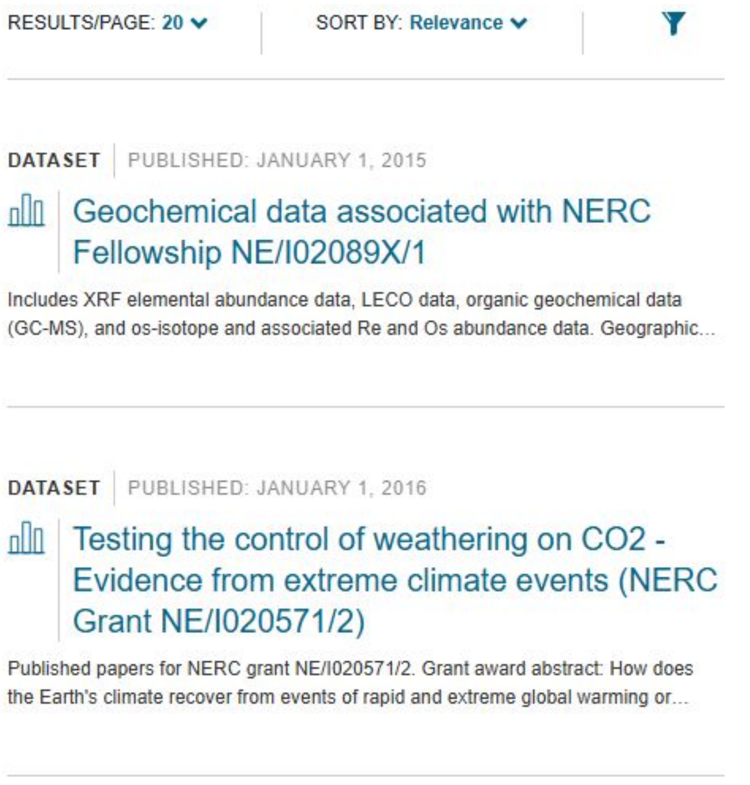
The data set provides concentrations for the air pollutants PM2.5 at grid level combining monitoring air quality data in a 'regression-interpolation-merging mapping' methodology and the observational values of the air quality monitoring stations used in the interpolation. It covers Europe. It provides estimates for human health related indicators of pollutants PM2.5 (annual average). …

Traditional Search Approach

- **Cycle of single-hop search + refinements**







DATASET | PUBLISHED: JANUARY 1, 2015

Geochemical data associated with NERC Fellowship NE/I02089X/1

Includes XRF elemental abundance data, LECO data, organic geochemical data (GC-MS), and os-isotope and associated Re and Os abundance data. Geographic...

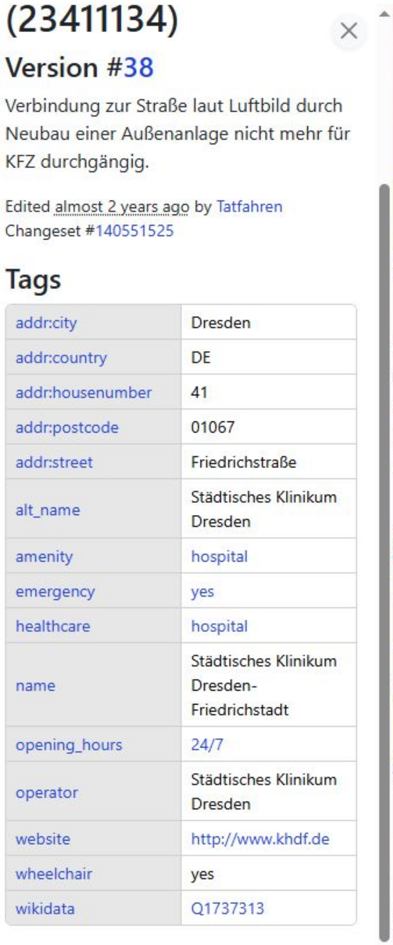
DATASET | PUBLISHED: JANUARY 1, 2016

Testing the control of weathering on CO2 - Evidence from extreme climate events (NERC Grant NE/I020571/2)

Published papers for NERC grant NE/I020571/2. Grant award abstract: How does the Earth's climate recover from events of rapid and extreme global warming or...

Challenges with traditional search approach

- Dependency on attributes / metadata quality (completeness / accuracy)



(23411134)

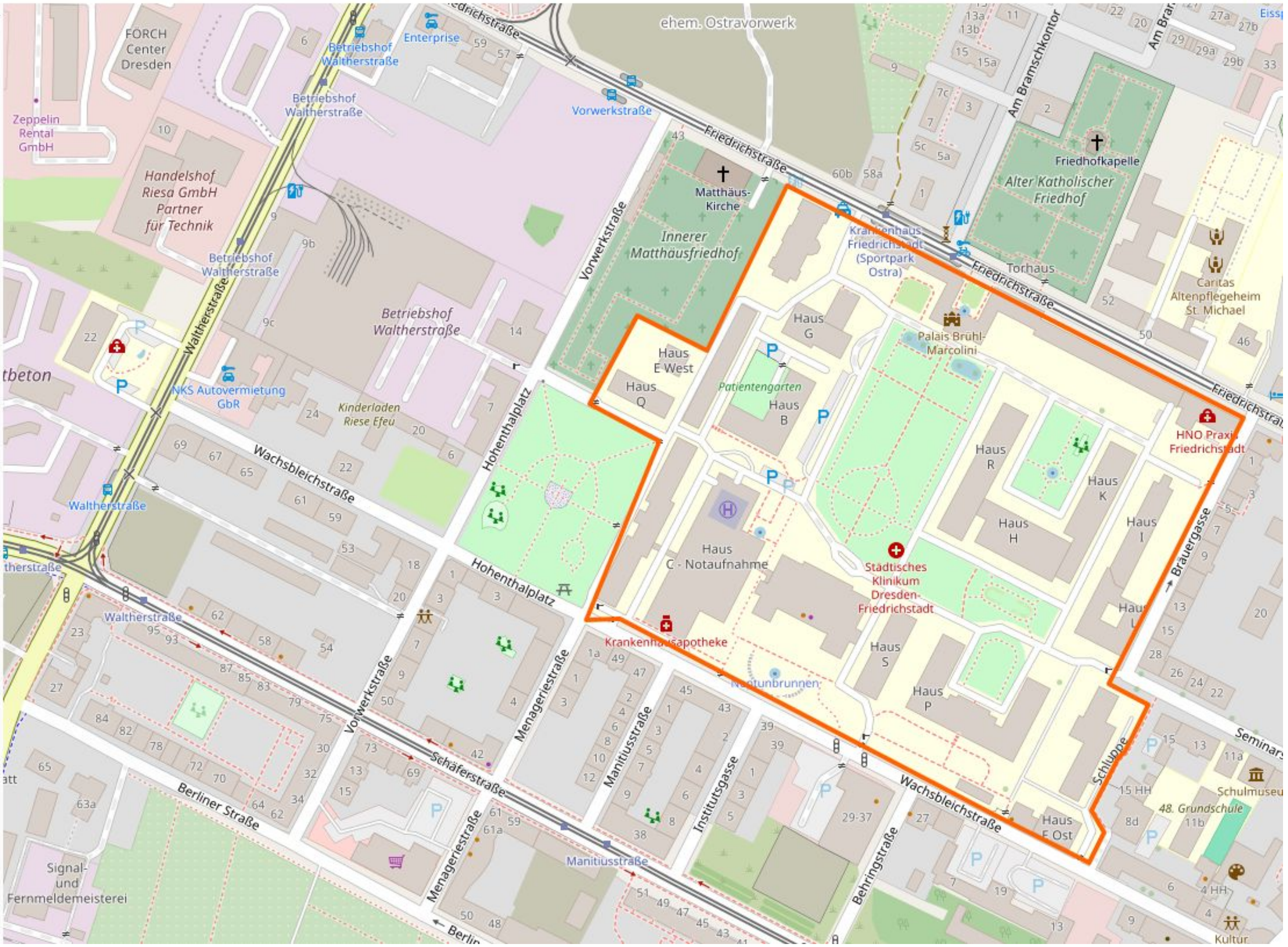
Version #38

Verbindung zur Straße laut Luftbild durch Neubau einer Außenanlage nicht mehr für KFZ durchgängig.

Edited almost 2 years ago by Tatfahren
Changeset #140551525

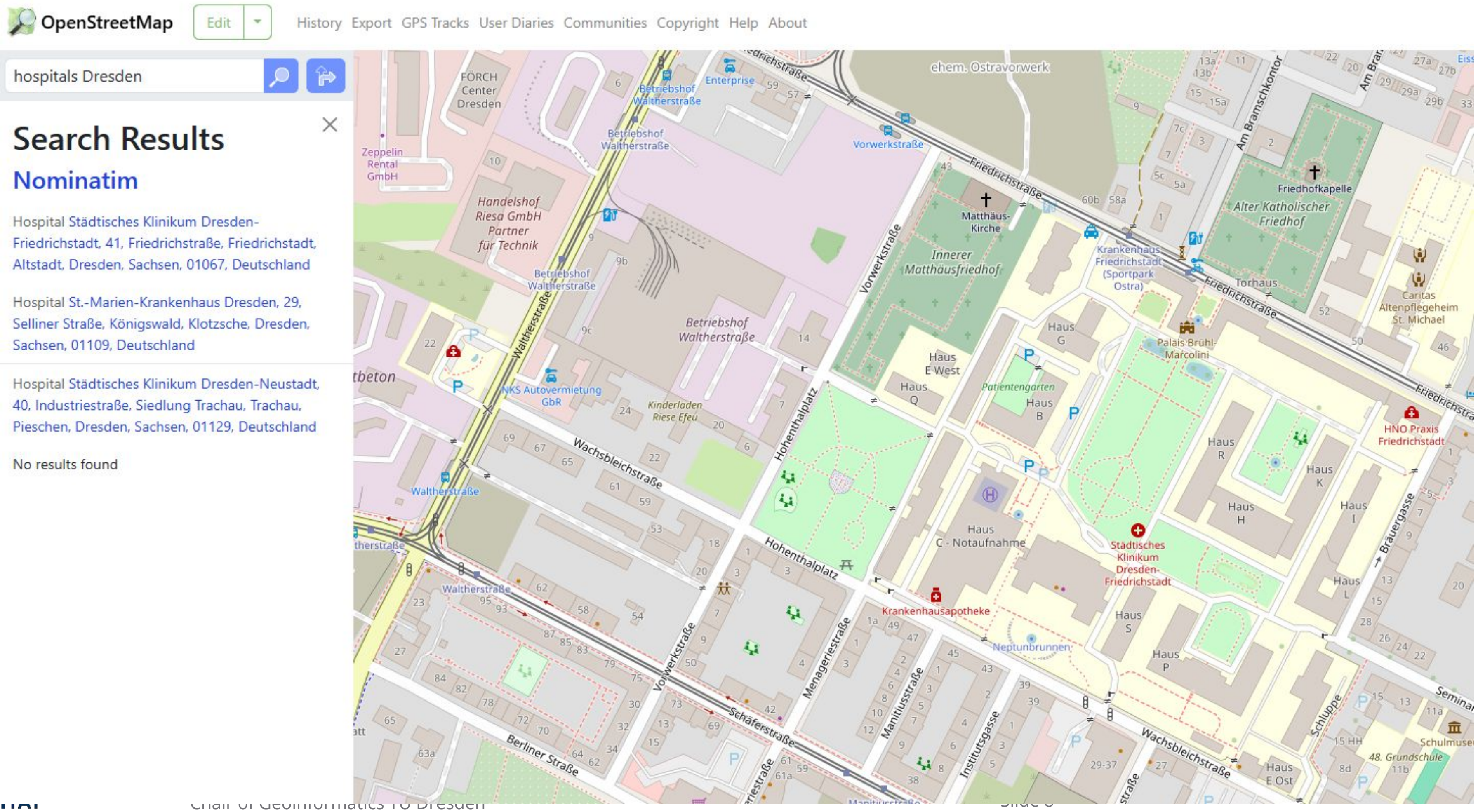
Tags

addr:city	Dresden
addr:country	DE
addr:housenumber	41
addr:postcode	01067
addr:street	Friedrichstraße
alt_name	Städtisches Klinikum Dresden
amenity	hospital
emergency	yes
healthcare	hospital
name	Städtisches Klinikum Dresden-Friedrichstadt
opening_hours	24/7
operator	Städtisches Klinikum Dresden
website	http://www.khdf.de
wheelchair	yes
wikidata	Q1737313



Challenges with traditional search approach

- Dependency on attributes



Search Results

Nominatim

Hospital Städtisches Klinikum Dresden-Friedrichstadt, 41, Friedrichstraße, Friedrichstadt, Altstadt, Dresden, Sachsen, 01067, Deutschland

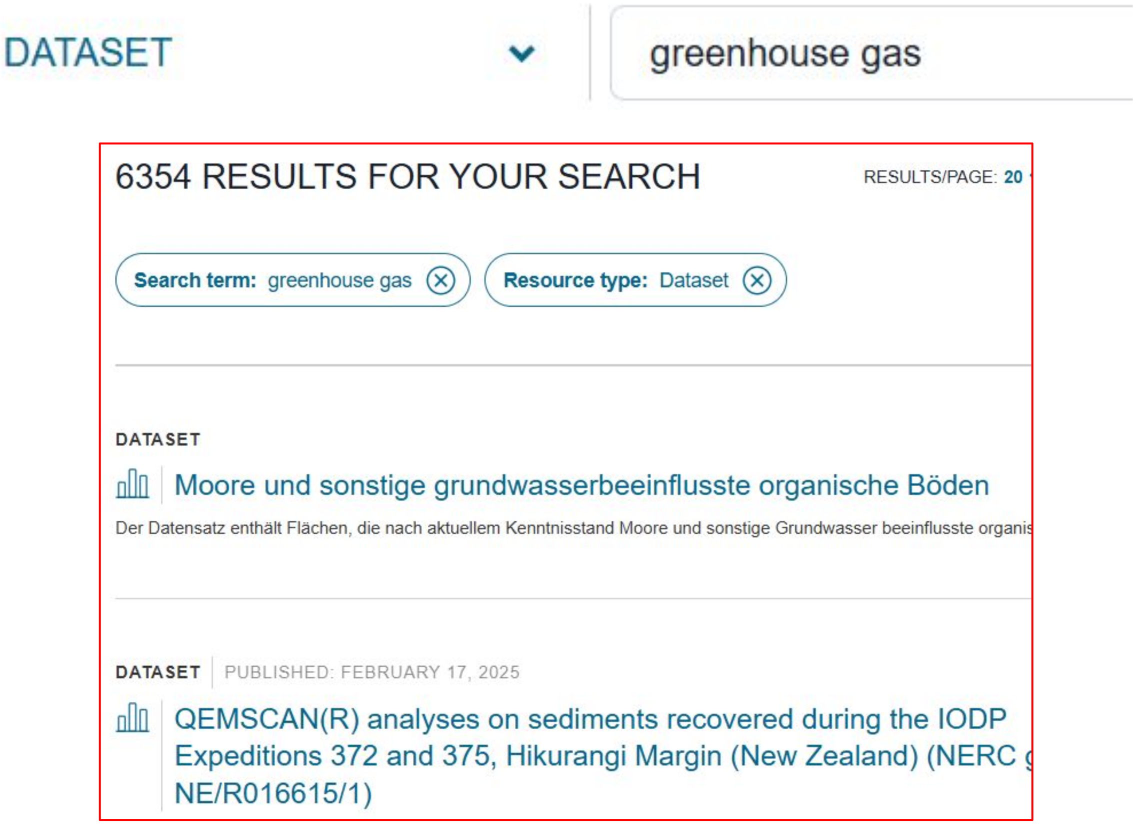
Hospital St.-Marien-Krankenhaus Dresden, 29, Selliner Straße, Königswald, Klotzsche, Dresden, Sachsen, 01109, Deutschland

Hospital Städtisches Klinikum Dresden-Neustadt, 40, Industriestraße, Siedlung Trachau, Trachau, Pieschen, Dresden, Sachsen, 01129, Deutschland

No results found

Challenges with traditional search approach

- Lexical Search => No semantic search



DATASET

greenhouse gas

6354 RESULTS FOR YOUR SEARCH

RESULTS/PAGE: 20

Search term: greenhouse gas ⊗ Resource type: Dataset ⊗

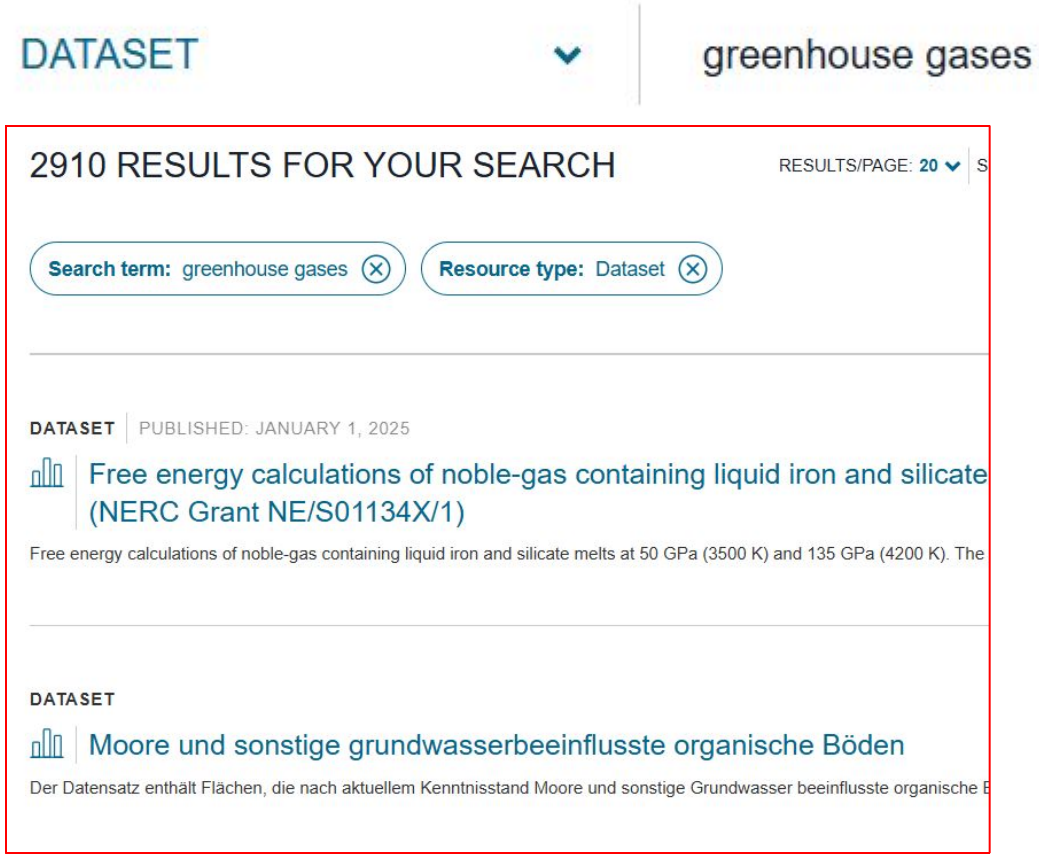
DATASET

Moore und sonstige grundwasserbeeinflusste organische Böden

Der Datensatz enthält Flächen, die nach aktuellem Kenntnisstand Moore und sonstige Grundwasser beeinflusste organis

DATASET | PUBLISHED: FEBRUARY 17, 2025

QEMSCAN(R) analyses on sediments recovered during the IODP Expeditions 372 and 375, Hikurangi Margin (New Zealand) (NERC g NE/R016615/1)



DATASET

greenhouse gases

2910 RESULTS FOR YOUR SEARCH

RESULTS/PAGE: 20 S

Search term: greenhouse gases ⊗ Resource type: Dataset ⊗

DATASET | PUBLISHED: JANUARY 1, 2025

Free energy calculations of noble-gas containing liquid iron and silicate (NERC Grant NE/S01134X/1)

Free energy calculations of noble-gas containing liquid iron and silicate melts at 50 GPa (3500 K) and 135 GPa (4200 K). The

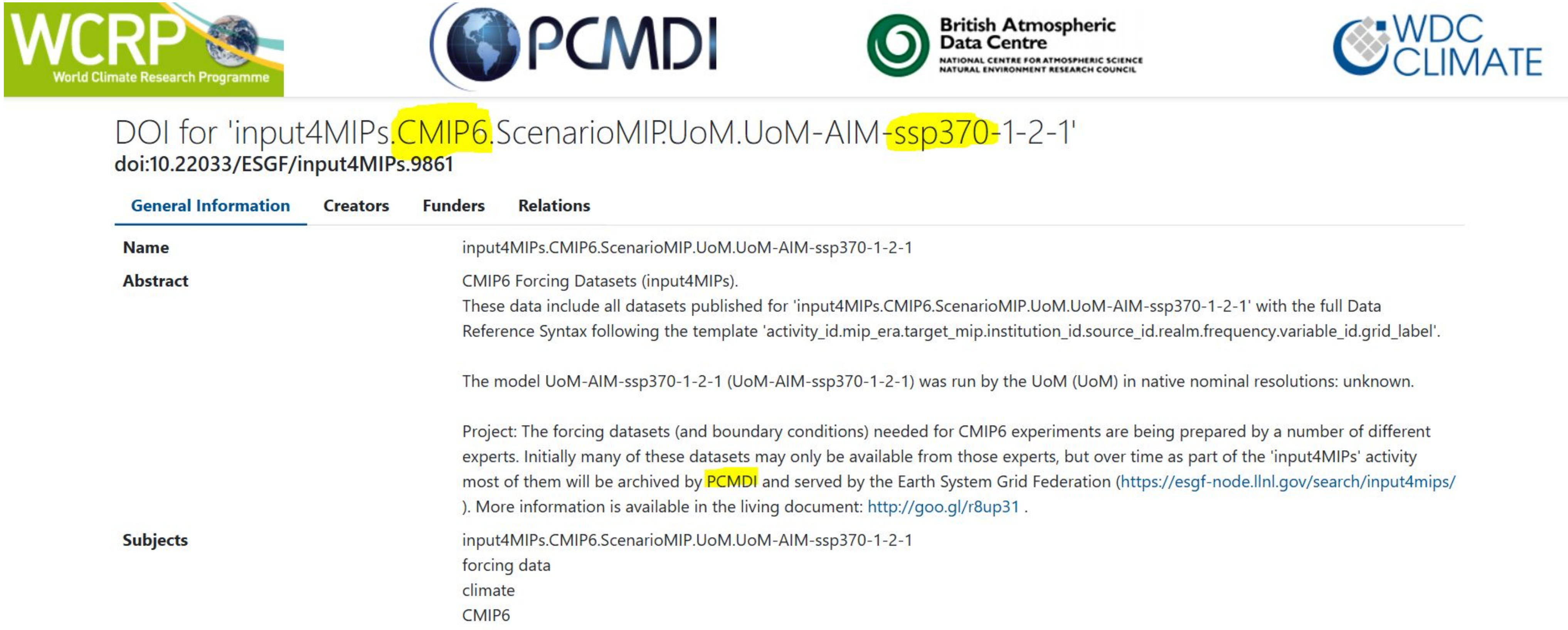
DATASET

Moore und sonstige grundwasserbeeinflusste organische Böden

Der Datensatz enthält Flächen, die nach aktuellem Kenntnisstand Moore und sonstige Grundwasser beeinflusste organische E

Challenges with traditional search approach

- Specific Terminology in ESS data


WCRP
World Climate Research Programme
PCMDI
British Atmospheric Data Centre
NATIONAL CENTRE FOR ATMOSPHERIC SCIENCE
NATURAL ENVIRONMENT RESEARCH COUNCIL
WDC CLIMATE

DOI for 'input4MIPs.CMIP6.ScenarioMIP.UoM.UoM-AIM-ssp370-1-2-1'
doi:10.22033/ESGF/input4MIPs.9861

General Information | Creators | Funders | Relations

Name	input4MIPs.CMIP6.ScenarioMIP.UoM.UoM-AIM-ssp370-1-2-1
Abstract	CMIP6 Forcing Datasets (input4MIPs). These data include all datasets published for 'input4MIPs.CMIP6.ScenarioMIP.UoM.UoM-AIM-ssp370-1-2-1' with the full Data Reference Syntax following the template 'activity_id.mip_era.target_mip.institution_id.source_id.realm.frequency.variable_id.grid_label'. The model UoM-AIM-ssp370-1-2-1 (UoM-AIM-ssp370-1-2-1) was run by the UoM (UoM) in native nominal resolutions: unknown. Project: The forcing datasets (and boundary conditions) needed for CMIP6 experiments are being prepared by a number of different experts. Initially many of these datasets may only be available from those experts, but over time as part of the 'input4MIPs' activity most of them will be archived by PCMDI and served by the Earth System Grid Federation (https://esgf-node.llnl.gov/search/input4mips/). More information is available in the living document: http://goo.gl/r8up31 .
Subjects	input4MIPs.CMIP6.ScenarioMIP.UoM.UoM-AIM-ssp370-1-2-1 forcing data climate CMIP6

Challenges with traditional search approach

- Context-less single-hop queries
- Lexical Search issues
- Specific Terminology in ESS data
- **Complex queries:**
  - 🌍🕒 Spatio-temporal queries:
    - „Historic buildings around…“, „“Heavy Precipitation Europe“, „Climate Projection 2020-2100“
  - ❓ Ambiguous entities:
    - „Radiation data“, „Buildings in Frankfurt Germany“…
  - 🌍 🤷 Vague spatial entities:
    - „…East coast…“, „North of Ireland“

Improvements with LLM-based search

- Improved capabilities for …

  … **query interpretation**

  … search **result interpretation**

  … **context-awareness**

  … **semantic search**

Scenario in this tutorial

Scenario: Design of a LLM-driven search architecture for geodata

Data:

- OpenStreetMap Data (buildings in Dresden, ~50k features used)

Data Pre-processing

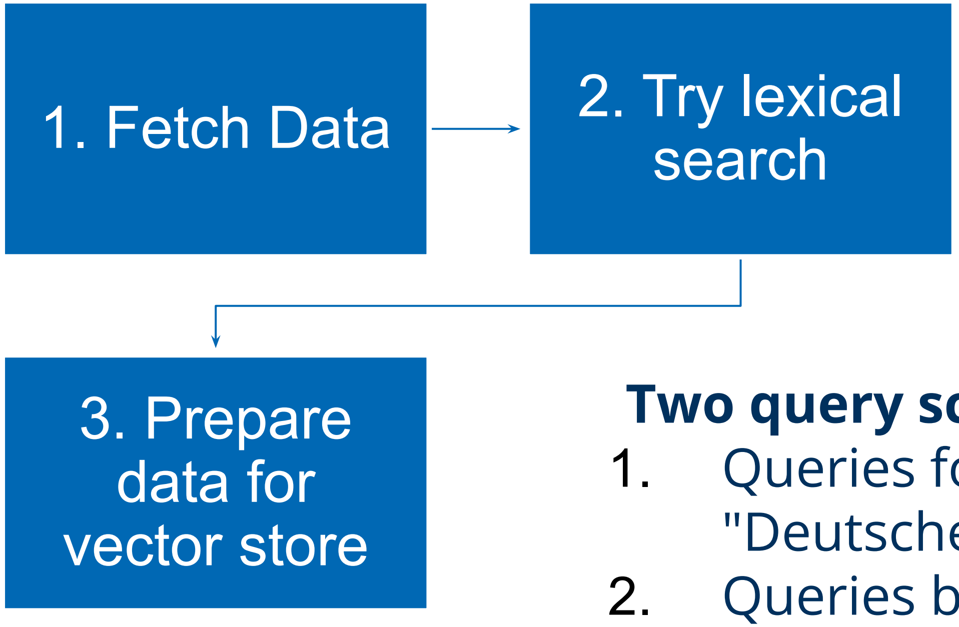
- Representing OSM data as embeddings
- Loading data into a vector store

LLM-based search:

- Retrieval of the data
- Using data as context

Scenario

Part 1:



1. Fetch Data

2. Try lexical search

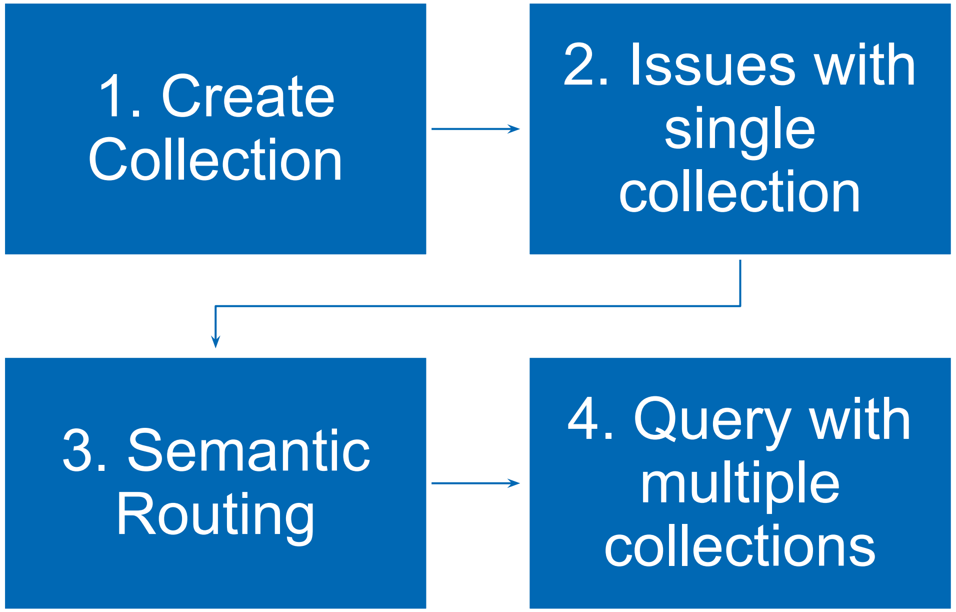
3. Prepare data for vector store

Two query scenarios

1. Queries for a **specific buildings** (by name) (e.g. "Deutsches Hygiene Museum")
2. Queries by **building type** (e.g. "museums in Dresden")

Scenario

Part 2:



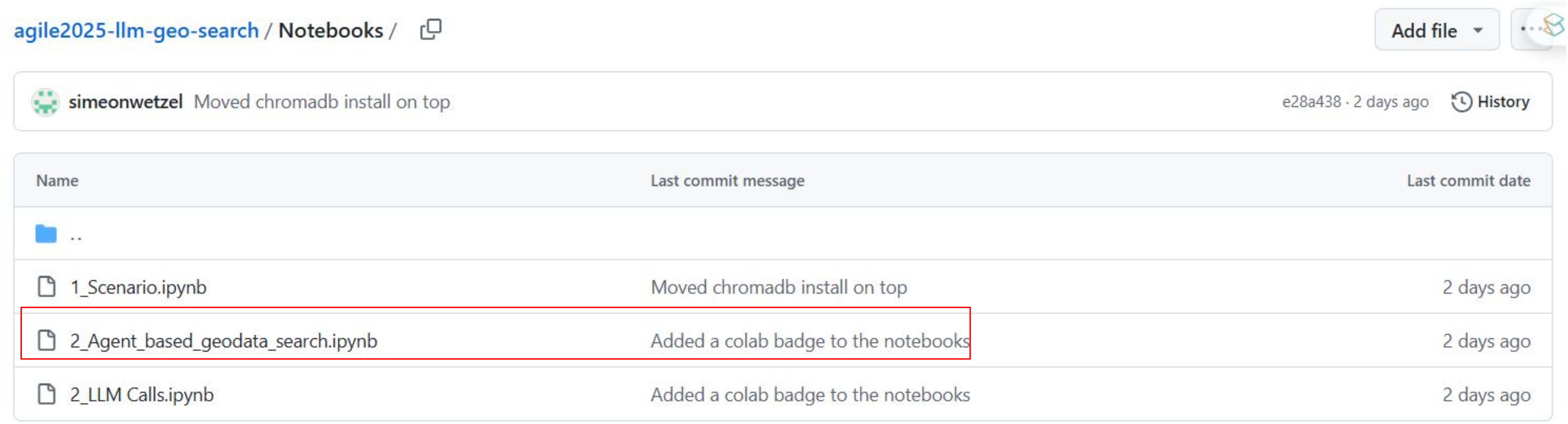
1. Create Collection

2. Issues with single collection

3. Semantic Routing

4. Query with multiple collections

Notebook:



agile2025-llm-geo-search / Notebooks /

simeonwetzel Moved chromadb install on top e28a438 · 2 days ago History

Name	Last commit message	Last commit date
..		
1_Scenario.ipynb	Moved chromadb install on top	2 days ago
2_Agent_based_geodata_search.ipynb	Added a colab badge to the notebooks	2 days ago
2_LLM Calls.ipynb	Added a colab badge to the notebooks	2 days ago

https://bit.ly/agile25-llm

API Keys

https://bit.ly/agile25-key