Sustav za analizu metapodataka otvorenih skupova podataka

A System for Metadata Analysis of Open Datasets

Sve veća dostupnost otvorenih podataka ne podrazumijeva i njihovu veću iskoristivost. Iako normirani formati meta podataka (npr. DCAT) omogućuju jednostavnije pronalaženje i tumačenje objavljenih pojedinačnih skupova podataka, dodatna vrijednost nalazi se u njihovom povezivanju i pronalaženju novih uvida i skrivenog znanja. Nagli uspon alata umjetne inteligencije temeljenih na velikim jezičnim modelima predstavlja obećavajući smjer za izradu na njima temeljenih alata, a koji bi običnim korisnicima pružili novi uvid u korištenje otvorenih podataka.

U ovom diplomskom radu potrebno je proučiti mogućnosti velikih jezičnih modela, na njima temeljenih alata i tehnika. Također je potrebno proučiti problematiku opisivanja skupova otvorenih podataka korištenjem norme DCAT i mogućnosti automatiziranog traženja veza između skupova podataka. Predložiti alat za dohvat i analizu meta podataka otvorenih skupova podataka, kao i za pružanje podrške korisnicima u povezivanju skupova podataka temeljene na analizi metapodataka. Naposljetku, potrebno je implementirati prototip sustava za portal CKAN korištenjem alata temeljenih na velikim jezičnim modelima i ocijeniti uporabljivost sustava.

## Otvorena pitanja

## Materijali za proučavanje

<https://www.deeplearning.ai/the-batch/issue-292/>

<https://jalammar.github.io/illustrated-transformer/>

<https://huggingface.co/models>

<https://huggingface.co/datasets>

<https://github.com/OpenDataMCP/OpenDataMCP>

## Plan rada

* Prouči SPARQL, usporedba sparql i LLM: <https://www.w3.org/TR/rdf-sparql-query/>
* Primjer: Ulice u Zagrebu i Umjetnici - daj mi skupove podataka (tablice) koje su mi potrebne kako bih kreirao SPARQL upit za input “navedi ulice u zagrebu nazvane po poznatim slikarima”

## Repozitorij

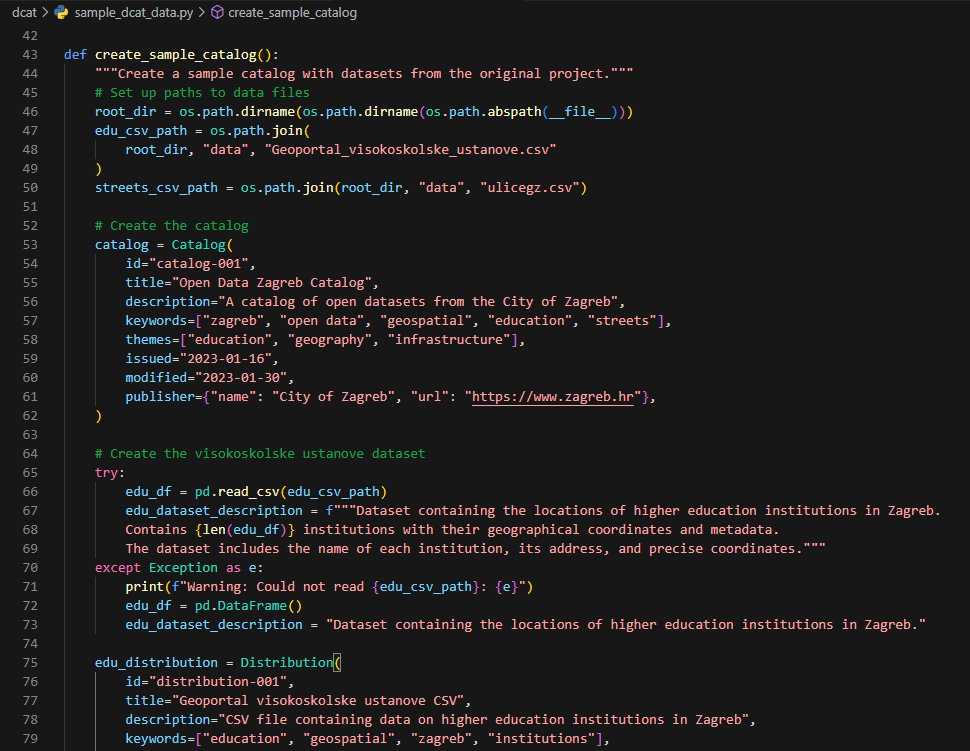
<https://github.com/lukahabus/open-data-assistant>

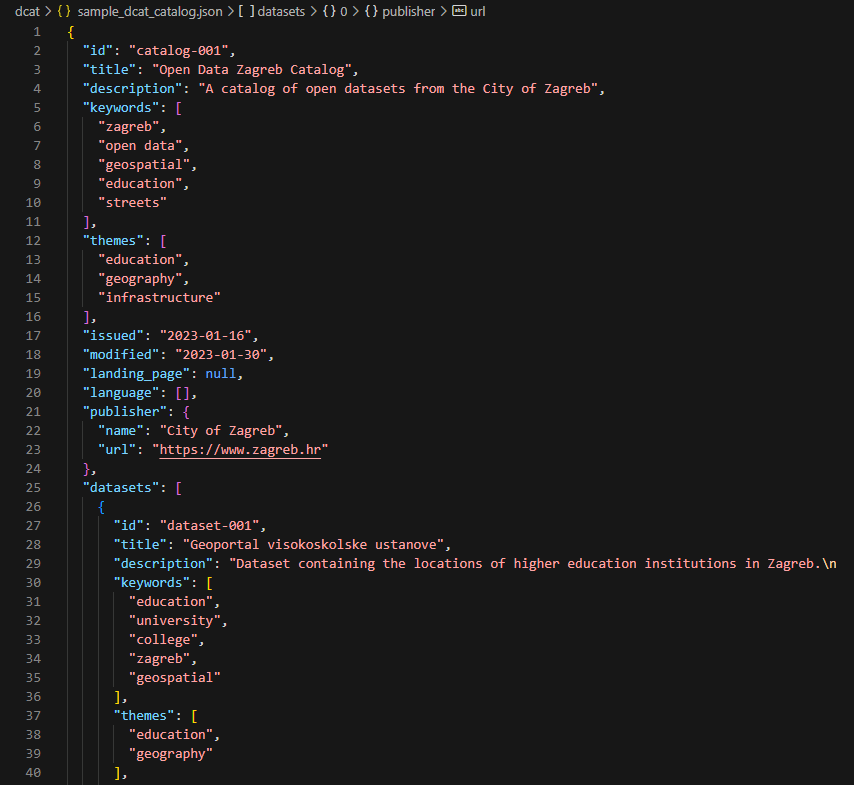
## Zapisnik

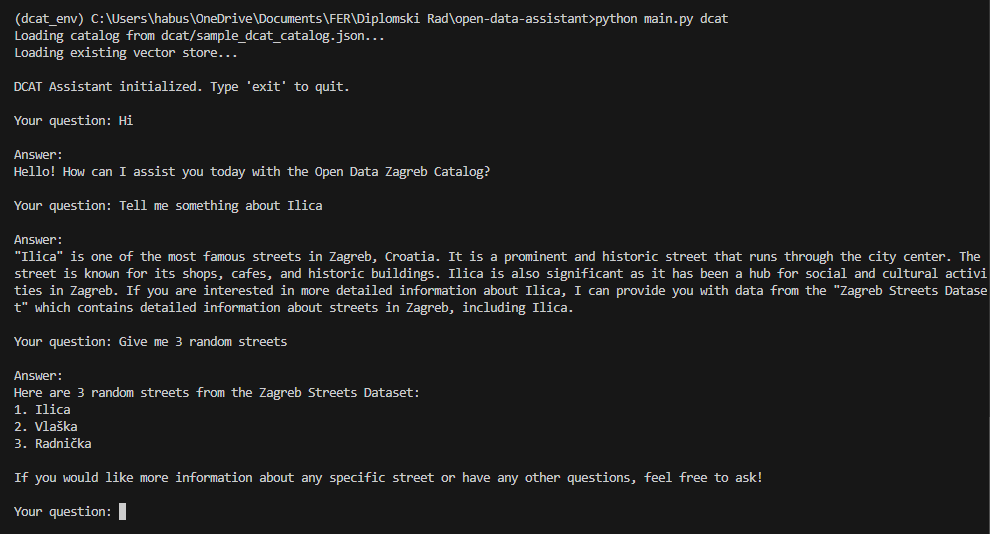
13.03.

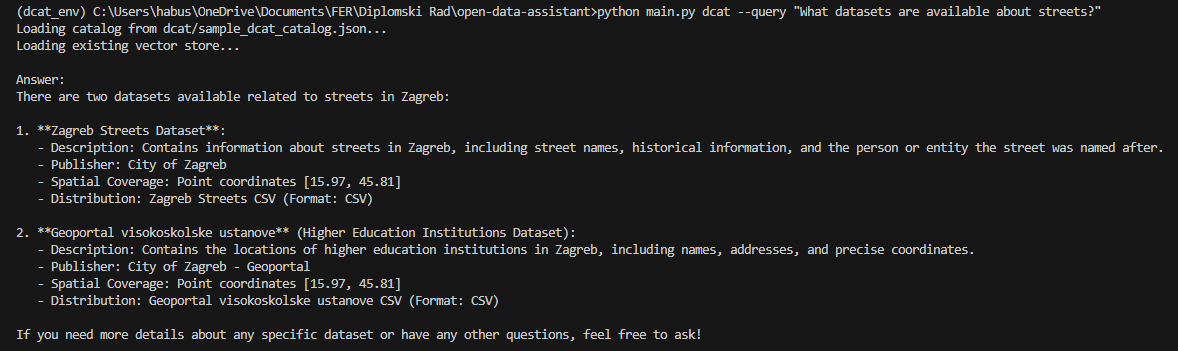
* proći kroz DCAT, izvući atribute, kako prilagoditi opis metapodataka za embedding
* složiti skup podataka sa DCAT opisima koji će služiti za testiranje
* pogledati rješenje s projekta, kako ga prilagoditi za korištenje naših testnih data setova
* usporedba s SQL LLM generiranjem upita, koje su sličnosti i razlike
* sljedeći sastanak 24.03.

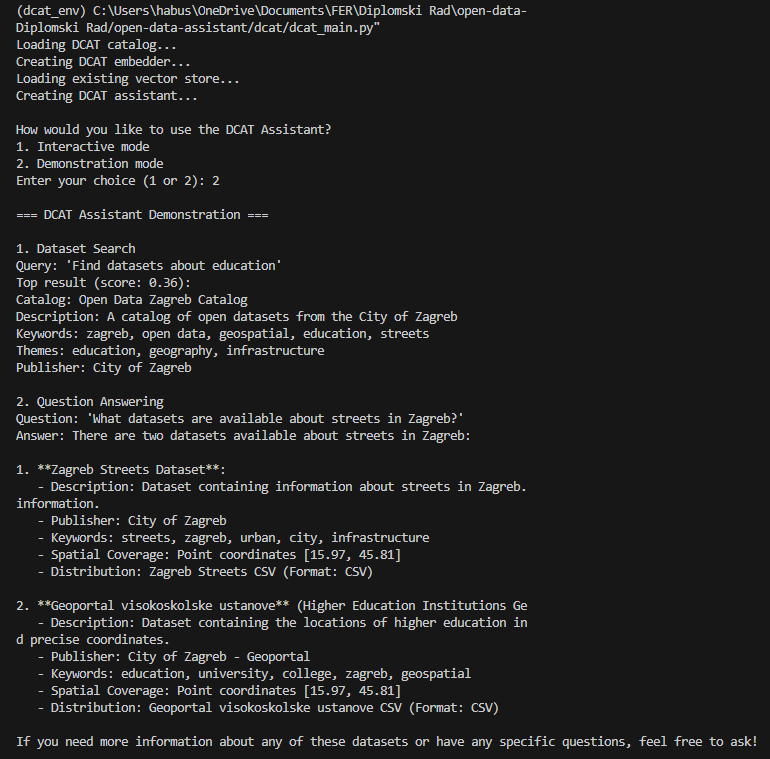
24.03.

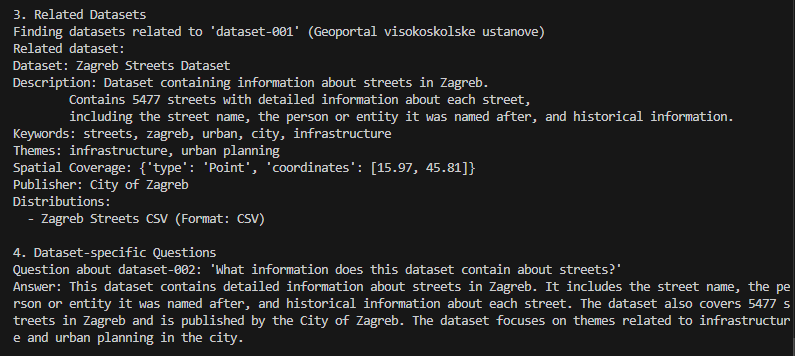


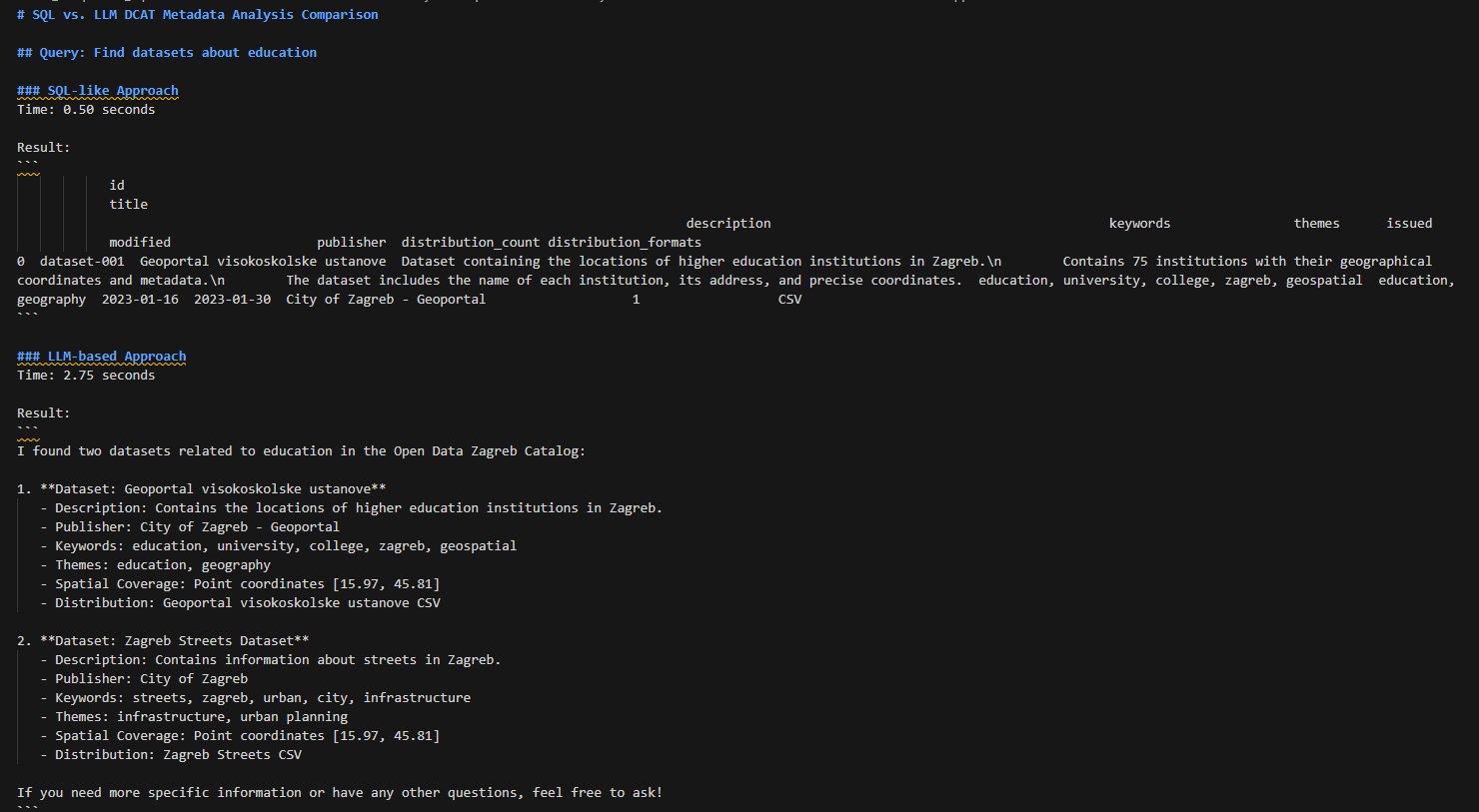


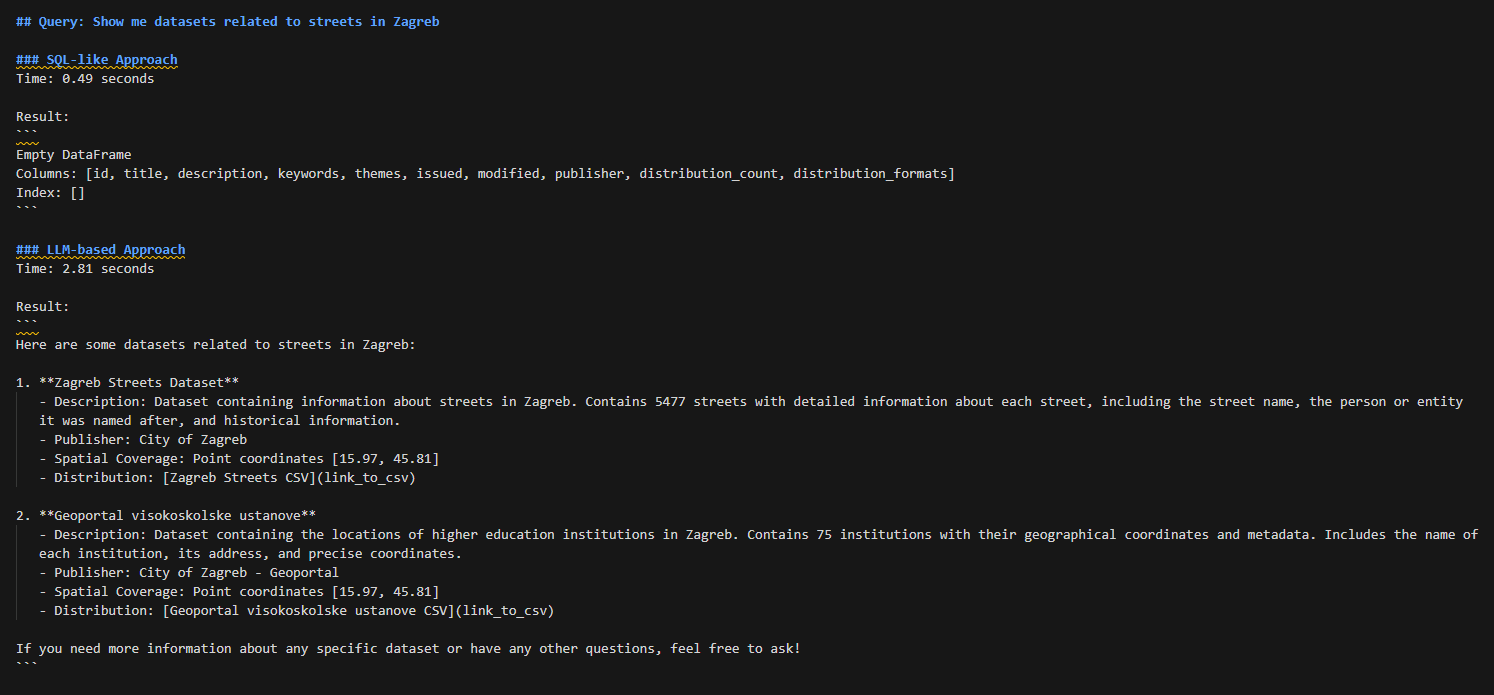


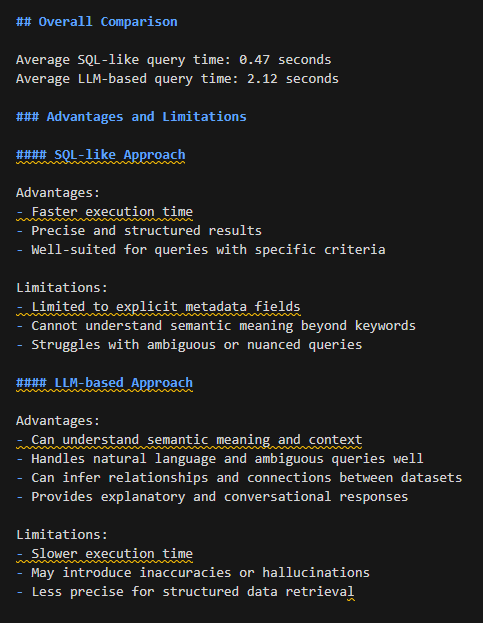


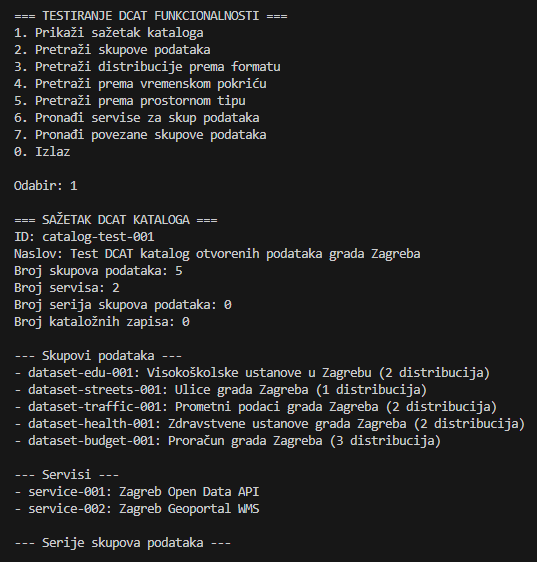


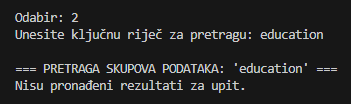


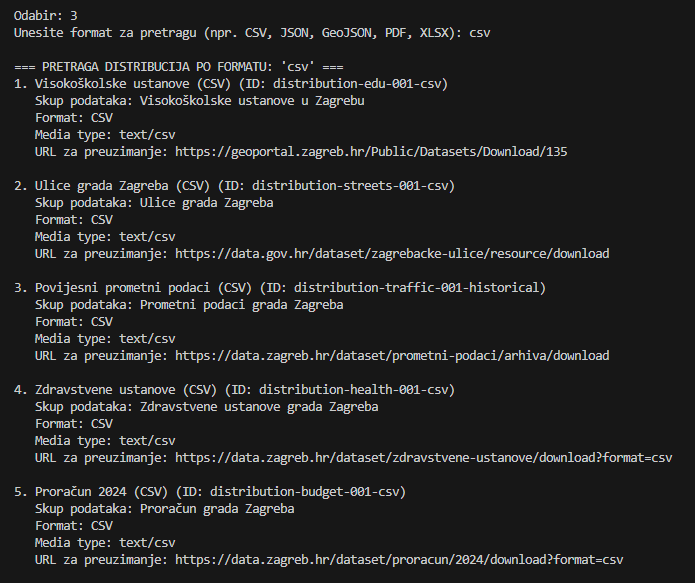




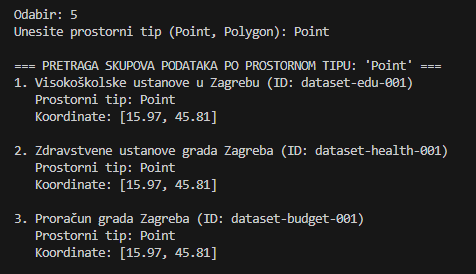


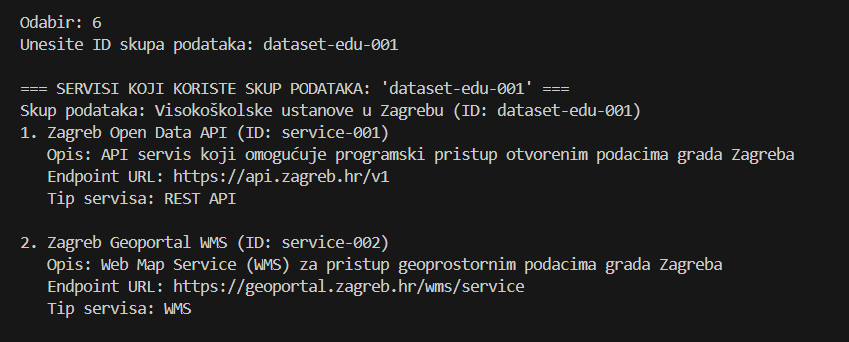


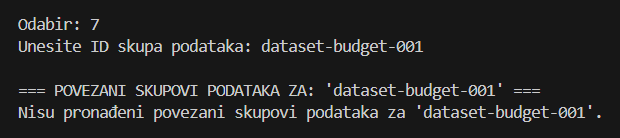




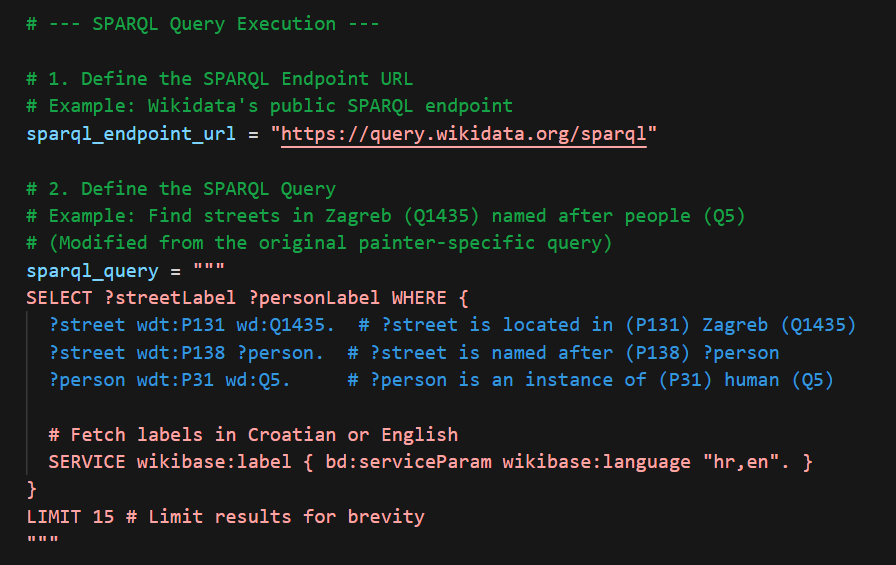


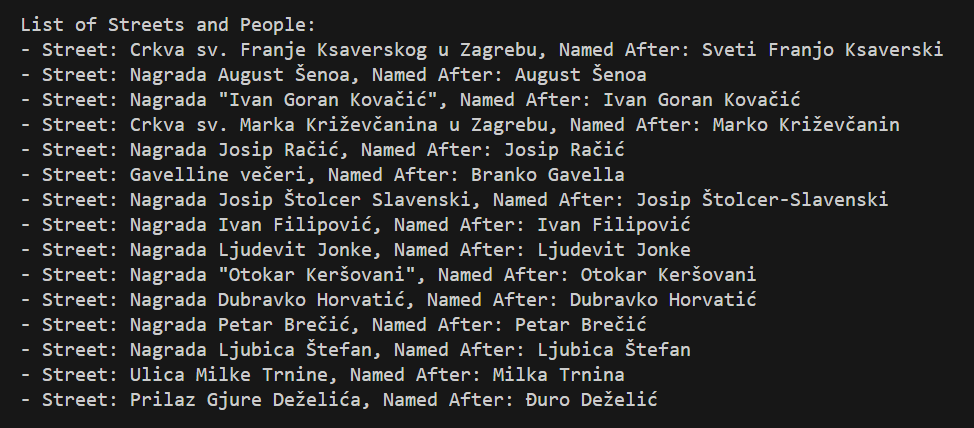






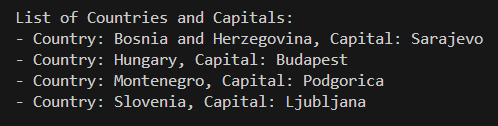
7.4.2025.

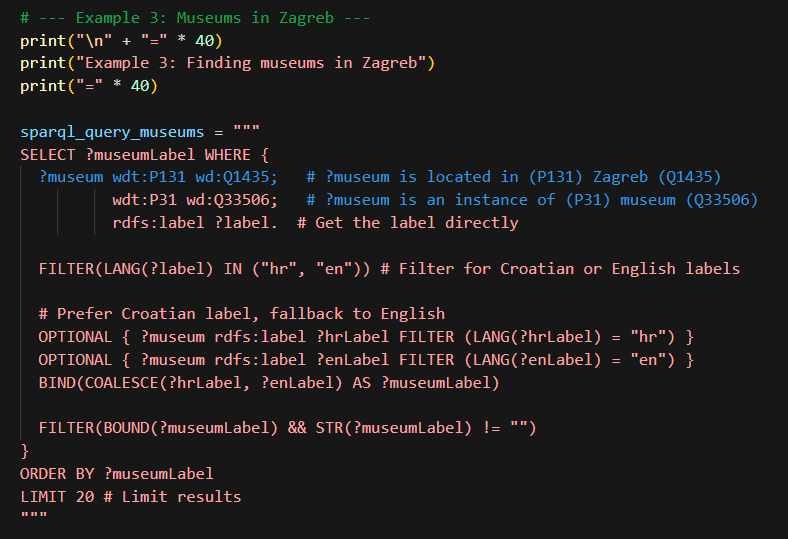


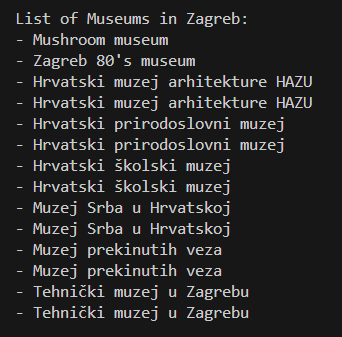


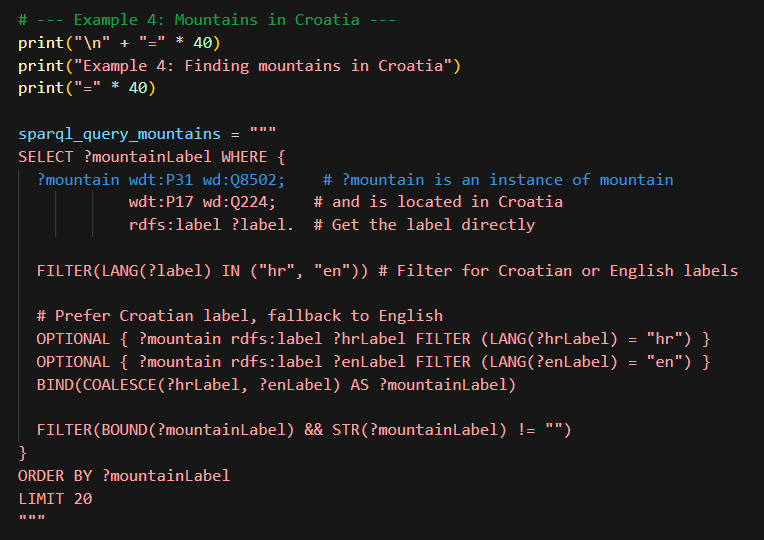


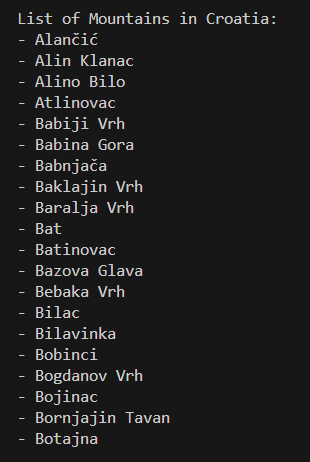


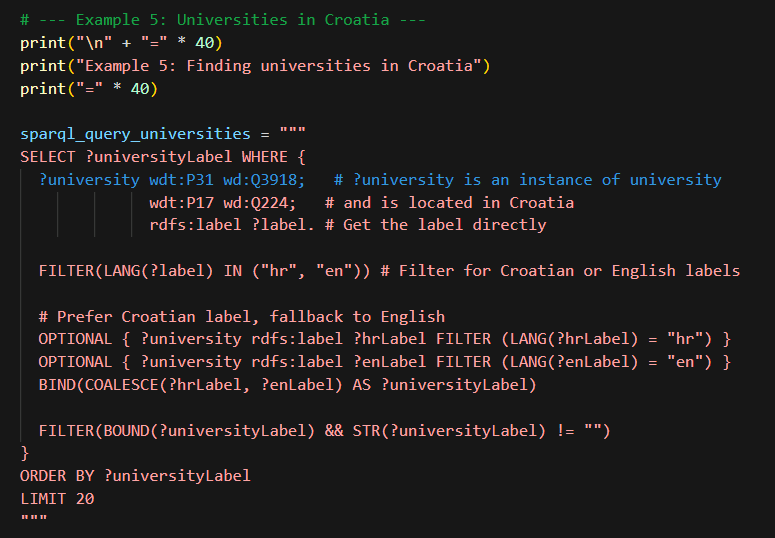




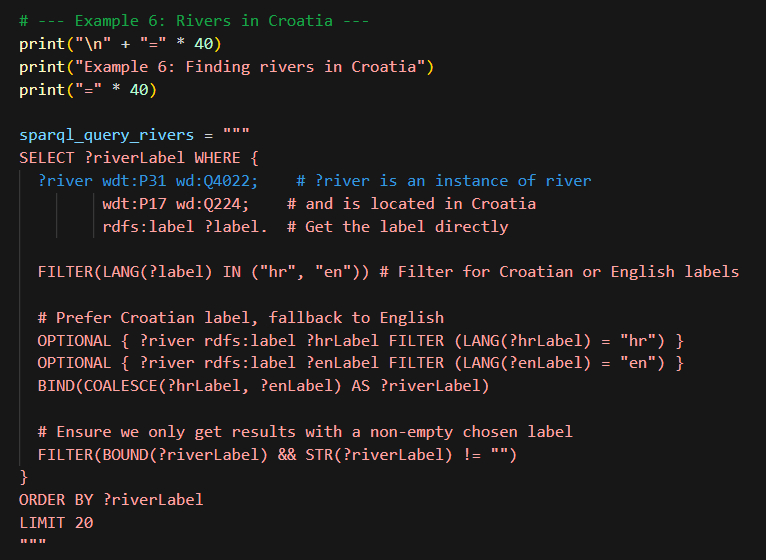


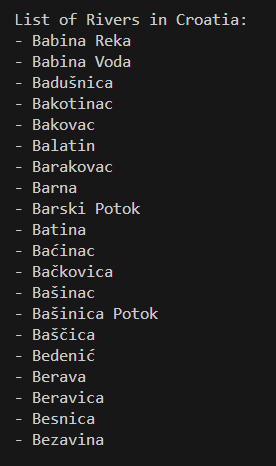


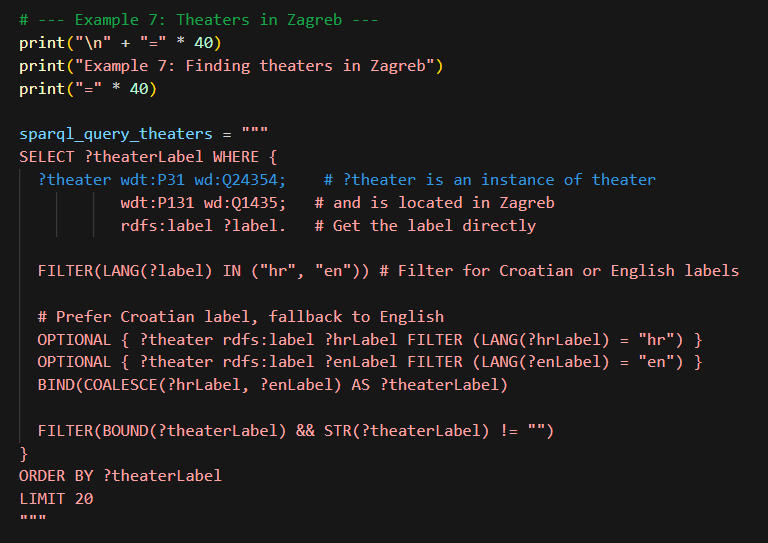


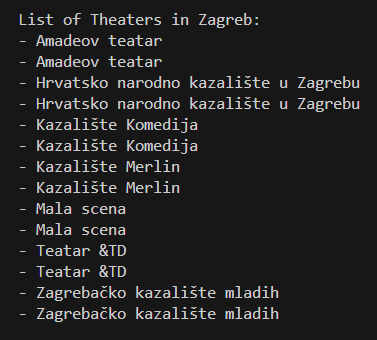






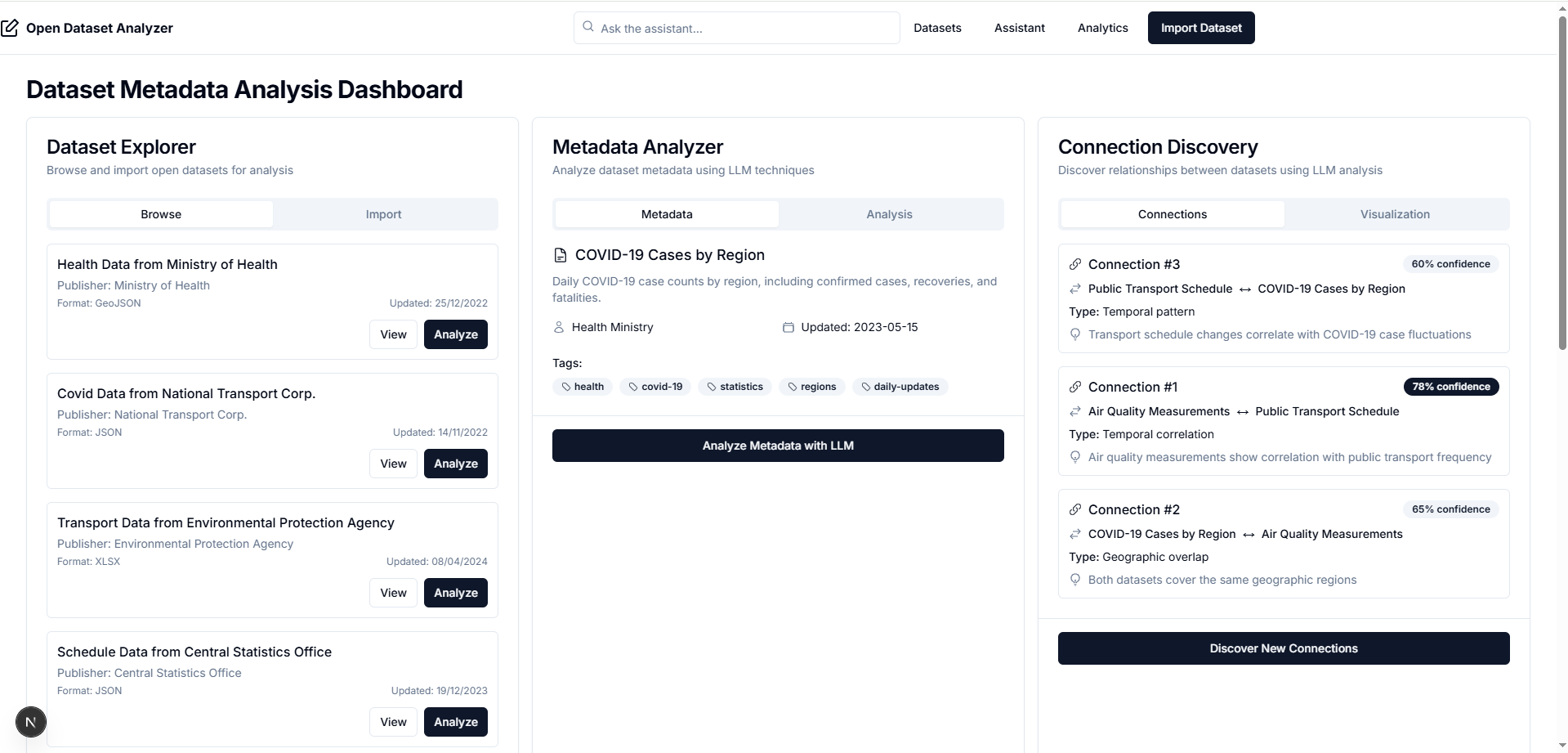


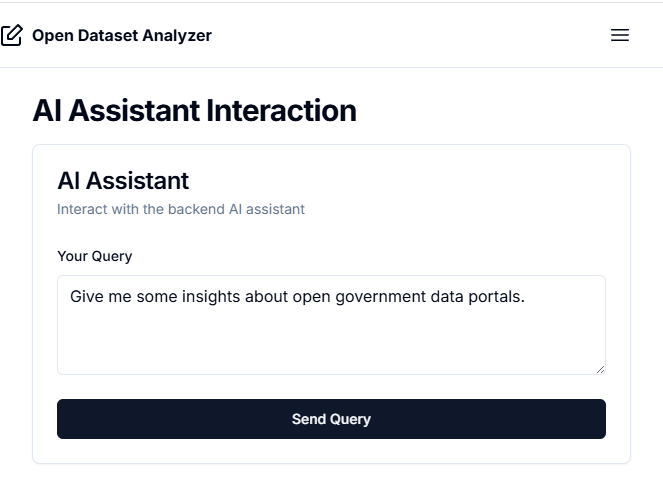
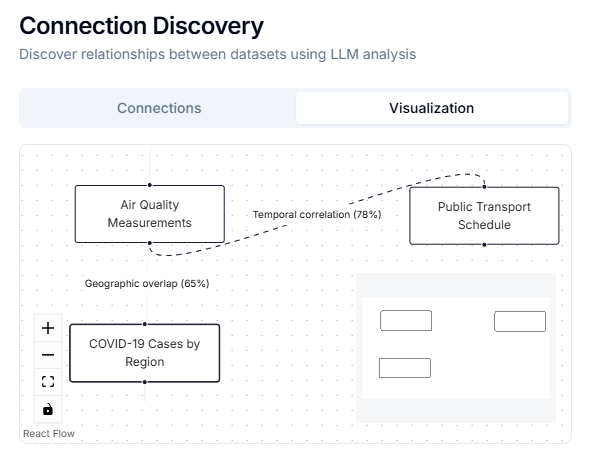


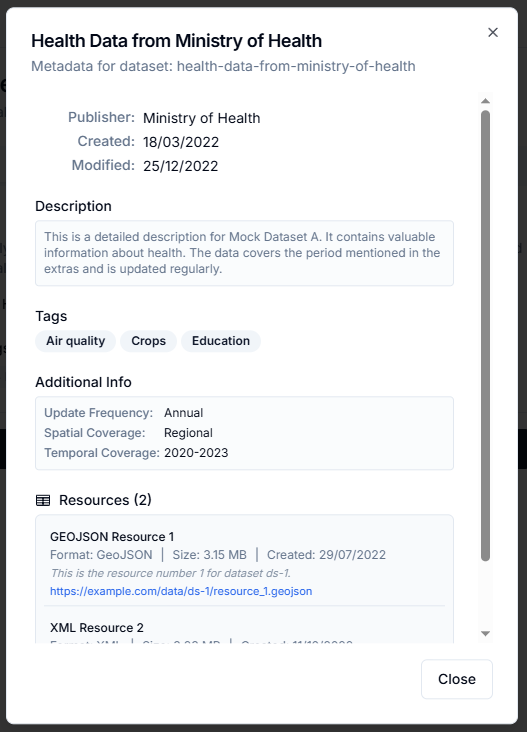


* SPARQL je moćan za precizno dohvaćanje provjerljivih, strukturiranih činjenica iz specifičnog izvora poput Wikidata, ali zahtijeva tehničko znanje i ne snalazi se dobro sa subjektivnošću. Rad s jezicima zahtijeva eksplicitno definiranje.
* LLM upiti su lakši za korištenje, razumiju prirodni jezik (uključujući hrvatski) i mogu bolje baratati dvosmislenošću, ali su manje pouzdani za činjeničnu točnost, nisu vezani za jedan izvor podataka i mogu generirati netočne informacije.

Web Aplikacija:

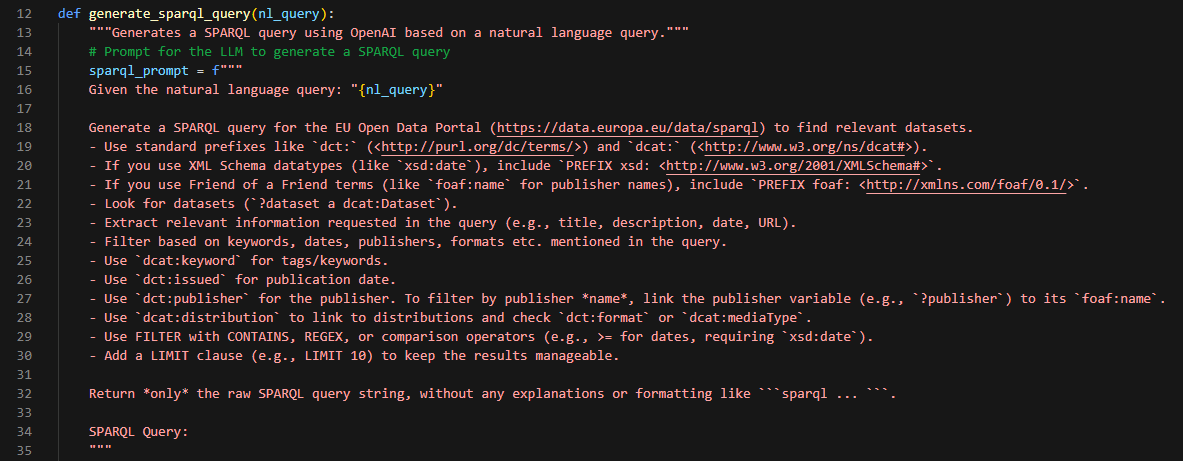


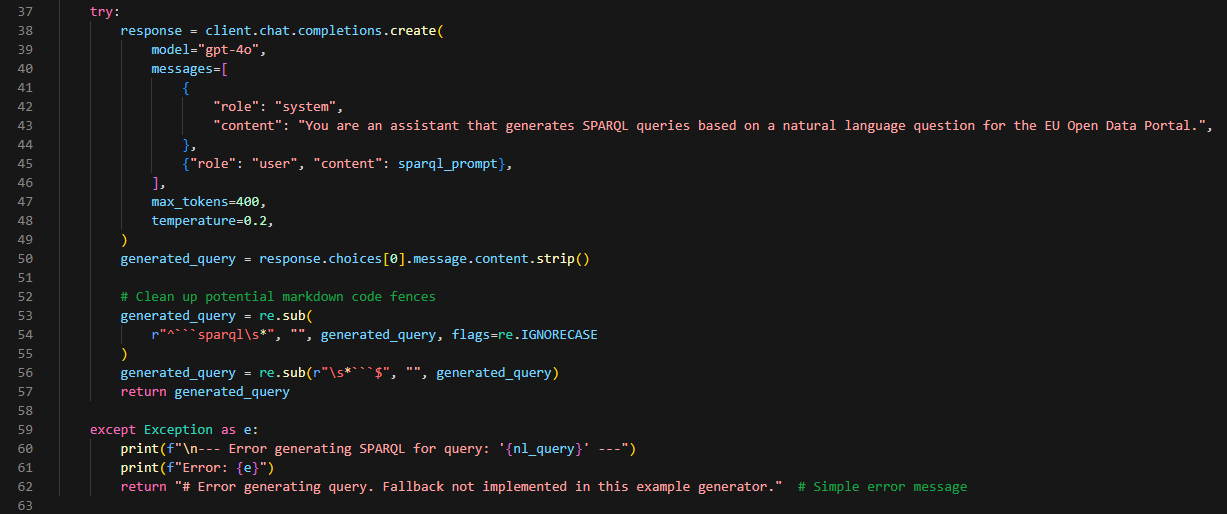


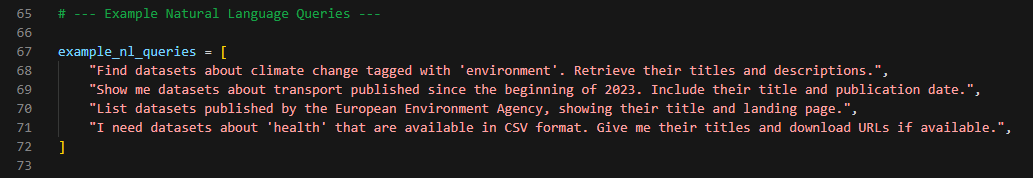


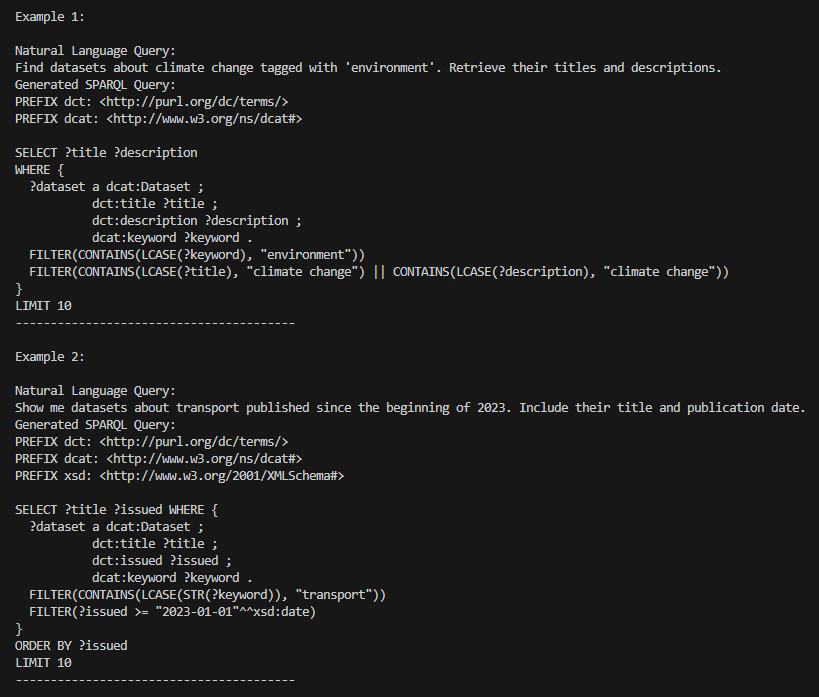
10.4.2025.

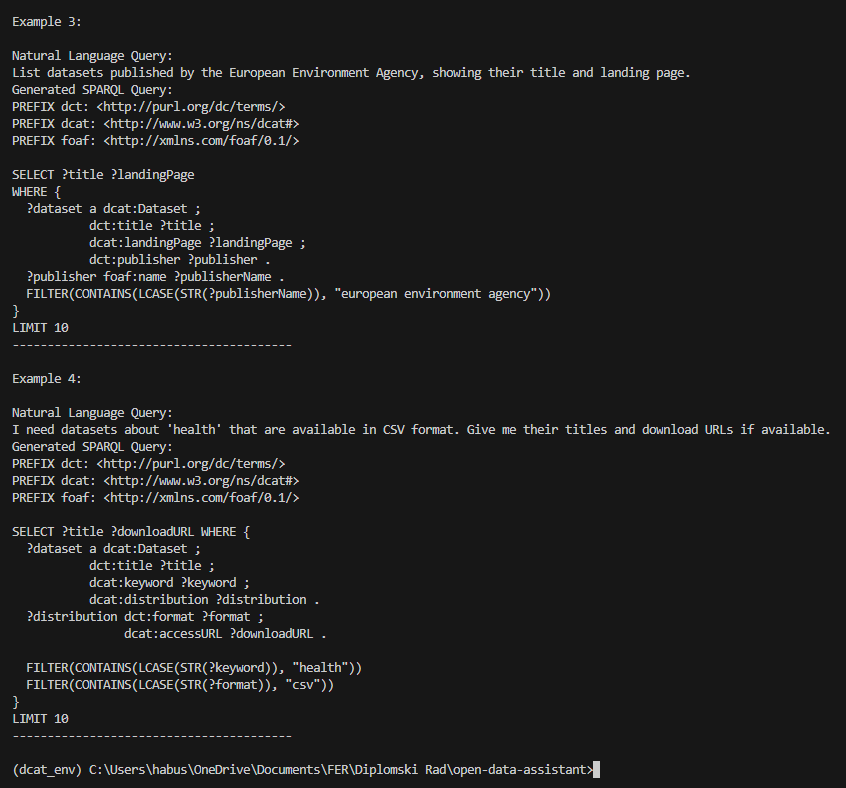
* sastanak za tjedan dana - do sada napravljeno primijenit na DCAT podatke s nekog od portala
  + pogledati dostupnost na hr portalu
  + pogledati neke portale s konzistentim korištenjem DCAT opisa skupova podataka . EU, engleski, nizozemski, USA
* <https://github.com/ckan/ckanext-dcat>
* **obrana diplomskog do uključivo 16.07.**

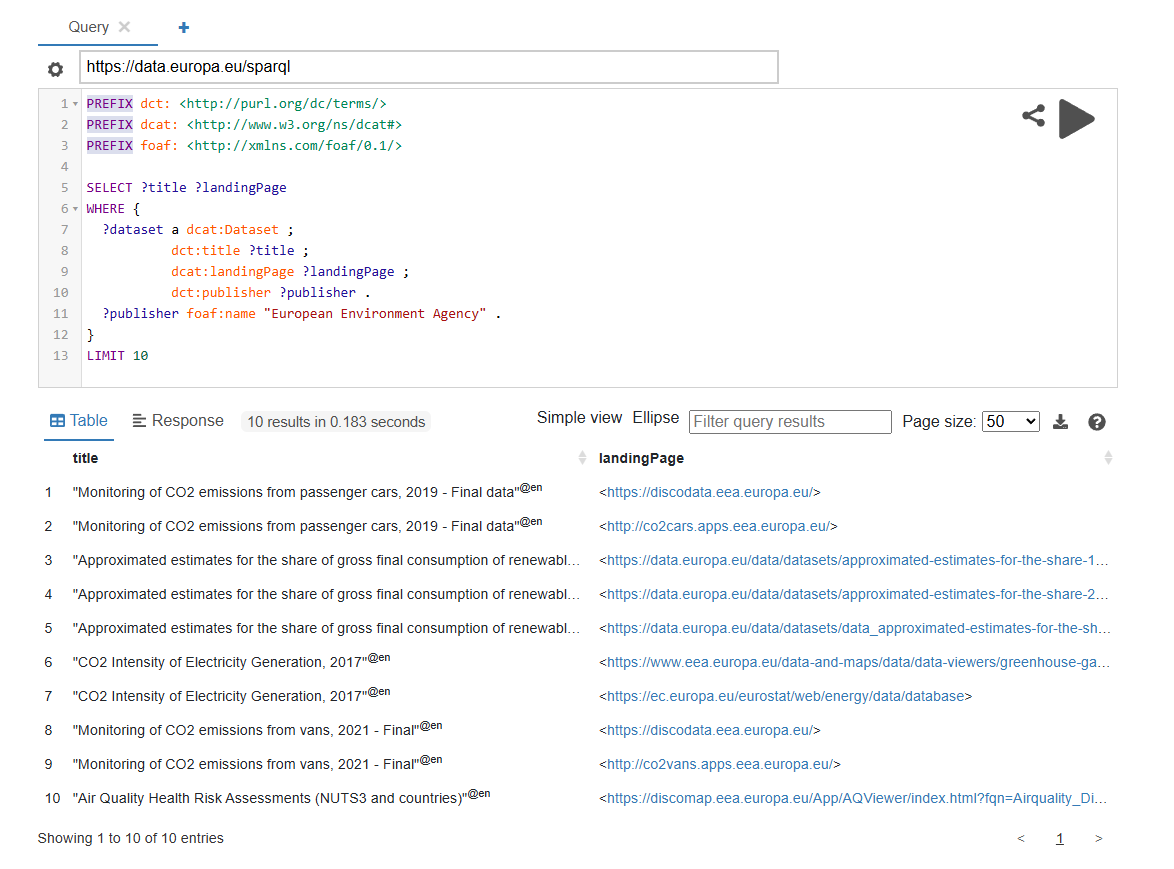


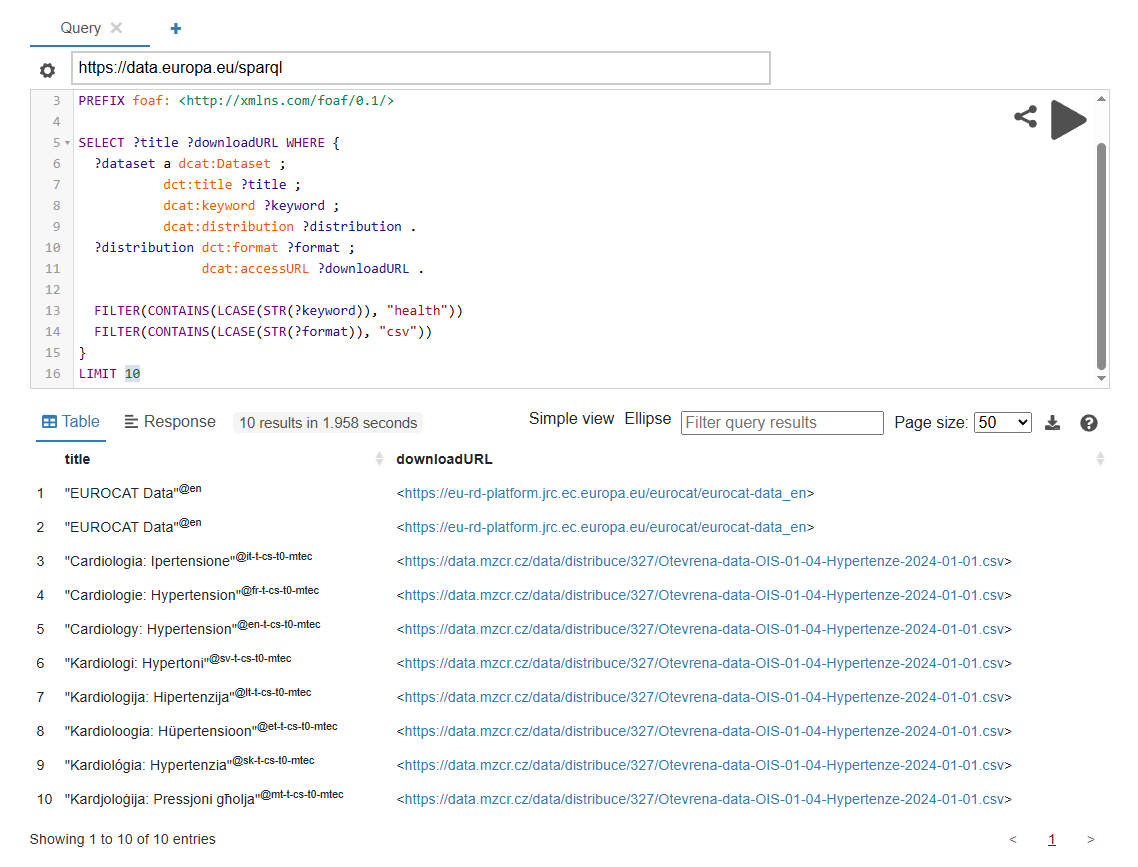








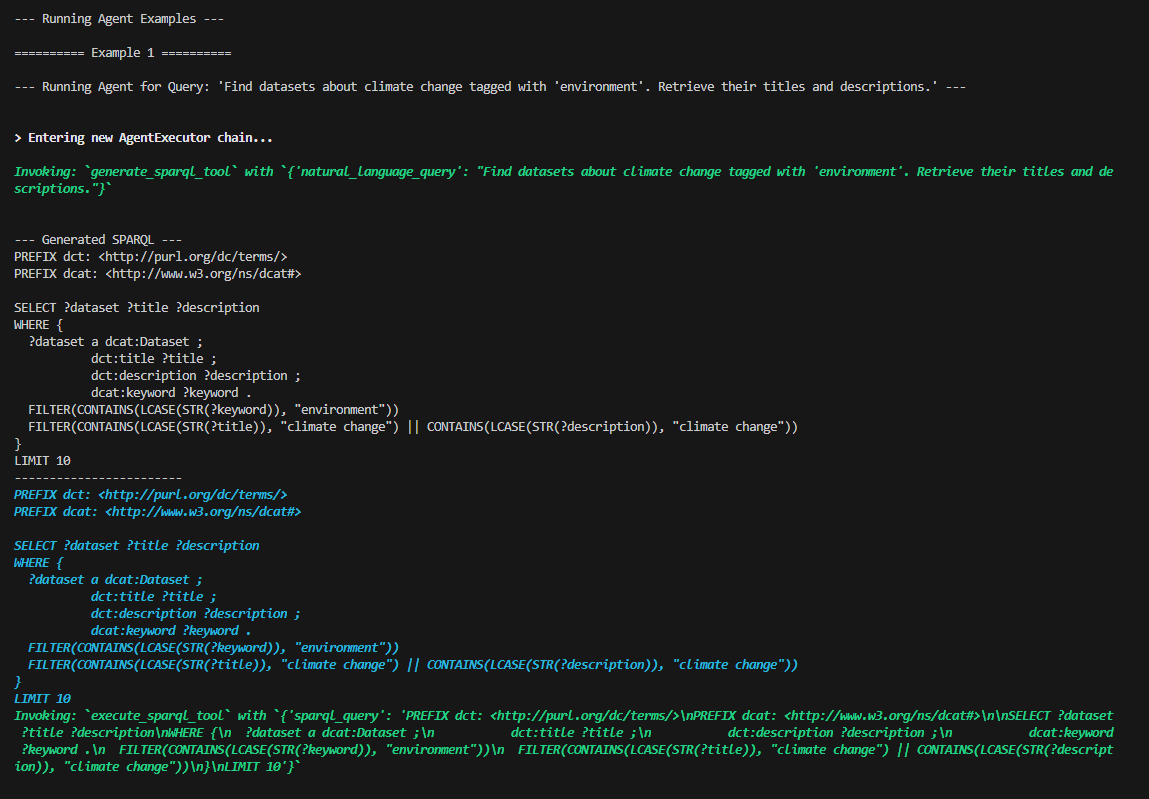


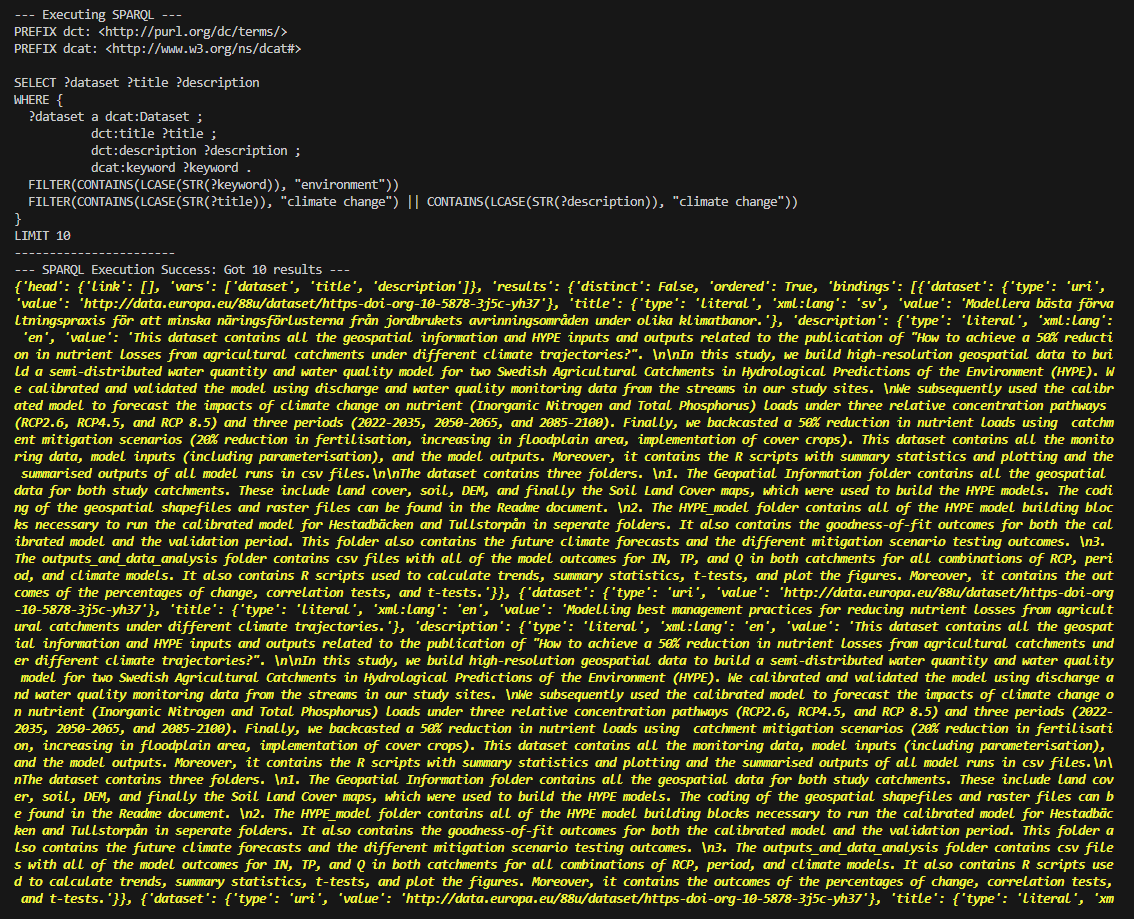


2025-04-18

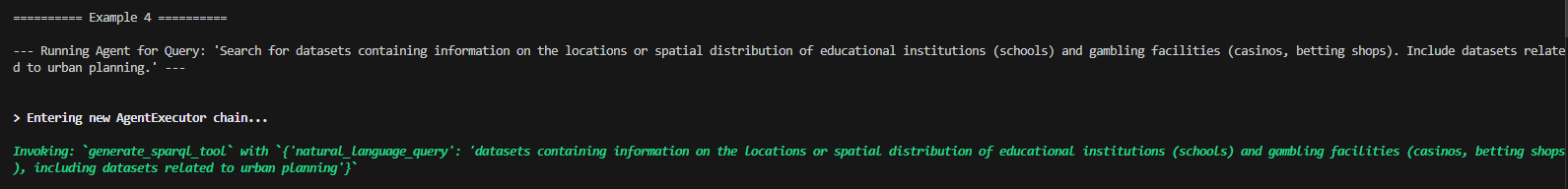
—--------------

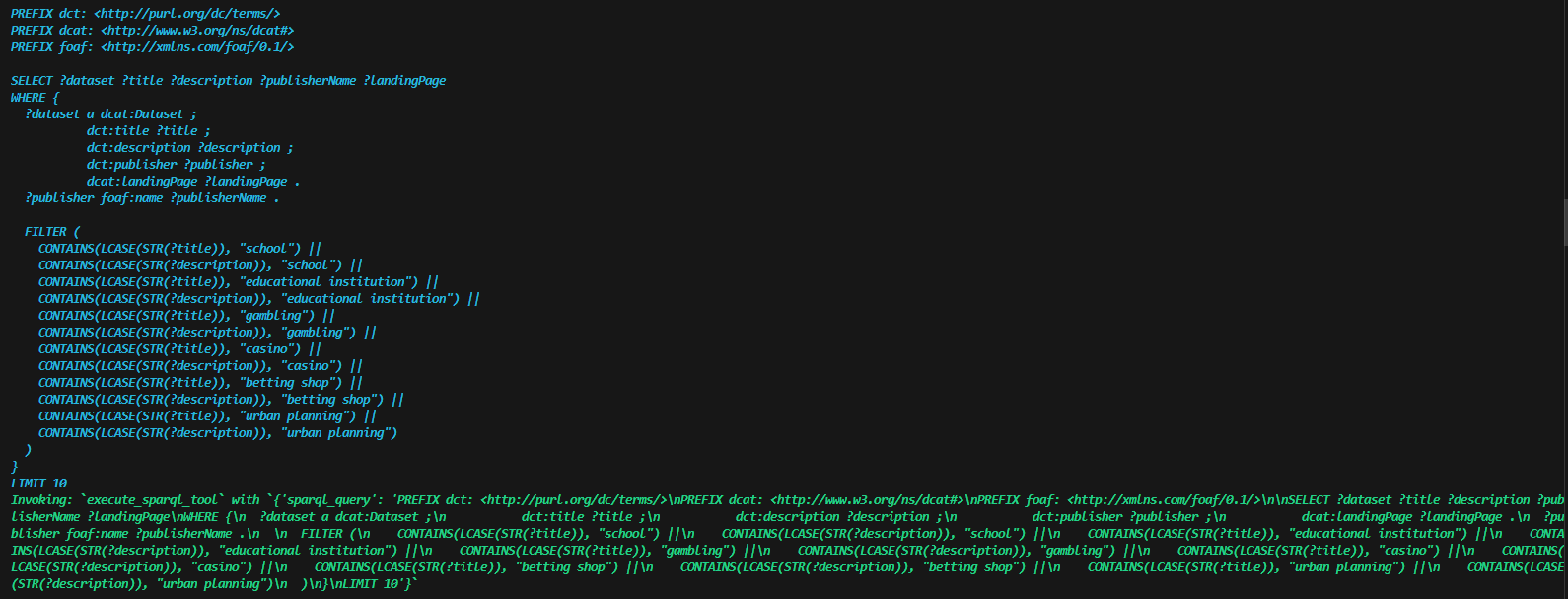
* razmisliti o “chain-of-thoughts” koji LLM treba izgenerirati (bez akcija, za sada) kako bi nam objasnio kako bi povezao dva (ili tri) dataseta iz upita:
  + stvaranje upita za dohvaćanje metapodataka data setova
  + analiza metapodataka
  + uparivanje podataka iz datasetova na osnovu neke logike
    - značenje podataka (stupci)
    - tipovi podataka (geojson, koordinate, imena ulica, …)
  + prezentacija podataka (rezultirajući dataset) - koji podaci iz oba (ili tri) skupa podataka su bitni korisniku (ako je ovaj korak uopće potreban)
  + sve ove scenarije razraditi na cca. 10-tak upita (koji su dovoljno različiti), koristiti raspoložive metapodatke s portala
  + razmisliti o samo-popravljanju SPARQL upita, kao kod generiranja SQL upita iz prirodnog jezika

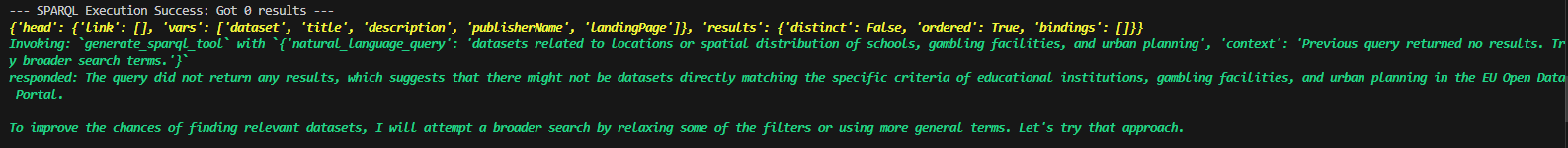


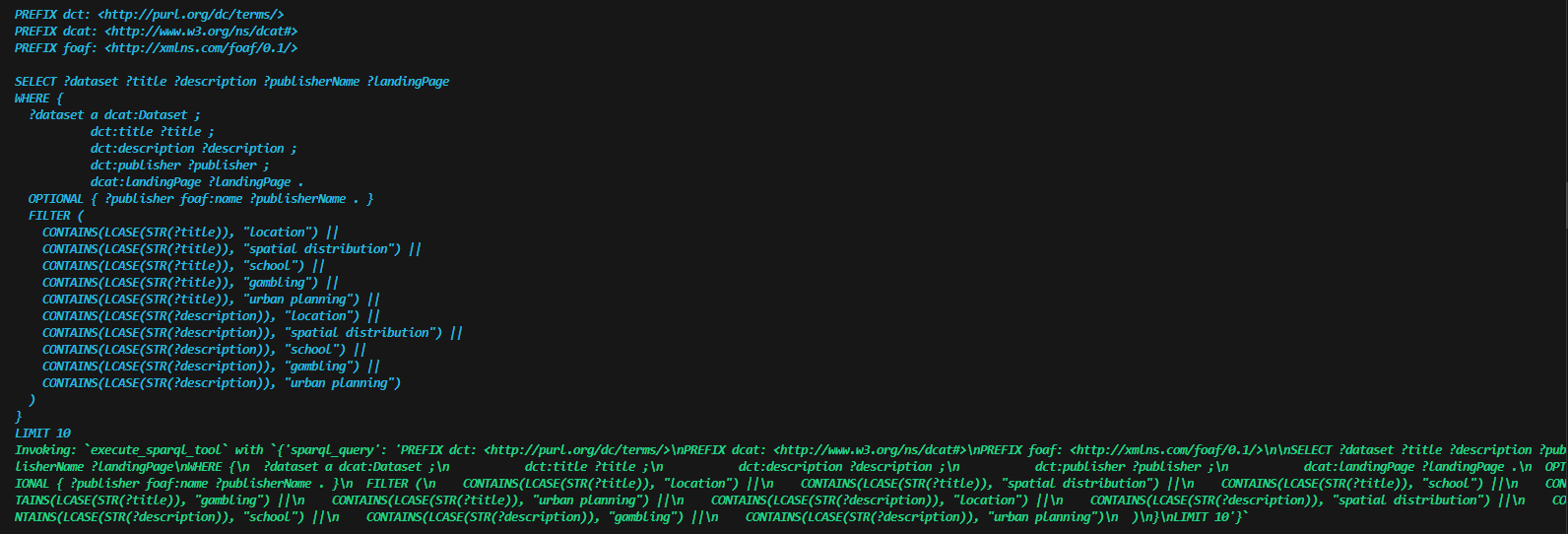






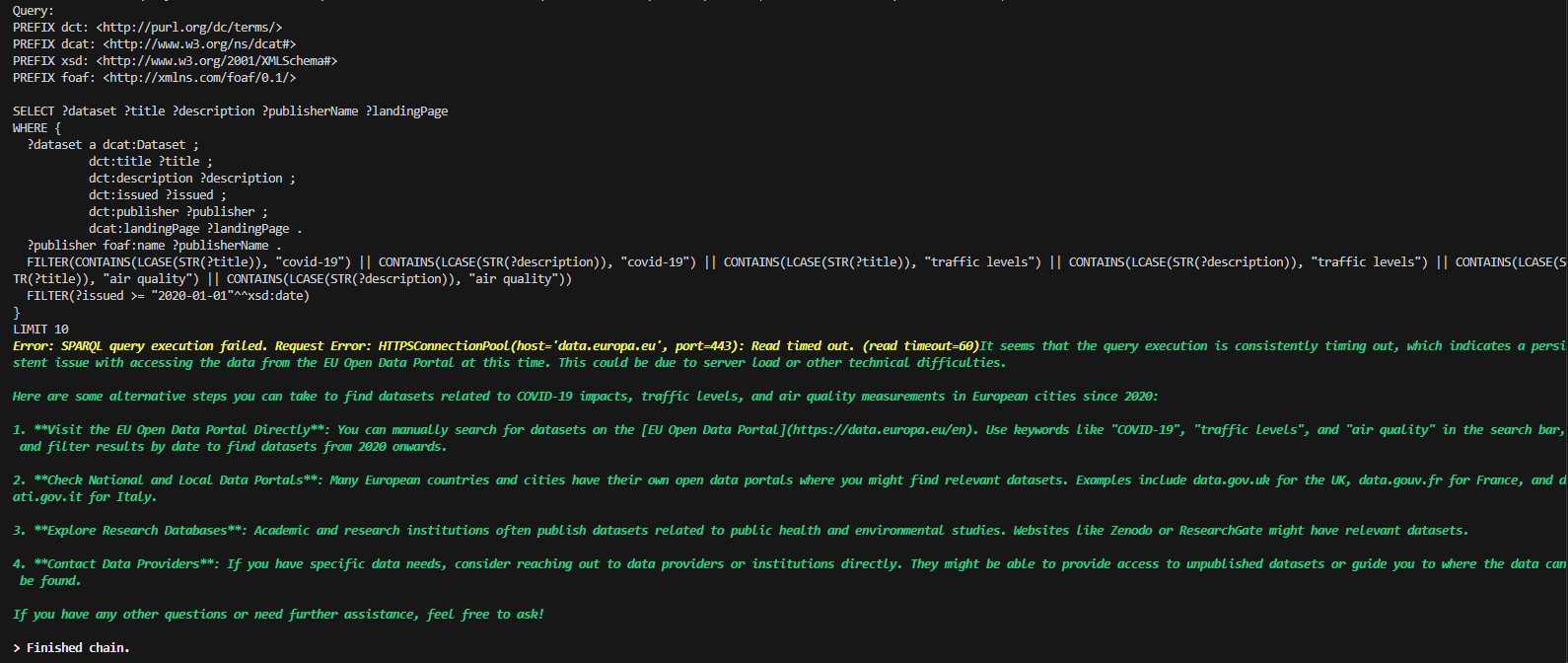


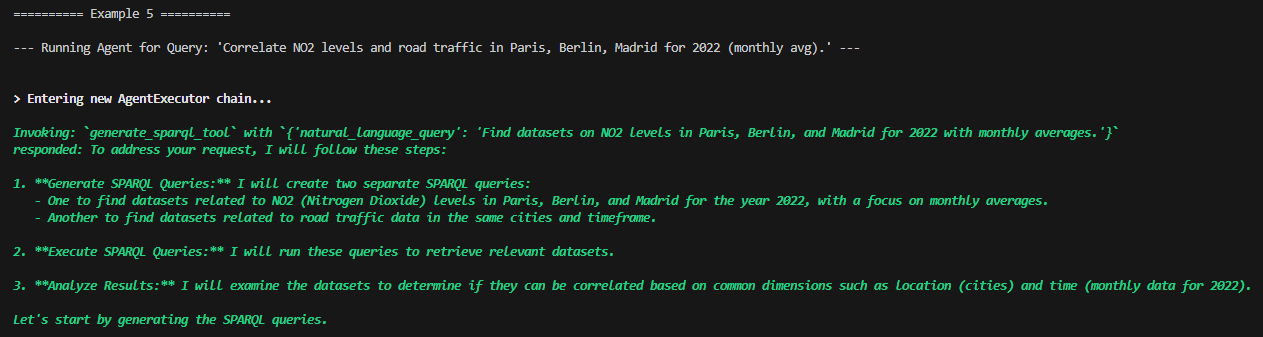


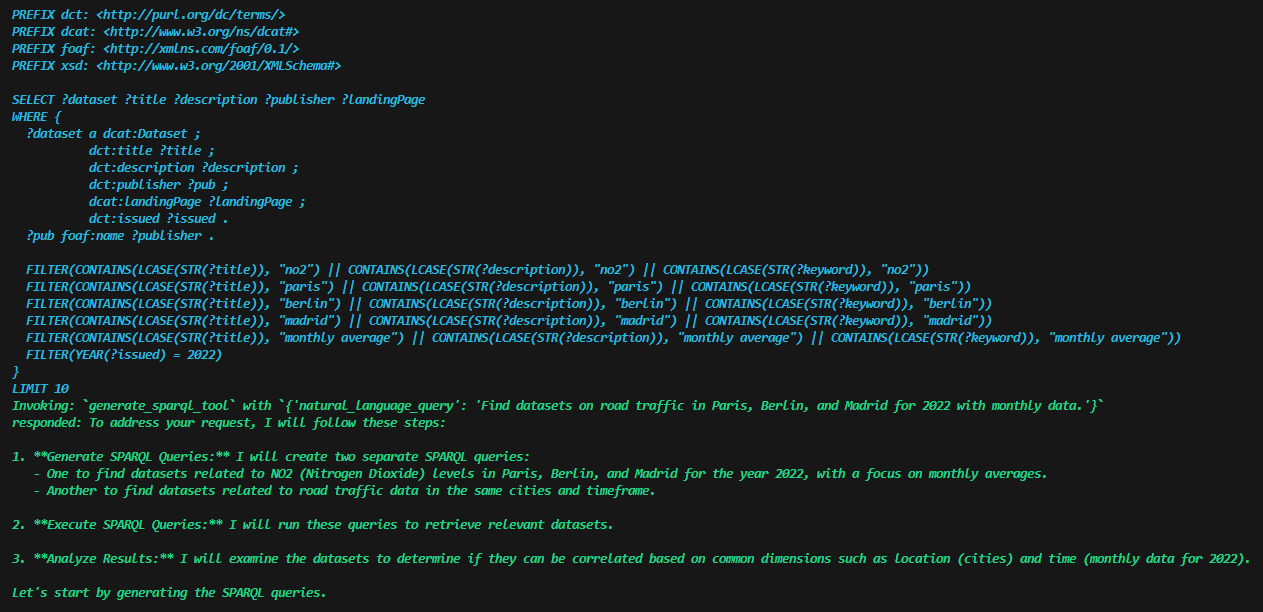




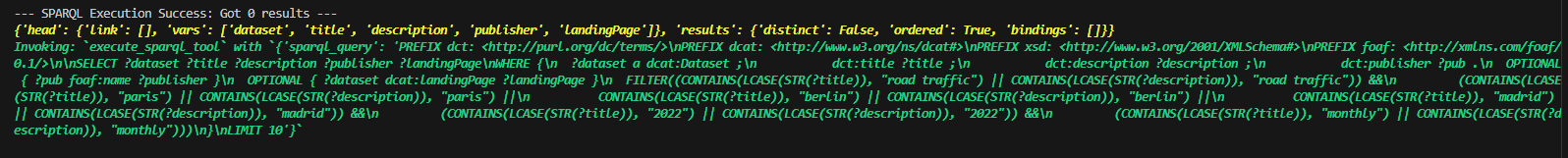


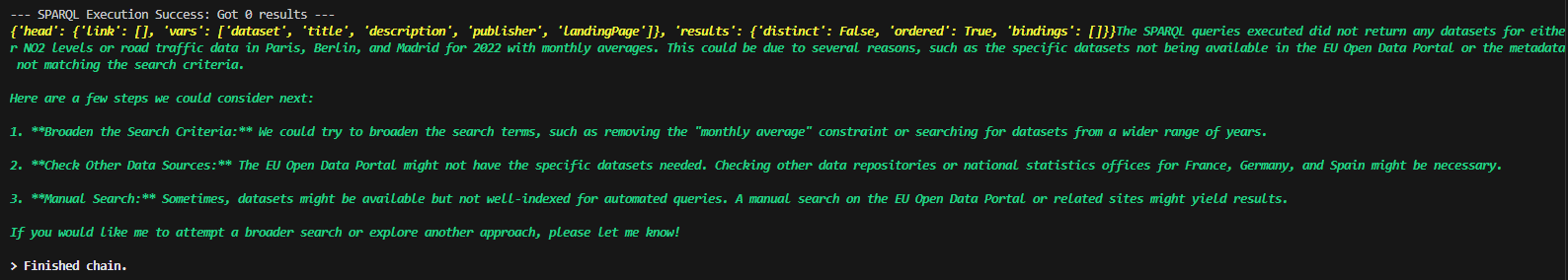


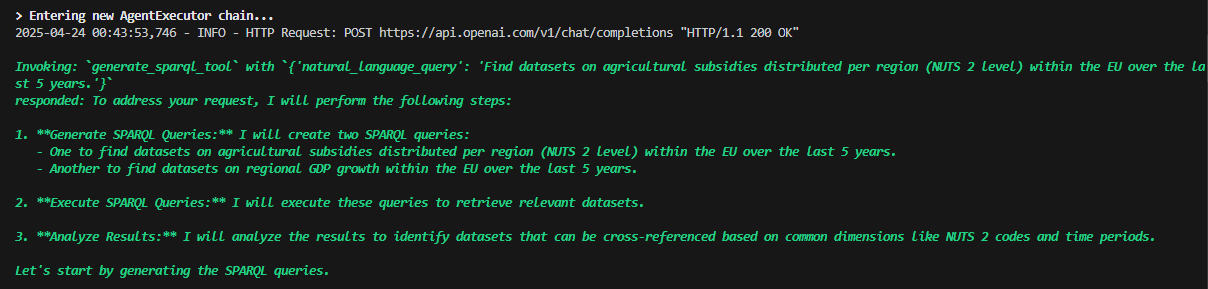


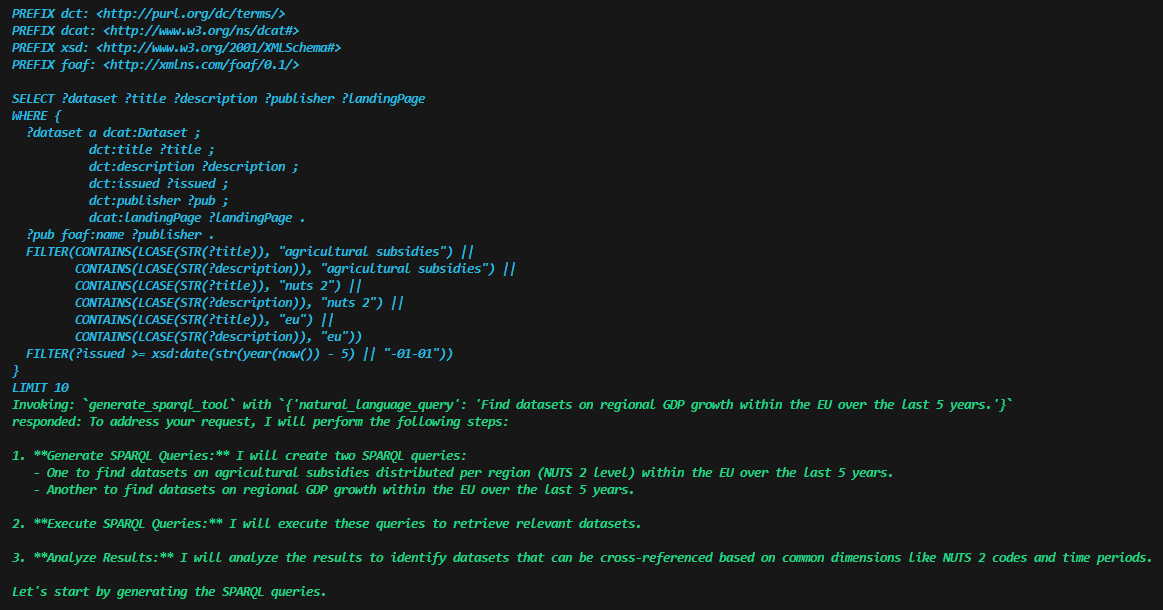




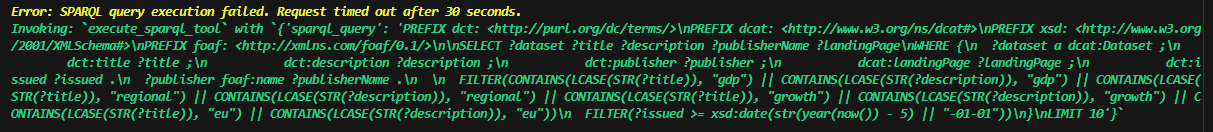




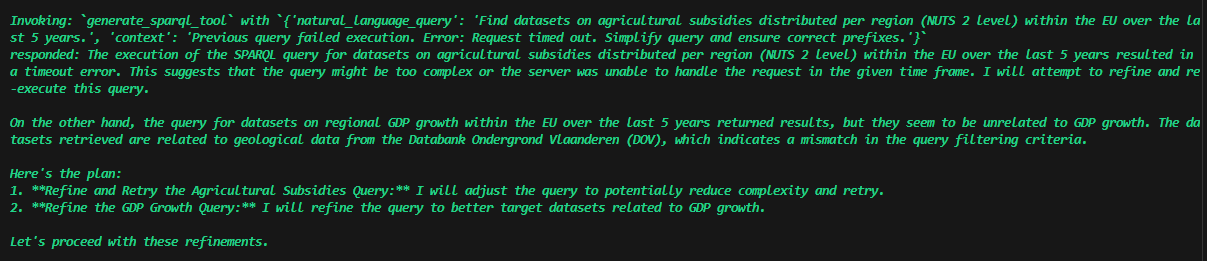


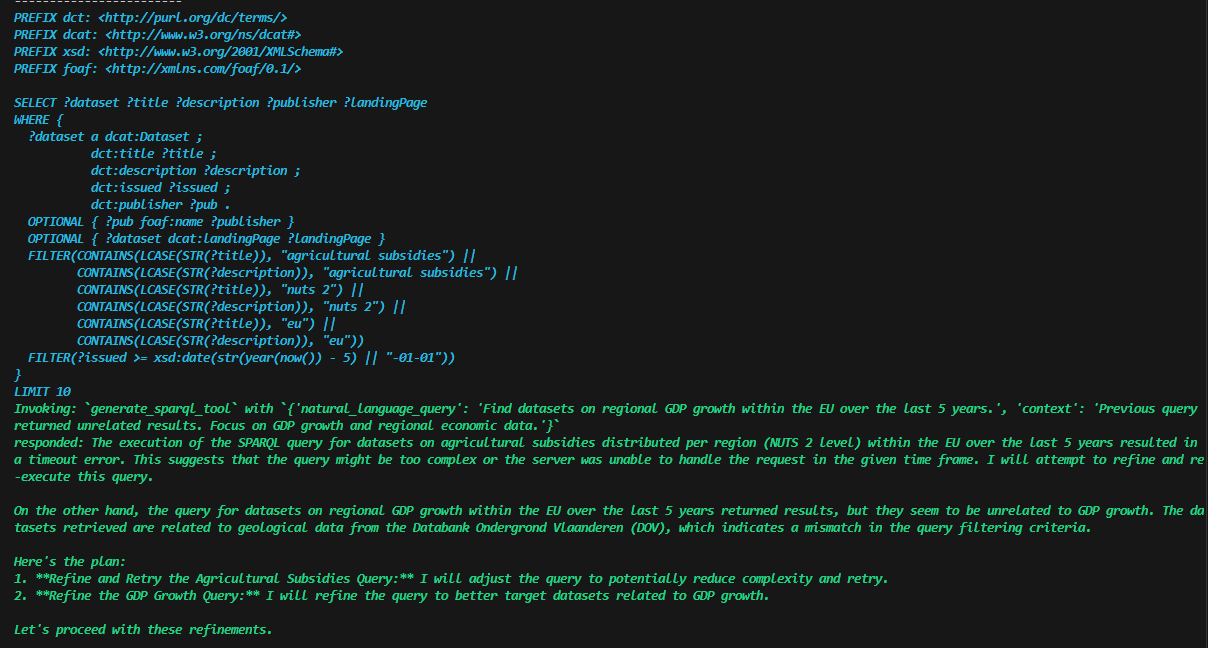


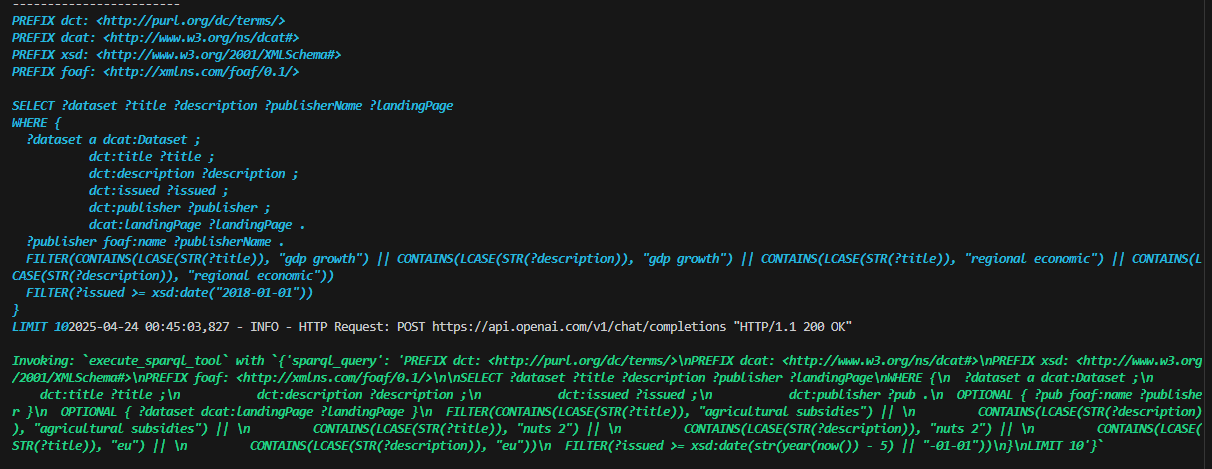


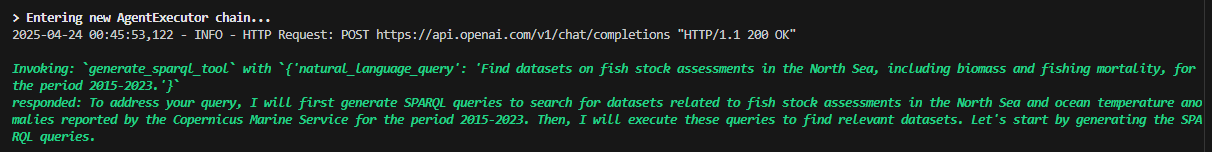


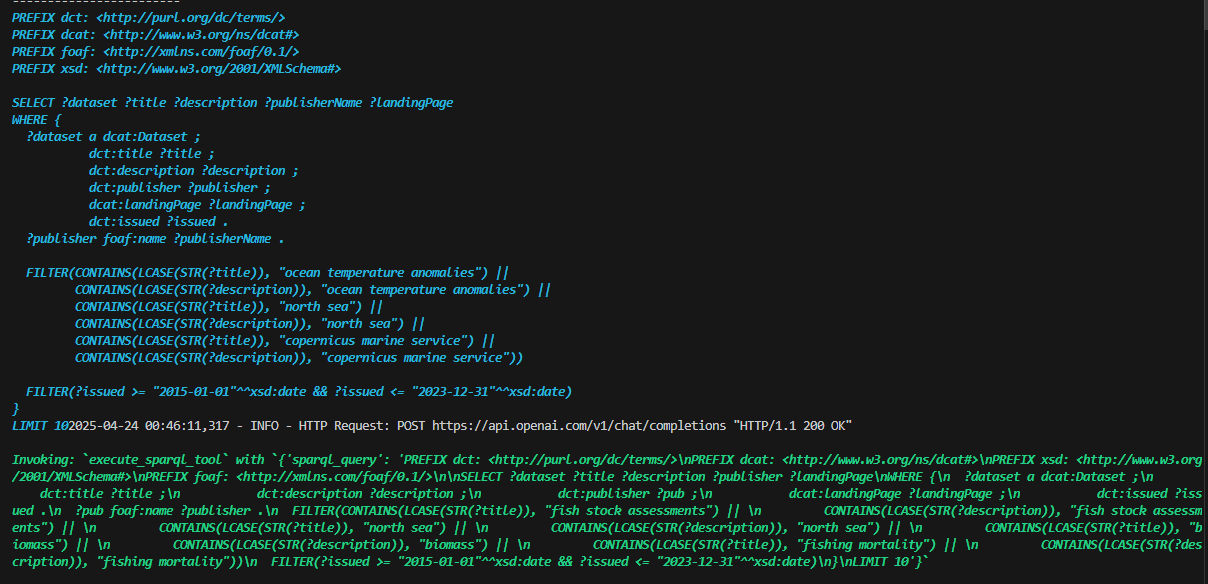


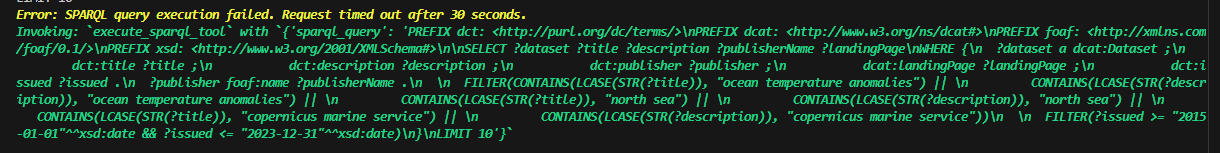






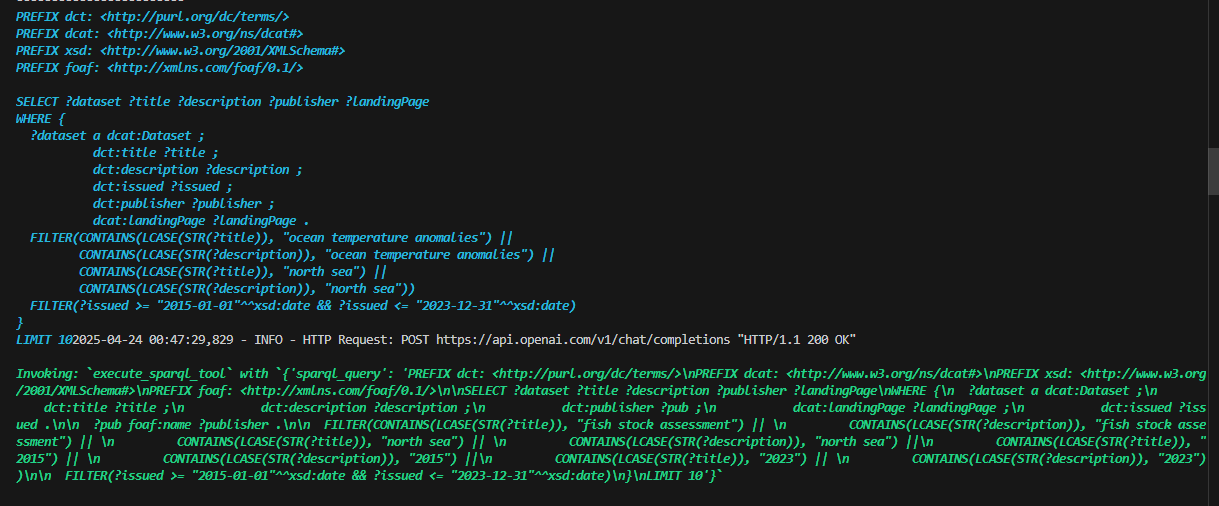


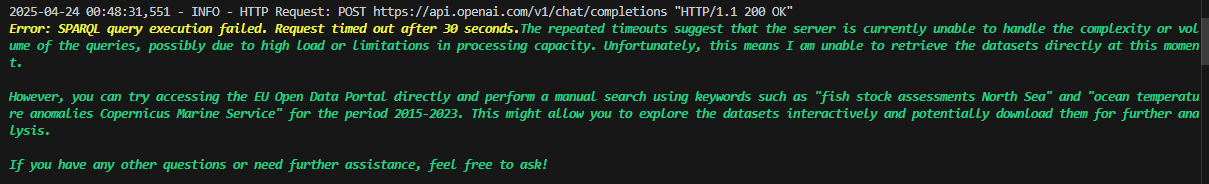


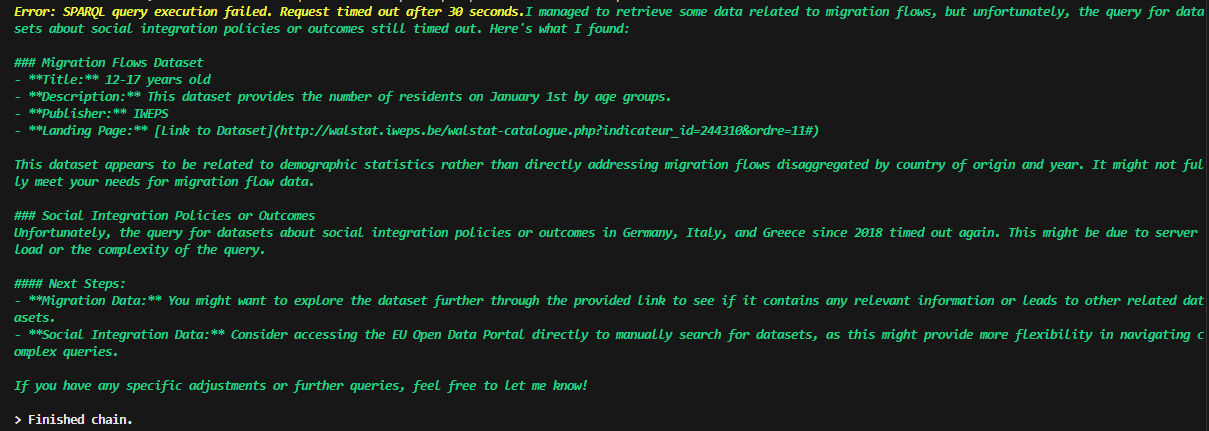
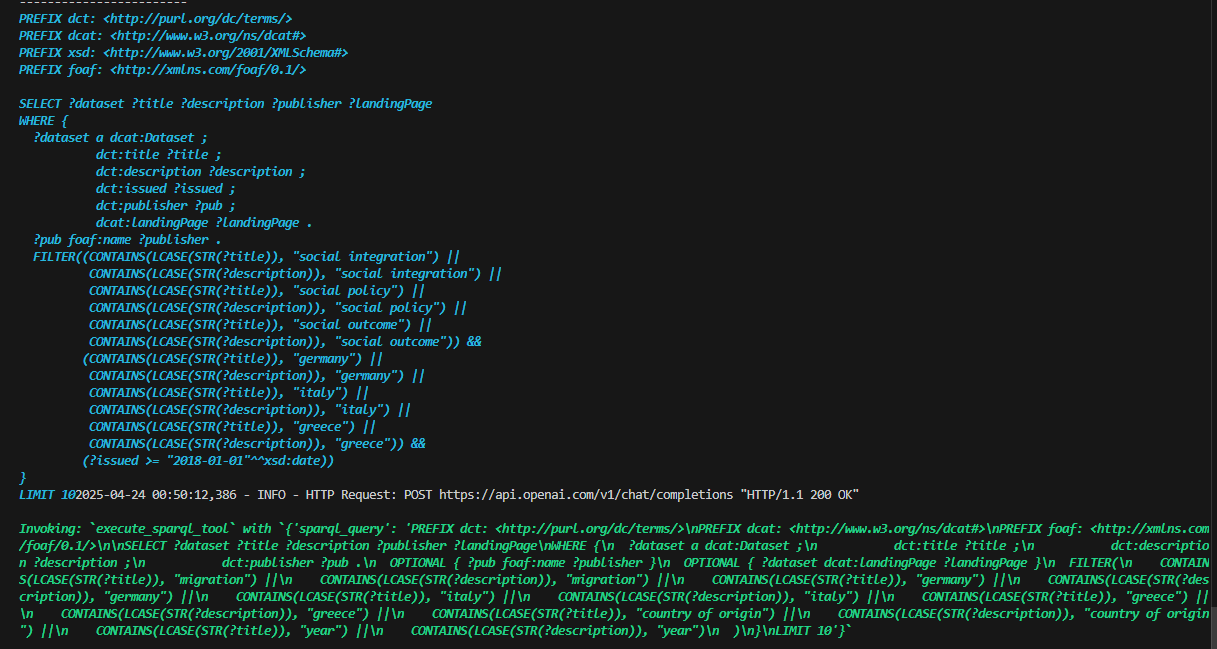
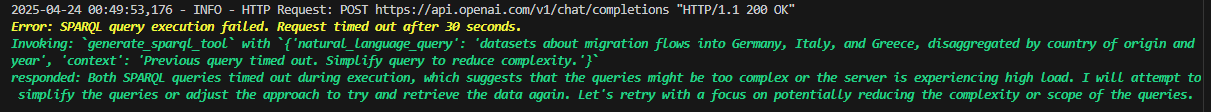
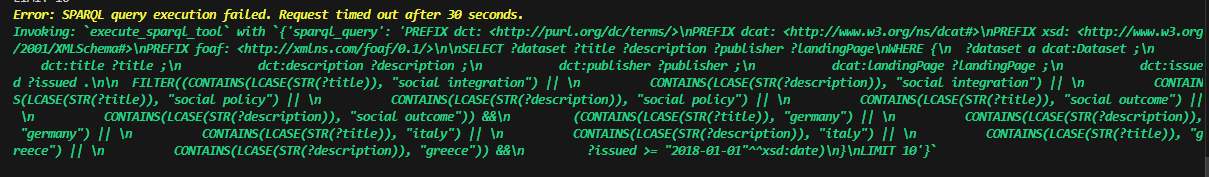
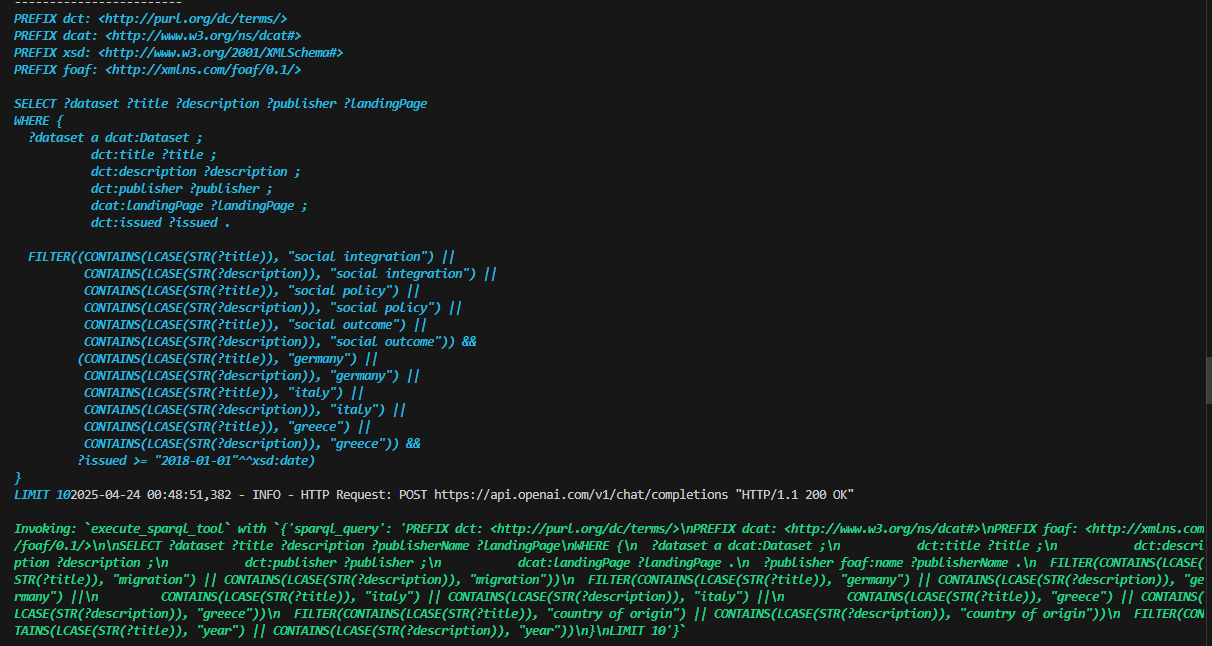
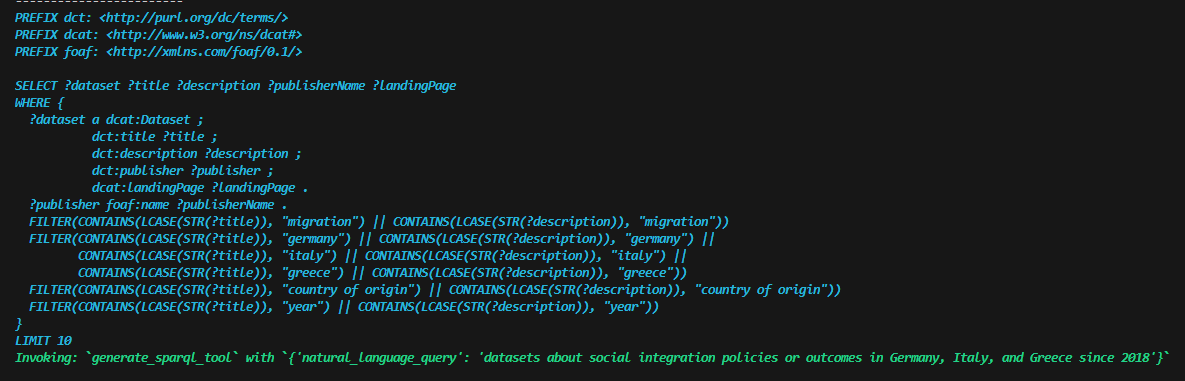
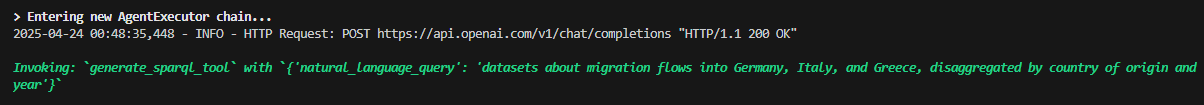




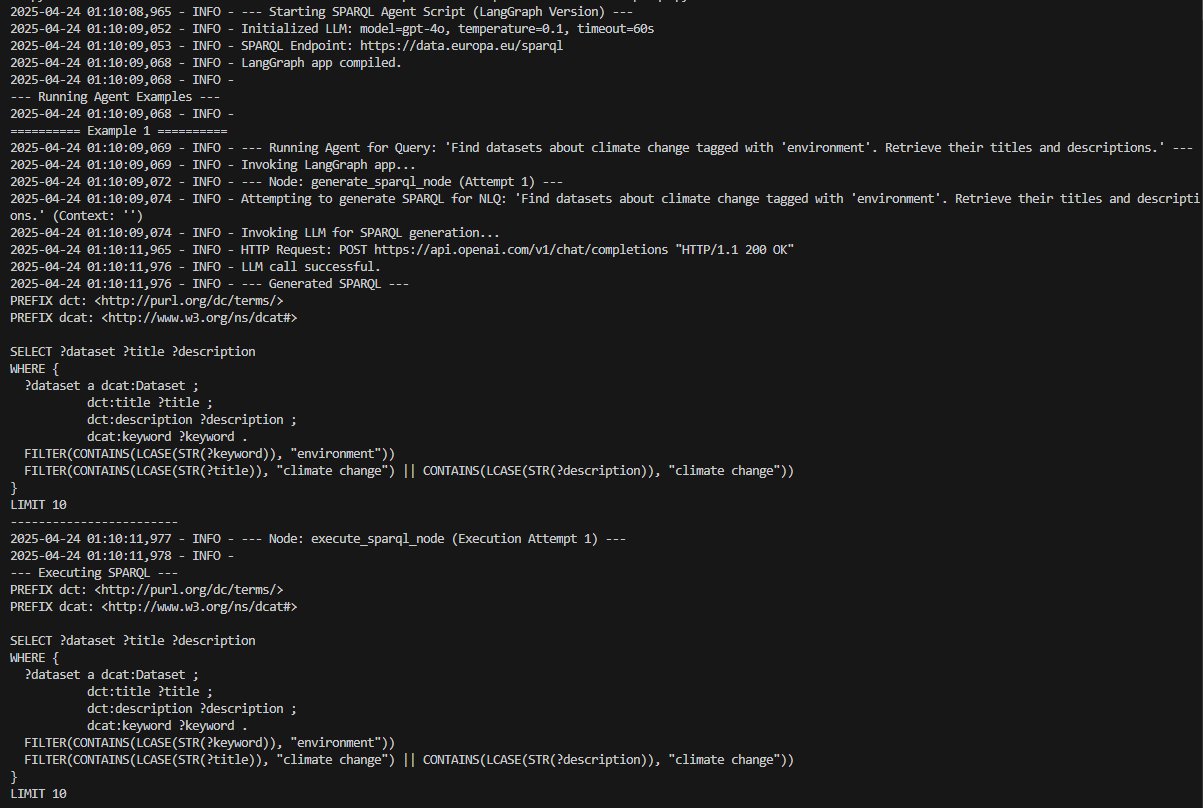


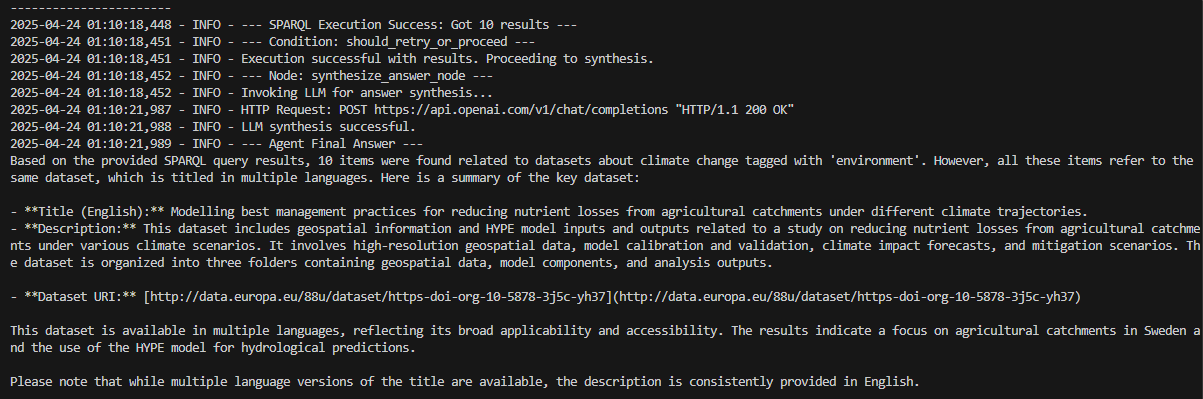






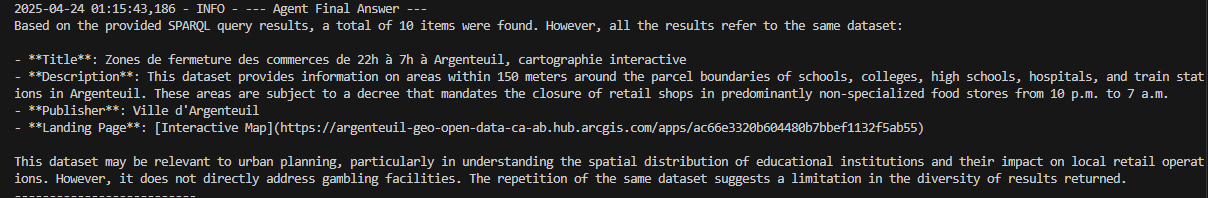
<https://walstat.iweps.be/walstat-catalogue.php?indicateur_id=244310&ordre=11#>





<https://data.europa.eu/data/datasets/https-doi-org-10-5878-3j5c-yh37?locale=en>





<https://argenteuil-geo-open-data-ca-ab.hub.arcgis.com/apps/ac66e3320b604480b7bbef1132f5ab55/explore>

2025-04-28

—--------------

* SPARQL upiti: poboljšati oporavak od greške - ako se desi greška, vratiti LLMu poruku s opisom greške, vidjeti kako će se ponašati glede oporavka
* isprobati pristup od jednostavnijeg ka složenijim upitima
* napraviti da ne vraća isti dataset u više jezika - u upitu navesti da vraća engleski ili “materinji” dataset ako nema engleskog
* API - vidjeti zašto ne generira plan?
* natjerati sustav da radi za složenije upite (još uvijek povezivanje samo dva dataseta u jedan par)
* počnite pisati rad (uvodi u tehnologije)

2025-05-09

—--------------

* Link na rad: <https://www.overleaf.com/read/pkgrgqmqbjks#fc86b6>

⇒ pogledati: <https://inria.hal.science/tel-01130622v1>

<https://arxiv.org/html/2410.06062v2>  
  
<https://chat.expasy.org/>

2025-05-26

—--------------

* nastavio sam s pisanjem rada
* dodao sam vektorsku bazu ChromaDB i osposobio RAG sustav

