Bsc

2025.03.24

Querying and Visualizing Air Pollution Data Using Linked Open Data and Python

Leyan CHENG & Yuxin GONG & Jiani XU

Catalogue

01 Introduction

02 Exploratory Analysis of the Database

03 Data Analysis: SPARQL Queries

04 Conclusion

01

Introduction

Introduction

- In recent years, air pollution has become a major environmental issue around the world. It doesn’t just affect the air we breathe—it also harms our health, nature, and even the climate. As cities grow and industries develop, sources of pollution like cars, factories, and wildfires are becoming more common.

- We now have better ways to search and understand data about air pollution from different places and topics. In this project, we use SPARQL to pull structured information from Wikidata, helping us explore and analyse.

- By combining this kind of smart data search with Python visualizations, we want to show a clearer picture of air pollution in open data and find patterns that can help raise awareness and guide better decisions.

Exploratory Analysis of the Database

02

Origin and Structure of the Data

Wikidata

a collaborative, open knowledge base

SPARQL

using SPARQL queries through the Wikidata Query Service

Customized queries

instead of relying on pre-built dataset

The data structure

a triple model: subject－predicate－object

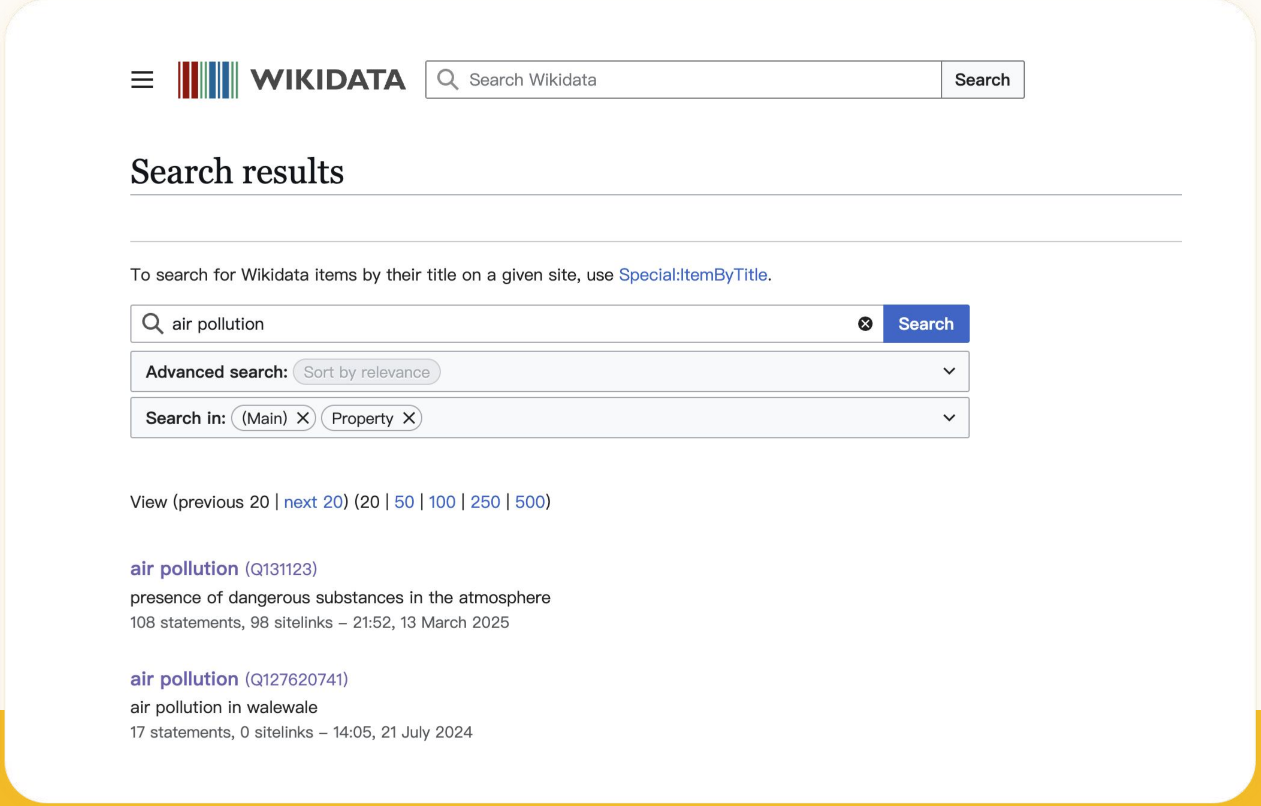
Results

The results of the queries were returned in tabular format

Analysis and visualization

This format enabled further analysis and visualization using Python

Data Related to Pollution



Search results

To search for Wikidata items by their title on a given site, use Special:ItemByTitle.

air pollution

Advanced search: Sort by relevance

Search in: (Main) ✕ Property ✕

View (previous 20 | next 20) (20 | 50 | 100 | 250 | 500)

air pollution (Q131123)
presence of dangerous substances in the atmosphere
108 statements, 98 sitelinks – 21:52, 13 March 2025

air pollution (Q127620741)
air pollution in walewale
17 statements, 0 sitelinks – 14:05, 21 July 2024

We started by searching for “air pollution” in the wikidata search bar to find the representation code.

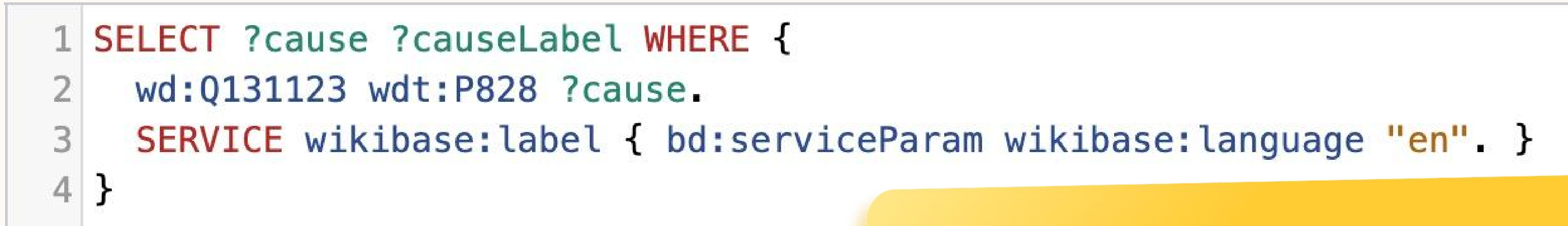
We selected different types of information for better understanding of air pollution.（Causes、Effects、Policies, laws and Scientific studies）

These linked data help us explore and visualise air pollution directly from multiple perspectives.

03

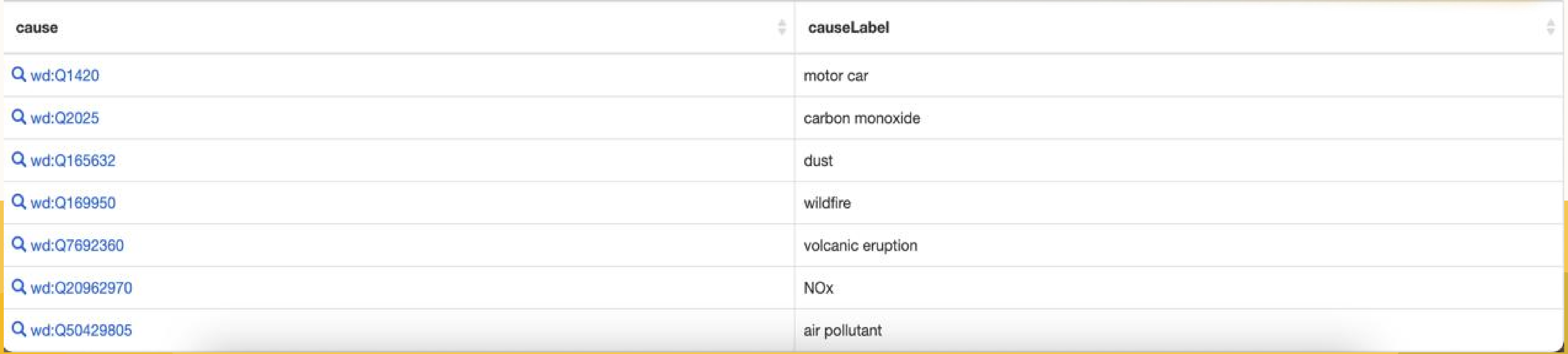
Data Analysis: SPARQL Queries

Sources of air pollution



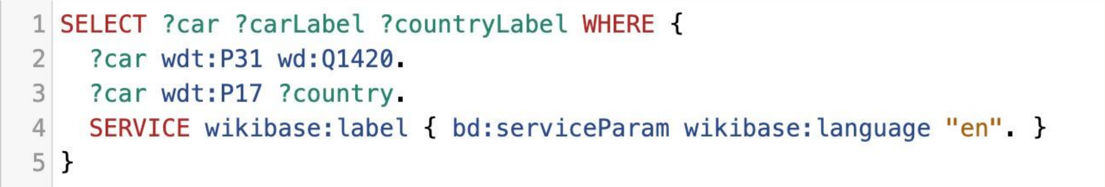
```
SELECT ?cause ?causeLabel WHERE {
  wd:Q131123 wdt:P828 ?cause.
  SERVICE wikibase:label { bd:serviceParam wikibase:language "en". }
}
```

we can see that the two main causes of air pollution are motor car and wildfire which correspond to the two types of human activities and natural sources of pollution respectively

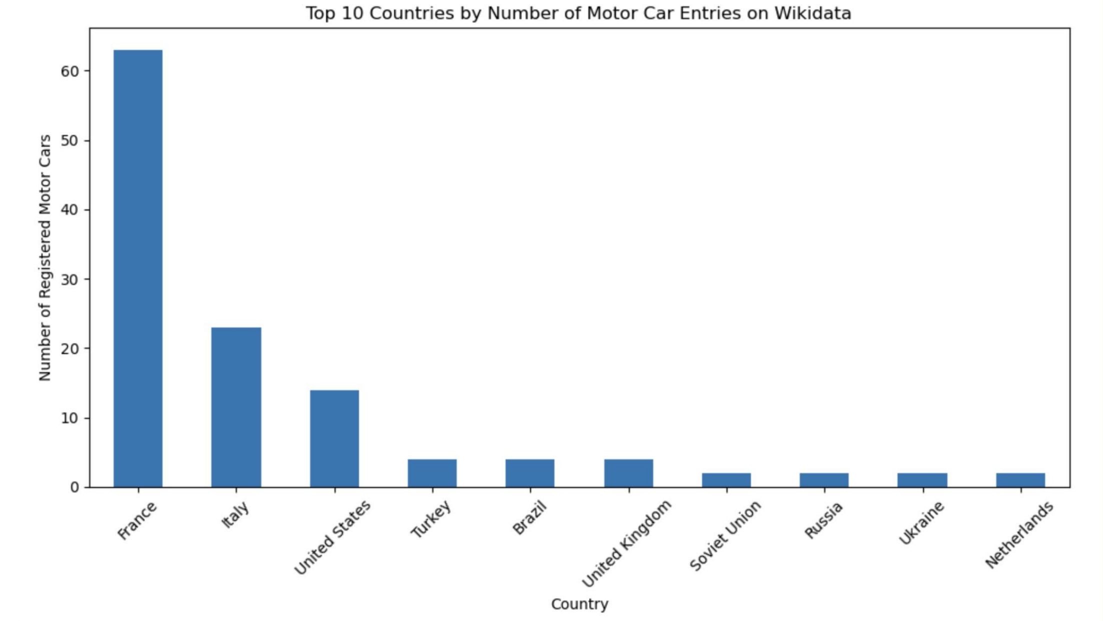


cause	causeLabel
wd:Q1420	motor car
wd:Q2025	carbon monoxide
wd:Q165632	dust
wd:Q169950	wildfire
wd:Q7692360	volcanic eruption
wd:Q20962970	NOx
wd:Q50429805	air pollutant

Enquire about the number of motor cars’ model in each country



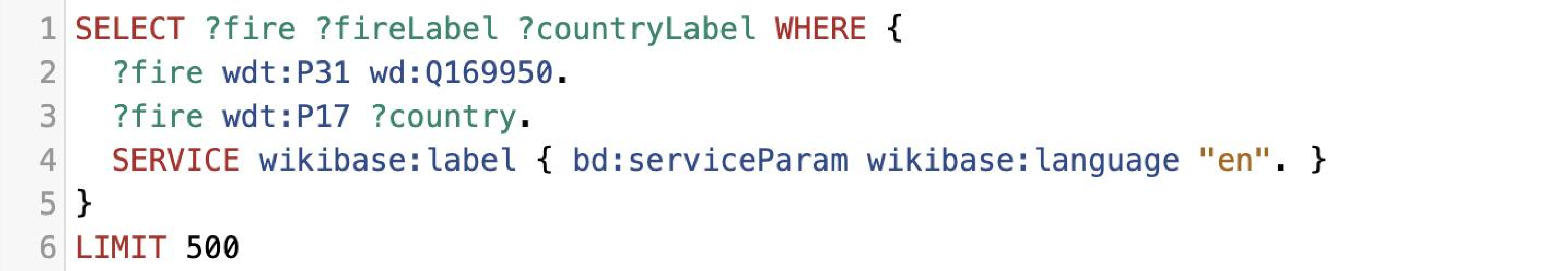
```
SELECT ?car ?carLabel ?countryLabel WHERE {
  ?car wdt:P31 wd:Q1420.
  ?car wdt:P17 ?country.
  SERVICE wikibase:label { bd:serviceParam wikibase:language "en". }
}
```



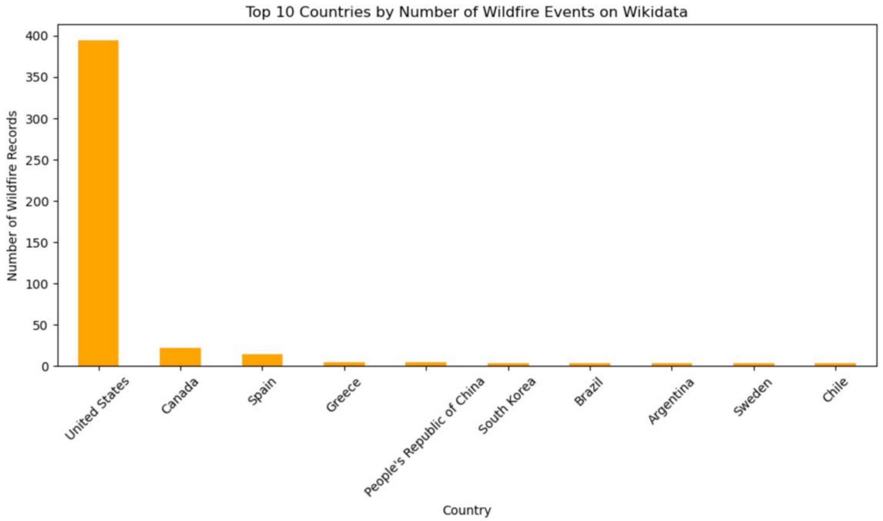
Motor vehicles from France are well recorded in semantic data sources. This suggests that transport-related emissions might be high. Countries like Italy and the United States also have many entries, which matches their known industrial growth and urban development. These countries often face problems like air pollution, greenhouse gas emissions, and city smog.

On the other hand, countries like Turkey, Brazil, and Russia have fewer records. This could mean that Wikidata has less data from these places, or that these countries have fewer motor vehicles.

Enquire about the number of wildfire in each country



```
SELECT ?fire ?fireLabel ?countryLabel WHERE {
  ?fire wdt:P31 wd:Q169950.
  ?fire wdt:P17 ?country.
  SERVICE wikibase:label { bd:serviceParam wikibase:language "en". }
}
LIMIT 500
```

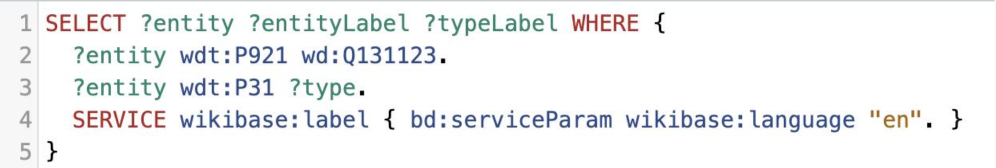


The large number of recorded wildfires in the United States indicates that wildfire seasons have the potential to have an even greater impact on regional and even global air pollution.

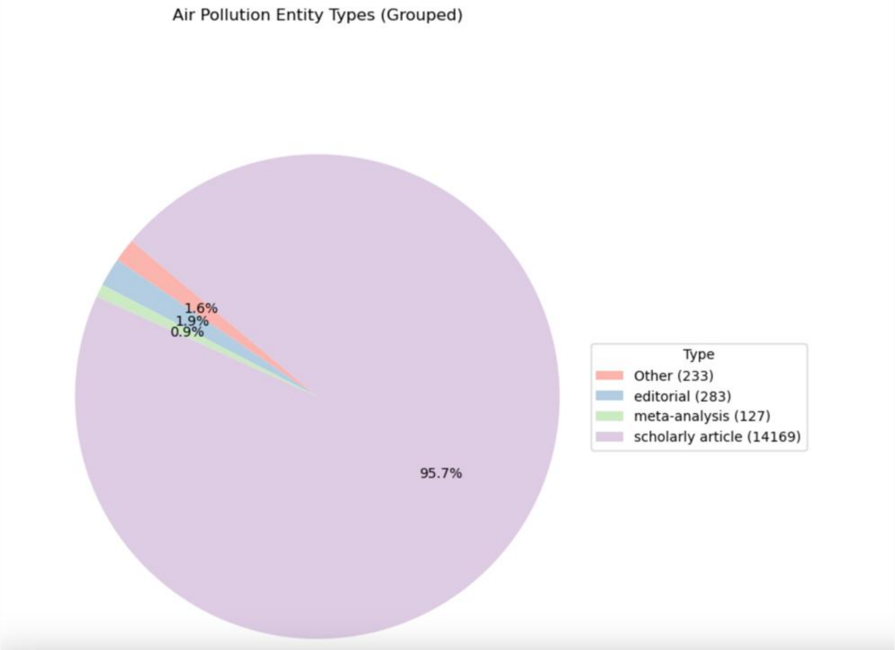
Countries such as Canada, Spain, Greece and Brazil are also known for the occurrence of seasonal or climate-induced wildfires, although they show fewer entries.

Wildfires are becoming more frequent as a result of climate change and deforestation, and their role in worsening air quality and contributing to global warming cannot be underestimated.

Air Pollution Topic Count number of typelabel



```
SELECT ?entity ?entityLabel ?typeLabel WHERE {
  ?entity wdt:P921 wd:Q131123.
  ?entity wdt:P31 ?type.
  SERVICE wikibase:label { bd:serviceParam wikibase:language "en". }
}
```

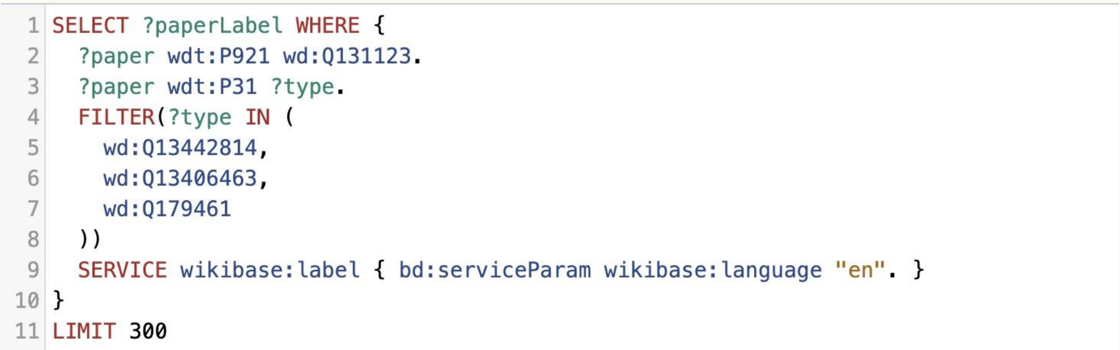


This pie chart shows the types of items related to air pollution. Most of them are academic articles, which make up 95.7% of the total. This means that air pollution is mostly studied in universities and research institutions, and many of the articles are peer-reviewed.

Other types, like editorials (1.9%), meta-analyses (0.9%), and others (1.6%), appear much less often.

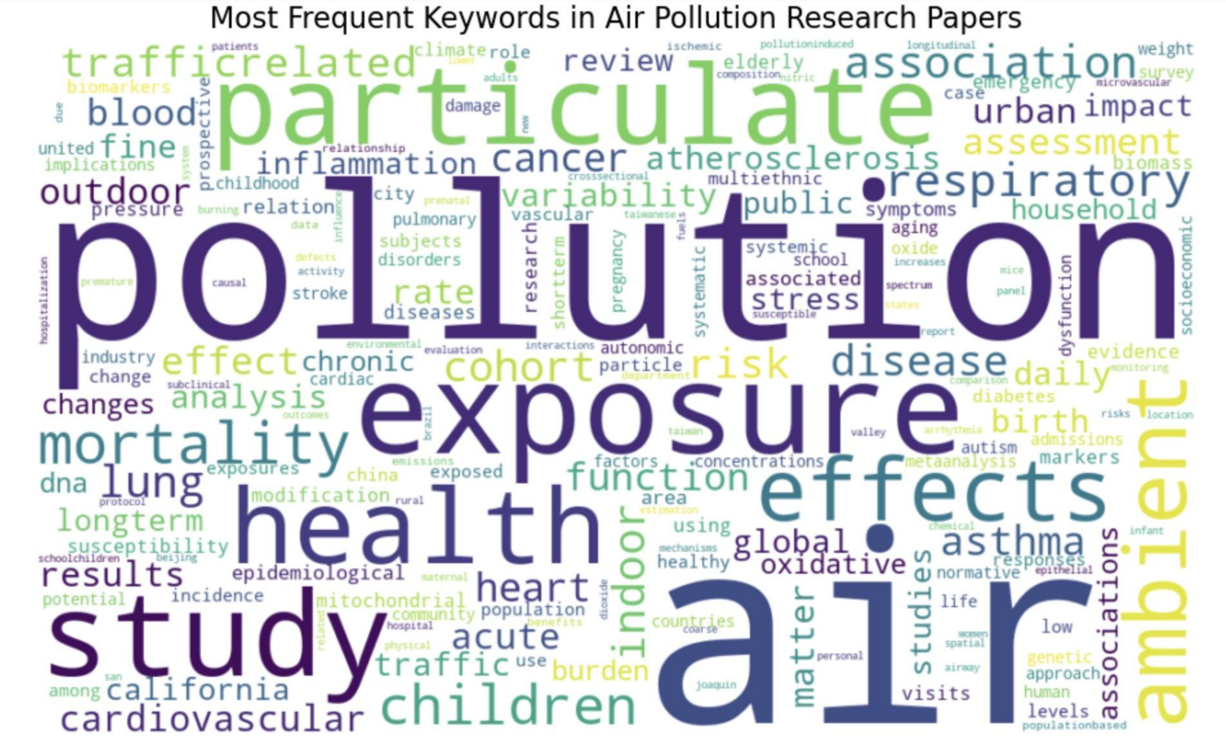
In general, the chart shows that academic articles are the main way people share research about air pollution, while other ways of sharing scientific information are not used as much.

Most Frequent keywords in Air Pollution Research Papers Title

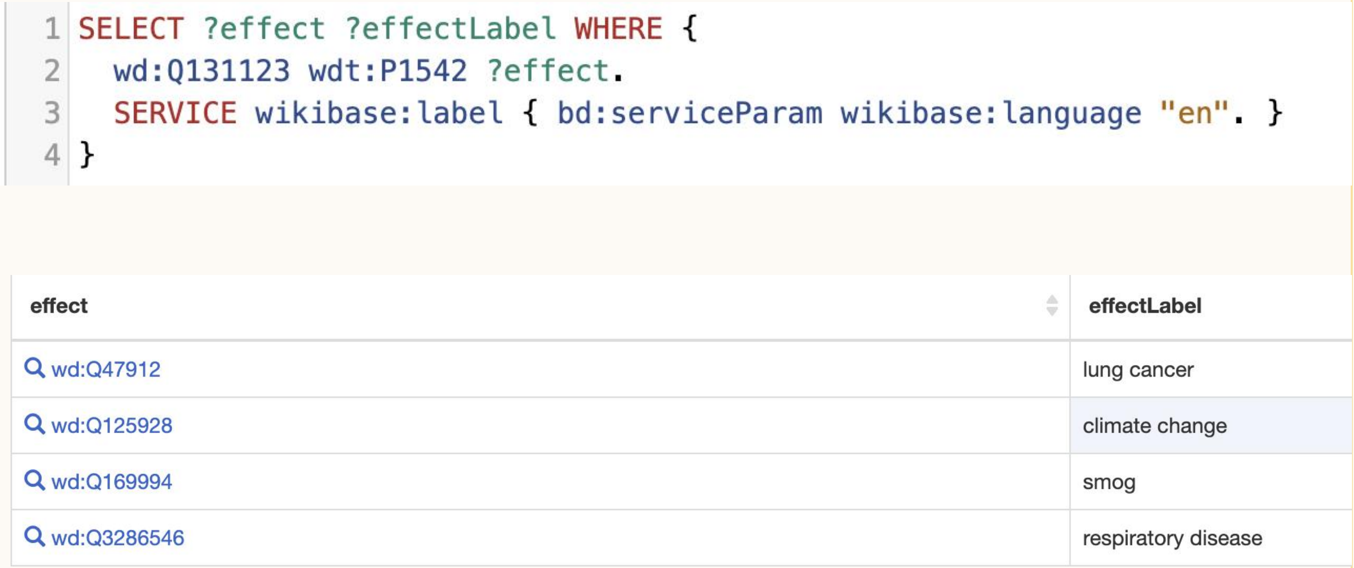


```
SELECT ?paperLabel WHERE {
  ?paper wdt:P921 wd:Q131123.
  ?paper wdt:P31 ?type.
  FILTER(?type IN (
    wd:Q13442814,
    wd:Q13406463,
    wd:Q179461
  ))
  SERVICE wikibase:label { bd:serviceParam wikibase:language "en". }
}
LIMIT 300
```

A keyword frequency analysis was conducted based on the titles of over 300 academic publications related to air pollution. The resulting word cloud highlights the most commonly recurring terms in this field of research.


Most Frequent Keywords in Air Pollution Research Papers

Impact of air pollution



```
SELECT ?effect ?effectLabel WHERE {
  wd:Q131123 wdt:P1542 ?effect.
  SERVICE wikibase:label { bd:serviceParam wikibase:language "en". }
}
```

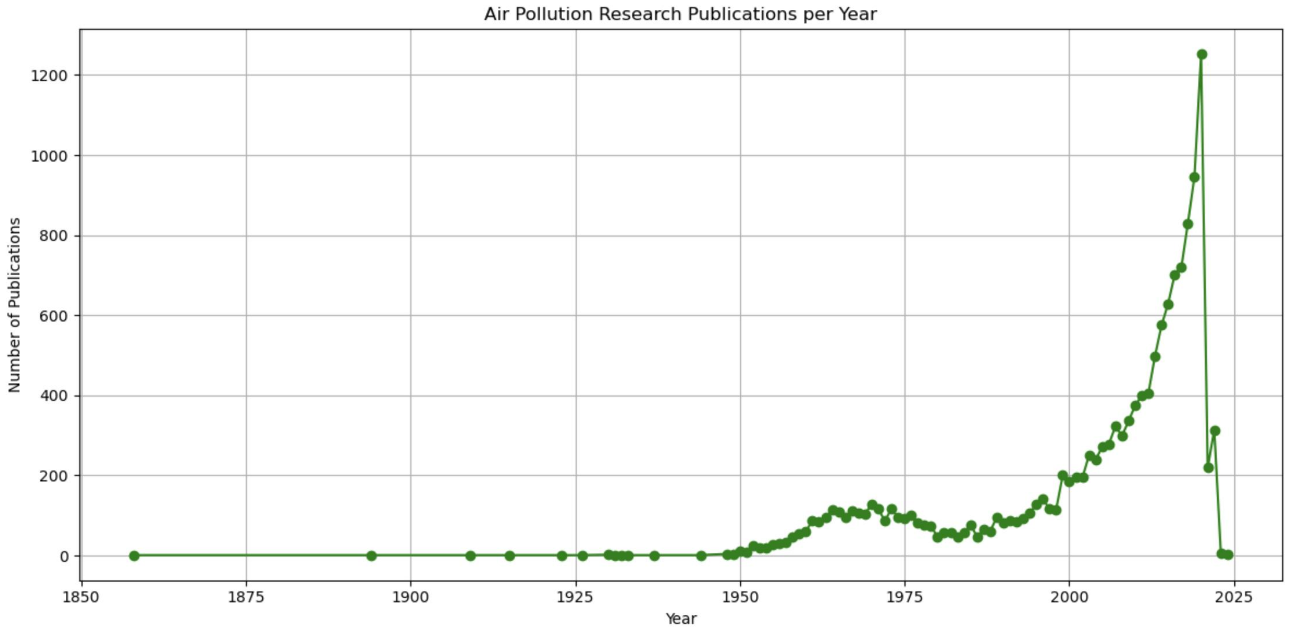
effect	effectLabel
wd:Q47912	lung cancer
wd:Q125928	climate change
wd:Q169994	smog
wd:Q3286546	respiratory disease

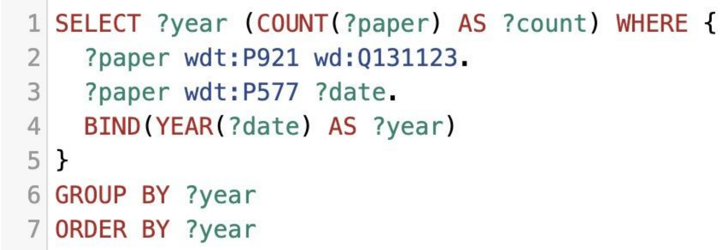
Real-life medical and environmental impacts are the most common consequences of air pollution.

Health impacts: lung cancer, respiratory disease → directly related to human health

Environmental impacts: climate change, smog → reflect the chain reaction of air pollution on natural systems

Air pollution research publications per year





```
SELECT ?year (COUNT(?paper) AS ?count) WHERE {
  ?paper wdt:P921 wd:Q131123.
  ?paper wdt:P577 ?date.
  BIND(YEAR(?date) AS ?year)
}
GROUP BY ?year
ORDER BY ?year
```

We analyzed the publication years of scholarly articles whose main subject is air pollution, as recorded in Wikidata. The data shows a clear upward trend in research output, especially after the year 2000.

Air pollution research has expanded rapidly, this growth reflects the rising global concern around environmental health, the development of international climate policies, and the availability of more environmental data for academic use.

04

Conclusion

Conclusion

Overall, the visualizations highlight how different data points, from human activities to natural events and academic knowledge, intertwine to shape our understanding of pollution. By combining these perspectives, this study shows that linked open data provides a rich, multidimensional view of environmental pollution - linking emission sources, event records, and knowledge representation. It also showcases various aspects of air pollution, allowing people to clearly understand the causes and effects.

THANK YOU