

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

GROUP MEMBER:

(listed in no particular order)

Jiani XU

Leyan CHENG

Yuxin GONG

Catalogue

1. Introduction

2. Chapter 1: Exploratory Analysis of the Database

- a. Origin and Structure of the Data
- b. Data Related to Pollution

3. Chapter 2: Data Analysis: SPARQL Queries

- a. Sources of air pollution
 - Enquire about the number of motor cars in each country
 - Enquire about the number of wildfire in each country
- b. Topics of air pollution
 - Air Pollution Topic Count number of typelabel
- c. Impact of air pollution
- d. Research Paper of air pollution
 - Most Frequent keywords in Air Pollution Research Papers
Title
 - Air pollution research publications per year

4. Conclusion

Introduction

In recent decades, air pollution has become one of the most pressing environmental challenges worldwide. It affects not only the atmosphere, but also human health, ecosystems and the global climate. As cities grow and industrial activities expand, the sources of air pollution - including motor vehicles, industrial emissions, and wildfires - are increasing.

With the rise of Linked Open Data (LOD) and Semantic Web technologies, researchers now have powerful tools to explore, query, and analyse air pollution-related data across different domains and regions. In this study, we use SPARQL to query structured knowledge from wiki data in order to study pollution-related entities, their causes, types and geographical distribution.

By combining semantic querying with Python-based data visualisation, this study aims to provide a clearer picture of the manifestations of air pollution in open datasets and highlight potential patterns and insights to support environmental awareness and data-driven decision-making.

Chapter 1: Exploratory Analysis of the Database

a. Origin and Structure of the Data

The data used in this project entirely came from Wikidata, which is a collaborative, open knowledge base maintained by the Wikimedia Foundation. Wikidata contains structured data on a wide range of topics, including environmental issues, and it supports querying through SPARQL, which is a powerful semantic query language designed for linked data. All the data used in this analysis was retrieved using SPARQL queries through the Wikidata Query Service. Instead of relying on pre-built dataset, we manage to use customized queries to extract relevant information about air pollution.

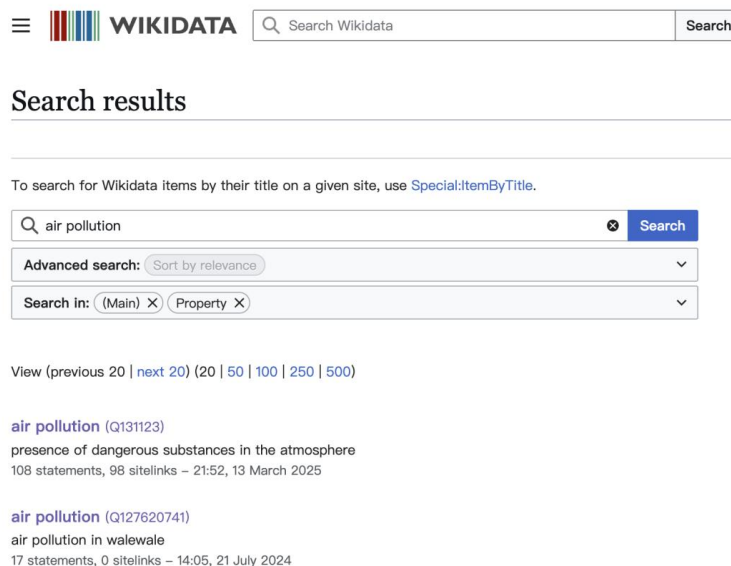
The data structure in Wikidata is based on a triple model: subject–predicate–object. For example, Motor car (Q1420) is a cause (P828) of air pollution (Q131123). Each entity (identified by a Q-number, Q131123 means air pollution) is linked to others through properties (identified by a P-number, P828 for “has cause”).

The results of the queries were returned in tabular format. Each row typically represents a data entity or relationship, and columns typically showing entity labels, types, countries, or counts. This format enabled further analysis and visualization using Python, especially with libraries like pandas, matplotlib, and plotly.

b. Data Related to Pollution

This project focuses on air pollution and the relevant data. All data was collected from Wikidata using SPARQL queries.

We started by searching for “air pollution” in the wikidata search bar. There are 2 same results, but we chose the first one. Because the first one had more information than the second one.



The screenshot shows the Wikidata search interface. At the top, there is a search bar with the text "Search Wikidata" and a "Search" button. Below the search bar, the heading "Search results" is displayed. A note states: "To search for Wikidata items by their title on a given site, use [Special:ItemByTitle](#)." Below this, there is a search input field containing "air pollution" and a "Search" button. Under the search bar, there are two dropdown menus: "Advanced search: Sort by relevance" and "Search in: (Main) X (Property) X". Below the search results, there is a pagination link: "View (previous 20 | next 20) (20 | 50 | 100 | 250 | 500)". Two search results are listed:

- [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 selected different types of information for better understanding of air pollution. The main categories of data include:

- **Causes:** We found several reasons caused air pollution, such as motor cars, wildfires, Pollutants. And we mainly analyzed motor cars and wildfires.
- **Effects:** Air pollution is linked to environmental problems like carbon dioxide emissions and methane. Some records also mention transport, brickworks and traffic congestion.
- **Policies, laws and Scientific studies:** We found several air pollution-related laws and treaties, such as the Clean Air Act in the United States. In addition, many scholarly articles and meta-analyses in Wikidata focus on the health effects of air pollution.

This linked data helped us directly explore and visualize air pollution from **multiple angles: environment, society, medicine, law and so on.**

Chapter 2: Data Analysis: SPARQL Queries

a. Sources of air pollution

```
1 SELECT ?cause ?causeLabel WHERE {  
2   wd:Q131123 wdt:P828 ?cause.  
3   SERVICE wikibase:label { bd:serviceParam wikibase:language "en". }  
4 }
```

- wd:Q131123: Refers to the Wikidata entity for **air pollution**.
- wdt:P828: Represents the property "**has cause**".
- ?cause: A variable that will return all causes linked to air pollution.
- ?causeLabel: The human-readable English label of each cause.
- SERVICE wikibase:label: Automatically fetches the English labels for entities.

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

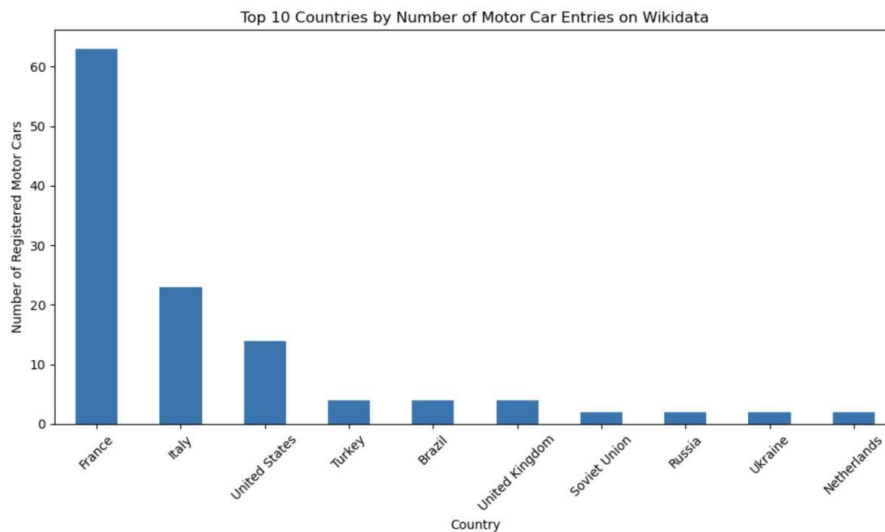
According to the results, we can see that the two main causes of air pollution are motor car (Q1420) and wildfire (Q169950) which correspond to the two types of human activities and natural sources of pollution respectively, so we will continue to focus our analyses according to these two directions as well.

● Enquire about the number of motor cars in each country

```
1 SELECT ?car ?carLabel ?countryLabel WHERE {  
2   ?car wdt:P31 wd:Q1420.  
3   ?car wdt:P17 ?country.  
4   SERVICE wikibase:label { bd:serviceParam wikibase:language "en". }  
5 }
```

- ?car: A variable representing an individual car entity.
- wdt:P31 wd:Q1420: This means the entity is an instance of (P31) a motor car (Q1420).
- wdt:P17: This property represents the country of origin or association.
- ?country: A variable for the country related to the car.
- ?carLabel and ?countryLabel: These return human-readable labels (names) in English.

- SERVICE wikibase:label: This line ensures that English labels are returned for all entities.



This graph shows the ten countries with the highest number of motor vehicles recorded on Wikidata.

While the chart does not directly reflect real-world vehicle ownership or pollution levels, because France may have a more active Wikidata editorial community, or its vehicle manufacturing and history information is widely available (e.g. local brands such as Renault, Citroen, etc.), it provides valuable information for understanding the potential environmental impacts associated with transport activities in different regions. French motor vehicles are well represented in the semantic data sources. This is indicative of both France's long automotive history and its potentially high level of transport-related emissions. Countries such as Italy and the United States also show a significant presence, which coincides with their known industrial and urban development. These countries are likely to face challenges related to air pollution, greenhouse gas emissions and urban smog, especially in densely populated cities with heavy vehicular traffic.

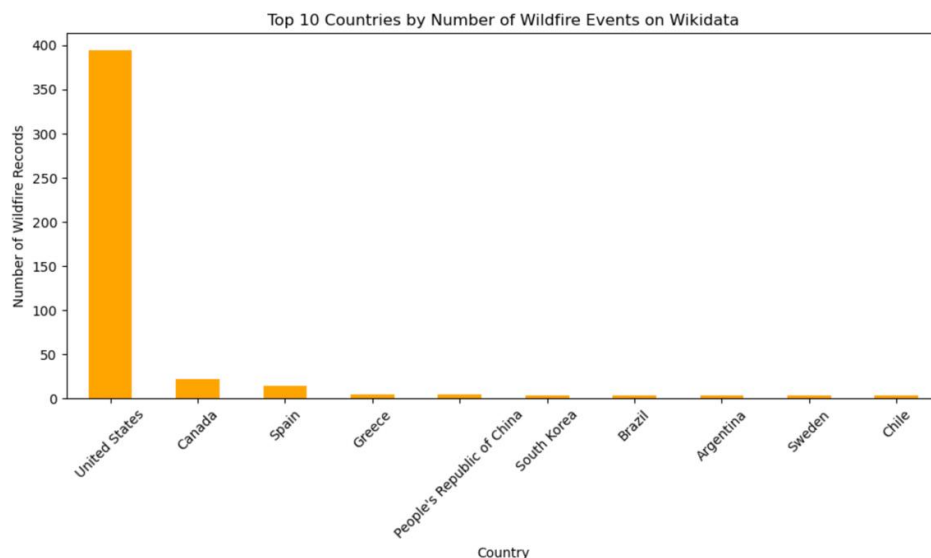
In contrast, countries such as Turkey, Brazil and Russia have relatively few. This may reflect the low representativeness of the data from Wikidata, or it may reflect the fact that motorisation is indeed low, especially in rural or developing areas.

The presence of the Soviet Union in the list highlights the importance of historical data, as legacy industrial activities may still have lasting environmental impacts.

● Enquire about the number of wildfire in each country

```
1 SELECT ?fire ?fireLabel ?countryLabel WHERE {  
2   ?fire wdt:P31 wd:Q169950.  
3   ?fire wdt:P17 ?country.  
4   SERVICE wikibase:label { bd:serviceParam wikibase:language "en". }  
5 }  
6 LIMIT 500
```

- ?fire: A variable representing a wildfire event.
- wdt:P31 wd:Q169950: Filters for entities that are an instance of wildfire (Q169950).
- wdt:P17 ?country: Gets the country where each wildfire occurred (P17 = country).
- SERVICE wikibase:label: Automatically retrieves English labels for the fire and country entities .
- LIMIT 500: Limits the results to 500 entries to avoid overload.



The figure above shows the top 10 countries with the highest number of recorded wildfire events in the Wikidata. Notably, the United States dominates this dataset with over 400 entries, while all other countries have less than 50 entries. 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.

b. Topics of air pollution

● Air Pollution Topic Count number of typelabel

```
1 SELECT ?entity ?entityLabel ?typeLabel WHERE {  
2   ?entity wdt:P921 wd:Q131123.  
3   ?entity wdt:P31 ?type.  
4   SERVICE wikibase:label { bd:serviceParam wikibase:language "en". }  
5 }
```

- ?entity: A variable representing an entity.
- wdt:P921 wd:Q131123: Filters for entities whose main subject (P921) is air pollution (Q131123).
- wdt:P31 ?type: Retrieves the instance type (P31) of each entity.
- ?entityLabel, ?typeLabel: Human-readable English labels for the entity and its type.
- SERVICE wikibase:label: Automatically returns English labels for the entities and their types.

Air Pollution Entity Types (Grouped)



This pie chart presents an overview of the grouping of entity types related to the topic of air pollution. The data shows that academic articles dominate, accounting for 95.7 per cent of the total records. This suggests that air pollution is a more researched topic in academia, with the vast majority of articles being peer-reviewed studies.

Other types, including editorials (1.9%), meta-analyses (0.9%), and miscellaneous (1.6%), appeared much less frequently. The small proportion of meta-analyses is particularly noteworthy as it highlights a potential gap in the field in terms of systematic reviews and pooled studies. Increasing the number of such studies could contribute to evidence-based decision-making and more comprehensive scientific synthesis.

Overall, this data suggests that scholarly articles are the primary medium for

disseminating research on air pollution, while other forms of scientific communication remain underrepresented.

c. Impact of air pollution

```
1 SELECT ?effect ?effectLabel WHERE {
2   wd:Q131123 wdt:P1542 ?effect.
3   SERVICE wikibase:label { bd:serviceParam wikibase:language "en". }
4 }
```

- wd:Q131123 → Refers to the Wikidata entity for air pollution
- wdt:P1542 → The property "has effect", which links a topic to its consequences or impacts
- ?effect → A variable that retrieves all effects associated with air pollution
- ?effectLabel → The human-readable name (label) of each effect
- SERVICE wikibase:label → Automatically displays results in English instead of QIDs

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

The impacts of air pollution are both medical and environmental:

Health impacts:

- Lung cancer, respiratory disease
- Directly affect human well-being and public health

Environmental impacts:

- Climate change, smog
- Show how air pollution triggers chain reactions in the natural environment

d. Research Paper of air pollution

● Most Frequent keywords in Air Pollution Research Papers Title

```
1 SELECT ?paperLabel WHERE {
2   ?paper wdt:P921 wd:Q131123.
3   ?paper wdt:P31 ?type.
4   FILTER(?type IN (
5     wd:Q13442814,
6     wd:Q13406463,
7     wd:Q179461
8   ))
9   SERVICE wikibase:label { bd:serviceParam wikibase:language "en". }
10 }
11 LIMIT 300
```

- ?paper wdt:P921 wd:Q131123: Selects papers whose main subject is air pollution.
- wd:Q131123: the Wikidata entity for air pollution.

- [illegible]

Most common keywords include:

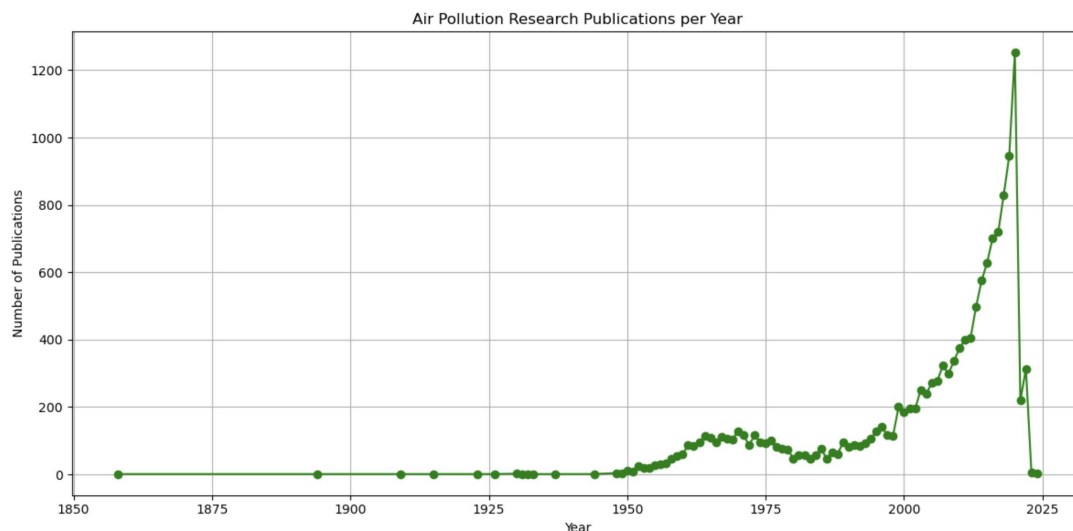
These keywords reflect the main themes in air pollution research:

- **Air pollution research publications per year**

- `wd:Q131123` → Refers to the Wikidata entity for air pollution
- `wdt:P1542` → The property "has effect", which links a topic to its consequences or

impacts

- `?effect` → A variable that retrieves all effects associated with air pollution
- `?effectLabel` → The human-readable name (label) of each effect
- `SERVICE wikibase:label` → Automatically displays results in English instead of QIDs



To understand how academic interest in air pollution has evolved over time, 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.

While isolated publications can be traced back as early as the 19th century (e.g., a single article in 1858), the volume remained relatively low until the 1950s. Starting from the 1960s and 70s — possibly due to increasing urbanization and the rise of environmental awareness — the number of publications began to grow steadily. Notably, from the year 2000 onward, air pollution research has expanded rapidly. The number of publications increased from around 200 in 1999 to over 1,200 in 2020, indicating a significant surge in global scientific attention.

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. The slight drop in 2021–2023 may be due to delayed data updates or reduced publication indexing in Wikidata, and does not necessarily reflect an actual decline in research activity.

Conclusion

Through the use of Linked Open Data and SPARQL queries, this project provides a multidimensional perspective on air pollution. By retrieving structured data from Wikidata, we explored its primary causes—such as motor vehicles and wildfires—as well as its wide-ranging effects on health and the environment. The analysis shows that countries like France and the United States have a high presence of motor vehicle data, suggesting potential hotspots of traffic-related pollution, while the United States also dominates in wildfire records, reflecting the impact of climate-induced natural disasters.

In addition, the project analyzed academic publications related to air pollution, revealing that over 95% of the entries are scholarly articles. This reflects the strong academic attention the topic receives but also points to a lack of systematic reviews and public-facing summaries, which are essential for translating research into action.

Overall, this study demonstrates how open, linked datasets and semantic technologies can support deeper insights into complex environmental issues. It highlights the importance of data-driven approaches in promoting environmental awareness, guiding policy decisions, and encouraging broader participation. We advocate for more open data sharing, interdisciplinary collaboration, and public engagement to collectively tackle the global challenge of air pollution.