‹ Responder 29

24 - Anonymous

02:37

Time to complete ›

1. Priority *

6

2. URL *

https://epa.gov/outdoor-air-quality-data/download-daily-data

3. Agency *

N/A

4. Event

5. Title

Daily Outdoor Air Quality Data- Download Daily Data

6. Crawled by the Internet Archive *

● Yes

7. Internet Archive URL

https://web.archive.org/web/*/https://www.epa.gov/outdoor-air-quality-data/download-daily-data

8. Description

9. Purpose or significance of data

Contains pollutant concentrations based on date and location. Historic data 1980-2016 (2017 does not return data) IMPORTANT: Way Back Machine does not work when you try to use the interactive data selector, therefore data download is not available.

10.

☐ Do not harvest. All data is small, unstructured, and on a page crawlable by the Internet Archive.

☑ Page contains dynamic content (e.g., links loaded by JavaScript).

☐ Page contains interactive visualizations.

☑ Data is accessible in structured file(s) that can be directly downloaded.

☐ Data is accessible over FTP.

☐ Data is accessible using a documented public API.

☑ Data is only accessible using search queries in a web form.

11. Recommended approach to harvesting data

A CSV file can be generated for each year for all sites in a single state. It contains ambient air pollution data collected by EPA, state, local, and tribal air pollution control agencies.

12. File formats

CSV

13. Estimated size in MB

1000

14. Related URLs

9857A837-BE41-4319-8C0D-F0D657F81A90 https://www.epa.gov/outdoor-air-quality-data 0708D63F-197F-4815-8C2E-B409C4872AB7 https://aqsdr1.epa.gov/aqsweb/aqstmp/airdata/download_files.html

15. Were you able to capture all of the data at this URL?

● Yes

○ No

16. Harvest method used

Python script used to query for and capture each csv file

17. Notes from Harvest

We grabbed all possible combinations of <pollutant>_<year>_<state>.csv. Also included is a file data/nodata.txt which lists all of the combinations which had no data available. When accessing the form https://www.epa.gov/outdoor-air-quality-data/download-daily-data directly, it should not be possible to choose any of the combinations listed in data/nodata.txt. Note that if the scrape.py script is run again at a later date, the following lines should be removed: # Remove 2017 from year_options del year_options[0] And you may also wish to delete some lines from data/nodata.txt, since downloads will only be attempted if the csv file does not exist and the file name is not present in data/nodata.txt -BinaryMan32

18. User certified that to the best of their knowledge this is a well-checked bag that will survive out of context of the site it was harvested from.

● Yes

19. Notes from Bagging

20. Notes from Describe