Responder 29

25 - Anonymous

02:15
Time to complete

1. Priority *

9

2. URL *

https://www3.epa.gov/enviro/facts/tri/p2.html

3. Agency *

Environmental Protection Agency

4. Event

5. Title

Pollution Prevention Search | Envirofacts | US EPA

6. Crawled by the Internet Archive *

- (●) Yes

7. Internet Archive URL

https://web.archive.org/web/*/https://www3.epa.gov/enviro/facts/tri/p2.html

8. Description

9. Purpose or significance of data

Report on pollution output from facilities across the United States. Searchable by Industry Sector, Chemical Group, State/Zip code, or year. Tables contain: Facility name, chemical, year, current year output, previous year output, percent difference and summary.

10.

- [ ] Do not harvest. All data is small, unstructured, and on a page crawlable by the Internet Archive.
- [x] Page contains dynamic content (e.g., links loaded by JavaScript).
- [ ] Page contains interactive visualizations.
- [x] 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.
- [x] Data is only accessible using search queries in a web form.

11. Recommended approach to harvesting data

other data available: https://www.epa.gov/enviro/data-downloads 1. Search by all chemicals, industries and states. Specify year. 2. Download data as either a PDF/CSV/XLSX file. 3. Repeat for all other years.

12. File formats

PDF, CSV, XLSX

13. Estimated size in MB

20

14. Related URLs

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

- (●) Yes
- ( ) No

16. Harvest method used

R script to read web pages for year, scrape table data

17. Notes from Harvest

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

checked and uploaded to s3 - John V

20. Notes from Describe