‹ Responder 29

13 - Anonymous

04:45

Time to complete ›

1. Priority *

1

2. URL *

https://iaspub.epa.gov/triexplorer/tri_release.chemical

3. Agency *

Environmental Protection Agency

4. Event

5. Title

Release Chemical Report | TRI Explorer | US EPA

6. Crawled by the Internet Archive *

● Yes

7. Internet Archive URL

Internet Archive URL: https://web.archive.org/web/*/https://iaspub.epa.gov/triexplorer/tri_release.chemical

8. Description

9. Purpose or significance of data

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

Autogenerate UI form interaction (maybe using nightwatch), then automate grabbing the data from subsequent report pages.

12. File formats

13. Estimated size in MB

14. Related URLs

https://iaspub.epa.gov/triexplorer/tri_release.chemical

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

● Yes

○ No

16. Harvest method used

Created a Perl script (tri_get.pl) that takes a list of URLs and chemical names from separate 'year' searches. This script takes for arguments the filename of URLs and the type of document (html, csv). There is a Chrome console.js script for extracting the URLs from chrome which you will paste into the file of URLs (which I named according to year.)

The Perl script will rename the file to a normalized chemical name with year (the http query string was too long for filenames.)

17. Notes from Harvest

Perl's LWP::Simple was not working properly, neither was wget. Curl works fine within the script and has successfully finished.

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

Data is a .CSV and a .HTML file containing the report for each chemical for each year. Zip file is 214 KB. Unzipped is 1.66 GB

20. Notes from Describe

https://iaspub.epa.gov/triexplorer/tri_release.chemical Description taken from https://www.epa.gov/toxics-release-inventory-tri-program/learn-about-toxics-release-inventory#What is the Toxics Release Inventory?