‹ Responder 29

28 - Anonymous

02:26
Time to complete ›

1. Priority *

5

2. URL *

https://epa.gov/iris

3. Agency *

N/A

4. Event

5. Title

Integrated Risk Information System

6. Crawled by the Internet Archive *

● Yes

7. Internet Archive URL

https://web.archive.org/web/*/https://www.epa.gov/iris

8. Description

9. Purpose or significance of data

US EPA's centralized database of chemical assessments, which identify and characterize the health hazards of chemicals found in the environment. Many years of work goes into each assessment. They are provided here as (1) a searchable database of HTML pages containing summary information about each chemical, presented in a uniform way; (2) full toxicological reports (up to ~100s of pages) and summaries (~10s of pages) which are downloadable PDF files linked from each assessment page; (3) drafts and other public documents related to the assessment process.

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

The IRIS website contains (1) a search interface for chemical assessments; (2) an organized set of pages and links to PDF files containing the substantive content of these assessments; (3) additional technical materials related to the development of the assessments; and (4) general description of the program and links to other EPA sites. Recommended to harvest items (2) and (3) by crawling HTML pages, including some useful tables of assessed chemicals, and download all linked files (including copies of the HTML pages). The content falling into category (4) above is almost certainly crawled by Internet Archive.

12. File formats

PDF, HTML

13. Estimated size in MB

480

14. Related URLs

FF8B3DF6-A117-474A-9FB8-F863BF34C9B2 https://cfpub.epa.gov/ncea/iris/search/basic/index.cfm D48932BE-189A-4704-BA83-BFAAD1F39A81 https://cfpub.epa.gov/ncea/iris/search/index.cfm

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

● Yes

○ No

16. Harvest method used

custom ruby scraper / v2: python scraper

17. Notes from Harvest

Harvested twice independently, oops. Uploaded additional zip file 14BDABB8-F78D-41BC-A160-C6E631E39081.v2.zip Original harvester notes: toxicological report summaries could not be retrieved for the following files: {"id":"Cyanogen_142","href":"https://cfpub.epa.gov/ncea/iris2/chemicalLanding.cfm?substance_nmbr=32","name":"Cyanogen"} {"id":"Hydrazine/Hydrazine_sulfate_282","href":"https://cfpub.epa.gov/ncea/iris2/chemicalLanding.cfm?substance_nmbr=352","name":"Hydrazine/Hydrazine_sulfate"} {"id":"Potassium_cyanide_392","href":"https://cfpub.epa.gov/ncea/iris2/chemicalLanding.cfm?substance_nmbr=92","name":"Potassium_cyanide"} {"id":"Potassium_silver_cyanide_393","href":"https://cfpub.epa.gov/ncea/iris2/chemicalLanding.cfm?substance_nmbr=93","name":"Potassium_silver_cyanide"} {"id":"Pyrene_410","href":"https://cfpub.epa.gov/ncea/iris2/chemicalLanding.cfm?substance_nmbr=445","name":"Pyrene"} {"id":"Sodium_cyanide_429","href":"https://cfpub.epa.gov/ncea/iris2/chemicalLanding.cfm?substance_nmbr=101","name":"Sodium_cyanide"} Second harvester notes (akokai): Captured all files outlined in the description.

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

changed harvest URL to v2 per the harvest notes. 3 files were causing errors during the validation process. 2017-03-04 13:44:30,536 - WARNING - data/data/IRIS_Assessments_Final/1,3,5-Trimethylbenzene/supporting-documents-trimethylbenzenes exists in manifest but not found on filesystem 2017-03-04 13:44:30,537 - WARNING - data/data/IRIS_Assessments_Final/1,2,3-

Trimethylbenzene/supporting-documents-trimethylbenzenes exists in manifest but not found on filesystem 2017-03-04 13:44:30,537 - WARNING - data/data/IRIS_Assessments_Final/1,2,4-Trimethylbenzene/supporting-documents-trimethylbenzenes exists in manifest but not found on filesystem 2017-03-04 13:44:30,537 - WARNING - data/data/IRIS_Assessments_Final/1,2,3-Trimethylbenzene/supporting-documents-trimethylbenzenes exists on filesystem but is not in manifest 2017-03-04 13:44:30,537 - WARNING - data/data/IRIS_Assessments_Final/1,3,5-Trimethylbenzene/supporting-documents-trimethylbenzenes exists on filesystem but is not in manifest 2017-03-04 13:44:30,537 - WARNING - data/data/IRIS_Assessments_Final/1,2,4-Trimethylbenzene/supporting-documents-trimethylbenzenes exists on filesystem but is not in manifest Removing the spaces at the ends of their filenames solved the issue and successfully validated -John V

20. Notes from Describe