‹ Responder 29

27 - Anonymous ⌄

07:11

Time to complete ›

1. Priority *

1

2. URL *

https://ofmpub.epa.gov/sor_internet/registry2/reusereg/searchandretrieve/

3. Agency *

Environmental Protection Agency

4. Event

5. Title

System of Registries | US EPA

6. Crawled by the Internet Archive *

● Yes

7. Internet Archive URL

https://web.archive.org/web/*/https://ofmpub.epa.gov/sor_internet/registry2/reusereg/searchandretrieve/

8. Description

9. Purpose or significance of data

This is a catalog of data descriptions. It does not seem to contain any actual data, and although many of the data sources listed have a tab that say "Dataset", nothing loads. It is unclear whether these are datasets that exist elsewhere or which have been deleted already. Per Matt Price we should archive the entirety of the catalog, since it contains useful descriptions of the datasets (which may or may not exist).

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

Python web crawling using Beautiful Soup and selenuim module

12. File formats

13. Estimated size in MB

14. Related URLs

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

● Yes

○ No

16. Harvest method used

Python web crawling using Beautiful Soup and selenuim module

17. Notes from Harvest

There is an data map (data_map.csv) to glue all files together. The detailed information about each item is the list are in /data/detail/ page and the file name is resource id.html . Please refer to data_map.csv file to get more information

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

Randomly checked the html files. Reviewed website. Bag validated. This seems to have descriptions only not actual data which is what the researcher noted.

20. Notes from Describe