‹ Responder 29

23 - Anonymous

02:47

Time to complete ›

1. Priority *

7

2. URL *

https://www3.epa.gov/storet/legacy/gateway.htm

3. Agency *

N/A

4. Event

5. Title

EPA - STORET Legacy Data Center

6. Crawled by the Internet Archive *

● Yes

7. Internet Archive URL

https://web.archive.org/web/*/https://www3.epa.gov/storet/legacy/gateway.htm

8. Description

9. Purpose or significance of data

Welcome to the STORET Legacy Data Center, site of the world's largest repository of ambient Water Quality Data. From this site you will be able to access a database that holds over 200 million water sample observations from about 700,000 sampling sites for both surface and ground water. This web site allows both scientists and the general public to access the historical data from the legacy STORET system. First-time users should narrow their search based on the options from the Query page, while experienced users may jump to the no-frills Advanced Query form for requesting data. Legacy STORET contains data of undocumented quality. Further, the data in this system is static, and all new data are being entered into Modernized STORET. Background information about the Office of Water and the history of STORET may be found by following the Purpose link. For more information on the layout of this site, please follow the Site Map link.

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 data sets for this legacy info is available here ftp://ftp.epa.gov/storet/exports . Each states data is supplied as an .exe file. This can be unzipped with Winzip on a PC. Once unziped, the files contain TXT files for the data, divided by county within each state. Be sure to also capture the documentation that helps decode the test data ftp://ftp.epa.gov/storet/exports/docs.

12. File formats

EXE, TXT

13. Estimated size in MB

250-500MB

14. Related URLs

https://www3.epa.gov/storet/legacy/gateway.htm

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

● Yes

○ No

16. Harvest method used

FTP crawling from https://github.com/edgi-govdata-archiving/harvesting-tools ; see "tools/run_me.sh"

17. Notes from Harvest

Data ultimately comes from <ftp://ftp.epa.gov/storet/exports/>. Everything under "exports", including "docs" is captured".

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

This might have been uploaded twice All data appear to be present in the uploaded file listed above

20. Notes from Describe

This dataset already existed in the CKAN, but without any linked data or tags, so I updated the existing record with this info/data.