Disk-Safe Harvesting of Data.gov Distributions: What a Catalog Search Actually Downloads

Jermaine Harris
Hutts Enterprises / Ferrus Labs
2026-09-20 · CC BY 4.0
Keywords: cs.DL, Data.gov, DCAT-US, open government data, catalog API
Source git 38fa640 · technical report

Abstract

Data.gov is a metadata catalog. This report measures what a disk-capped harvester obtained from Catalog API v4 searches on 2026-09-20, plus two curated agency CSVs fetched by known IRMA and ECOS URLs. Across harvest manifests, 231 distributions committed successfully (about 452 MB of ok bytes), 230 listings had no downloadURL, 85 were skipped by policy (HTML, PDF, FTP, size), and 6 failed with transport errors. The on-disk corpus is 556 MB. The result is an empirical description of catalog yield, not a claim that Data.gov hosts federal microdata.

1. Introduction

Research question. If an operator searches Data.gov by topic and follows only downloadURL values under explicit disk caps, what fraction of listings yield files, and what kinds of files appear?

Contributions:

- C1. Manifest status counts from this campaign (results/gov/harvest_inventory.json).
- C2. Shape facts for EIA RECS 2009/2015 public-use files that did download.
- C3. Operational skip rules (HTML, PDF, FTP shapefiles, 80 MiB/file, keep 25 GiB free).

Companion reports 06-08 analyze NPS visitation, FWS ECOS species, and National Register extracts from the same corpus.

2. Background

Catalog API v4 (https://api.gsa.gov/technology/datagov/v4/) returns DCAT-US records. Distributions may list downloadURL or only accessURL. The harvester used DATAGOV_API_KEY from a local env file and wrote under Documents/Research Data/datagov. Default caps: 80 MiB per file, 0.2-0.4 GiB per topic run, abort if free disk would fall below 25 GiB. The operator disk was about 50 GiB free (95 percent full) at harvest time.

3. Method

We parsed every runs/*/manifest.jsonl line written by datagov_fetch.py on 2026-09-20. Status values were grouped without re-downloading. File counts are find over files/. Byte totals are stat on the tree. RECS shapes are header-plus-row counts on the two EIA CSVs that landed. No Catalog API calls were repeated for this paper. Hardware: operator workstation, 12th Gen Intel Core i5-12600KF, 64 GiB RAM, Linux 6.17.

4. Results

Table 1. Manifest outcomes (source: results/gov/harvest_inventory.json).

status	n
ok	231
no downloadURL	230
skipped (policy)	85
already on disk	30
transport error	6
no distribution	2

Table 2. Corpus size.

measure	value
successful download bytes (ok)	452409165
files under files/	231
entire datagov tree bytes	555589278

Table 3. EIA Residential Energy Consumption Survey public-use files that downloaded.

file	rows	columns	bytes
recs2009_public.csv	12084	940	28917457
recs2015_public_v4.csv	5686	759	19372309

About half of catalog hits in this run were listings without a direct file URL. Policy skips included Census TIGER FTP zips and NOAA HTML graph pages. USDA catchment weather CSVs later totaled 24 files and 1,614,992 data rows (results/gov/usda_weather_meta.json); those files are a single study, not a national climate archive.

5. Discussion and threats

Yield is query-dependent. Searches for unemployment produced Iowa Socrata CSVs; weather with org_slug=noaa produced HTML. A different query set would change Table 1. ok_bytes counts this campaign's successful transfers, not unique bytes after later copies into historyrise/ and animalstart/. Manifest exists means a later search hit the same destination path. We did not validate CSV schemas beyond RECS column counts. 2025 NPS rows in a sibling file may be a partial year; that threat is discussed in paper 06.

6. Related work

Data.gov documents the catalog as metadata (resources.data.gov/catalog-api). DCAT-US 3.0 distinguishes downloadURL from accessURL. CKAN package_search is legacy and was not used. This paper does not claim GSA harvest-pipeline internals from datagov-harvester source.

7. Ethics

API keys were not written to git. Disk caps avoided filling a 95 percent-full root filesystem. No production WordPress or Ferrum data were modified.

8. Conclusion

A disk-safe Data.gov client mostly retrieves whatever agencies already published as HTTP

files. In this campaign, listings without downloadURL were about as common as successful downloads. The useful tables for later papers (NPS, FWS, NY/MD registers, RECS) came from a mix of catalog hits and known agency URLs.

9. Artifact appendix

Rerun: python3 harness/govdata/analyze.py. Inputs: the corpus path in that script. Outputs: results/gov/harvest_inventory.json and RECS *_shape.json. Harvester: Programming/datagov-research/datagov_fetch.py.

10. References

- Data.gov Catalog API. https://resources.data.gov/catalog-api/
- DCAT-US 3.0 distribution. https://resources.data.gov/standards/catalog/dcat-us-3/distribution/
- EIA RECS public-use files (2009, 2015 v4), downloaded 2026-09-20.