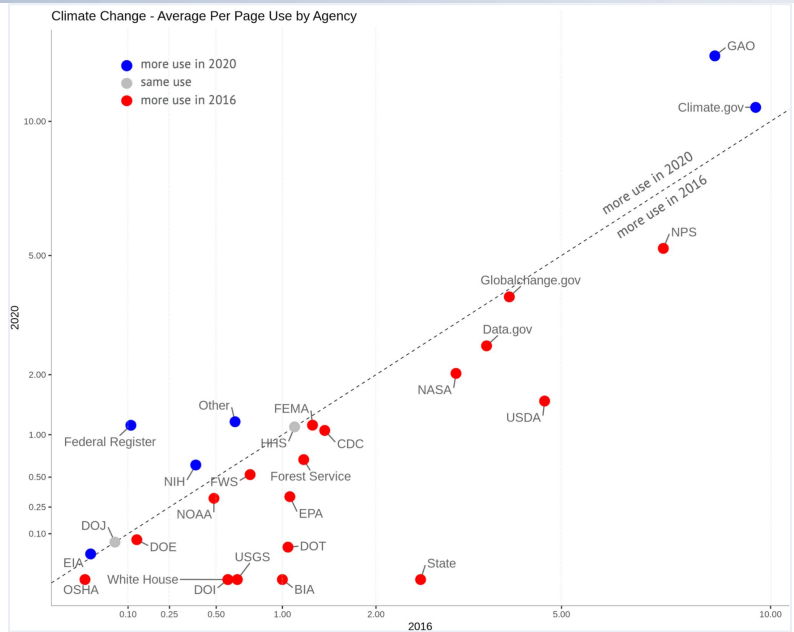
Extending “Visualizing changes to US federal environmental agency websites, 2016–2020”

Original study by Eric Nost , Gretchen Gehrke, Grace Poudrier, Aaron Lemelin, Marcy Beck, and Sara Wylie, on behalf of the Environmental Data & Governance Initiative, PLOS One, 2021

Extension forensics study and presentation by:
Lesley Frew
November 28, 2022

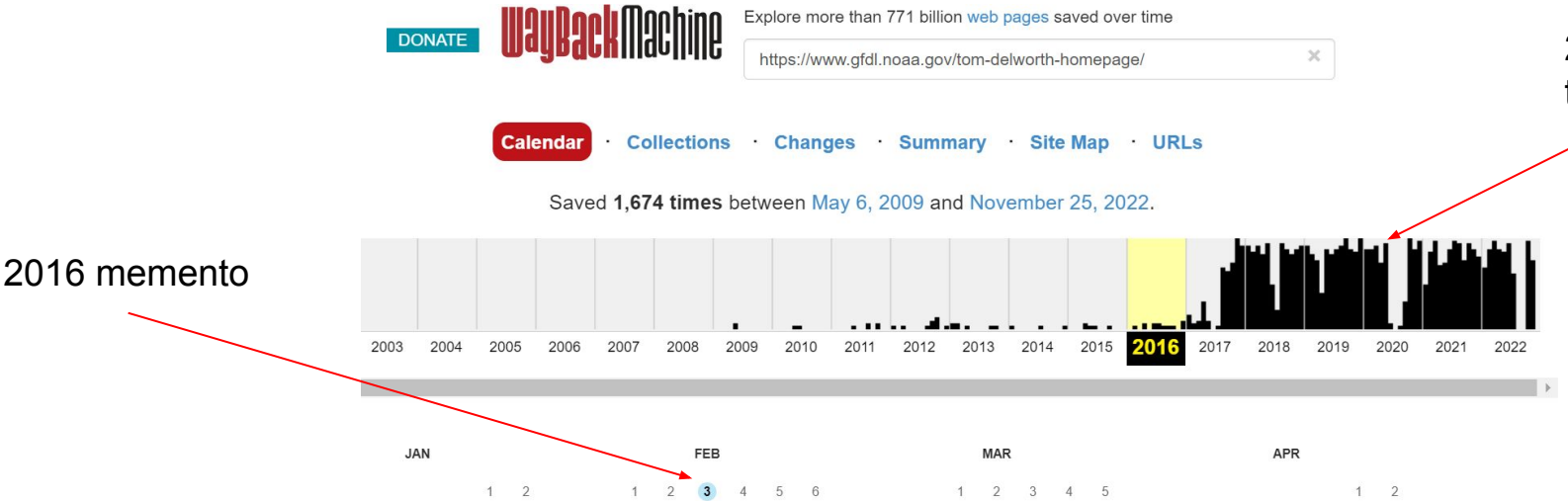
“Climate change” terms were used on federal web pages more in 2016 than in 2020



Nost et al., Visualizing changes to US federal environmental agency websites,, Figure 5

Can we learn more about the deletions by examining multiple web archives to get more paired mementos?

40,378 URI-Rs → 9,144 paired mementos

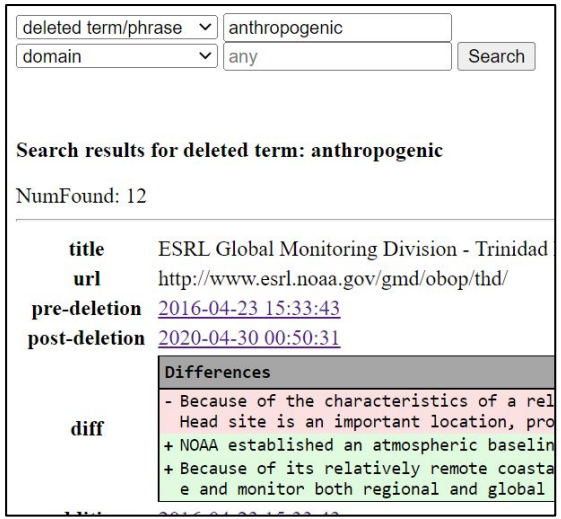


2020 memento to make a pair

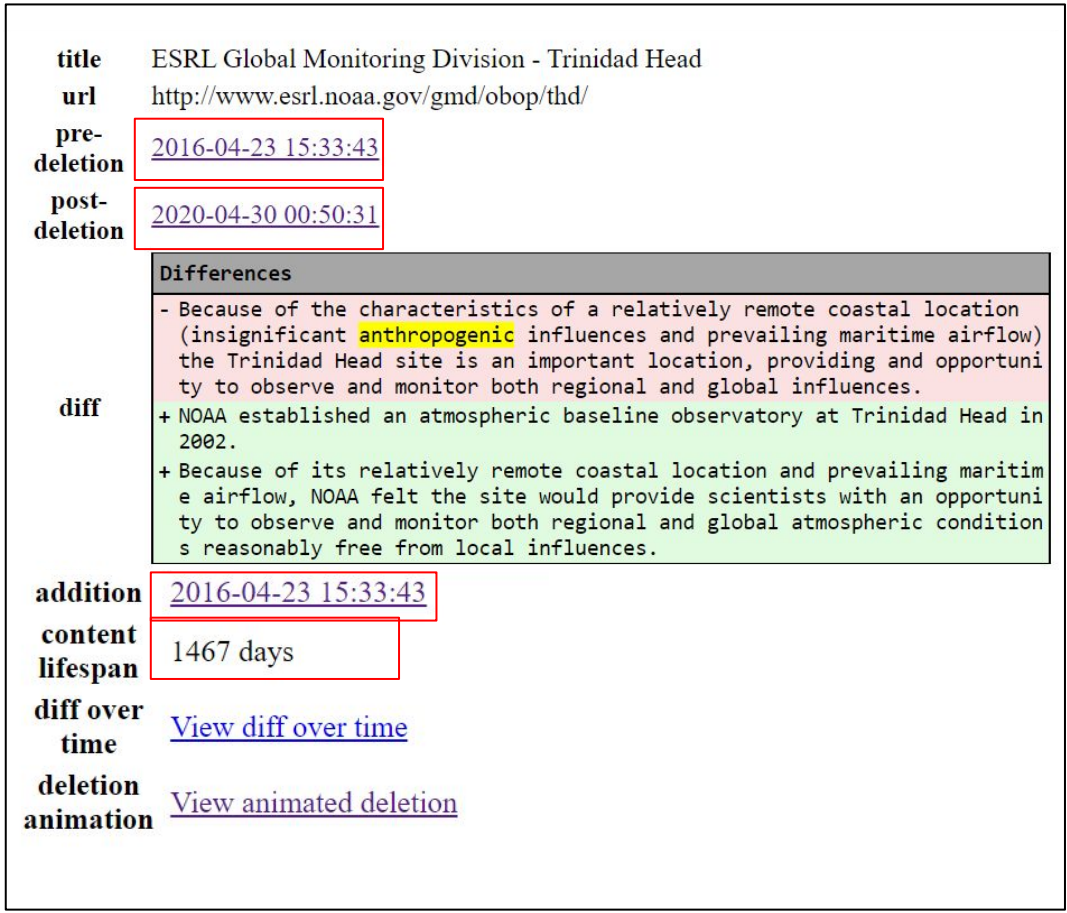
2016 memento

https://web.archive.org/web/20160615000000*/https://www.gfdl.noaa.gov/tom-delworth-homepage/

How can we increase the granularity of the deleted terms and phrases?



zoom in

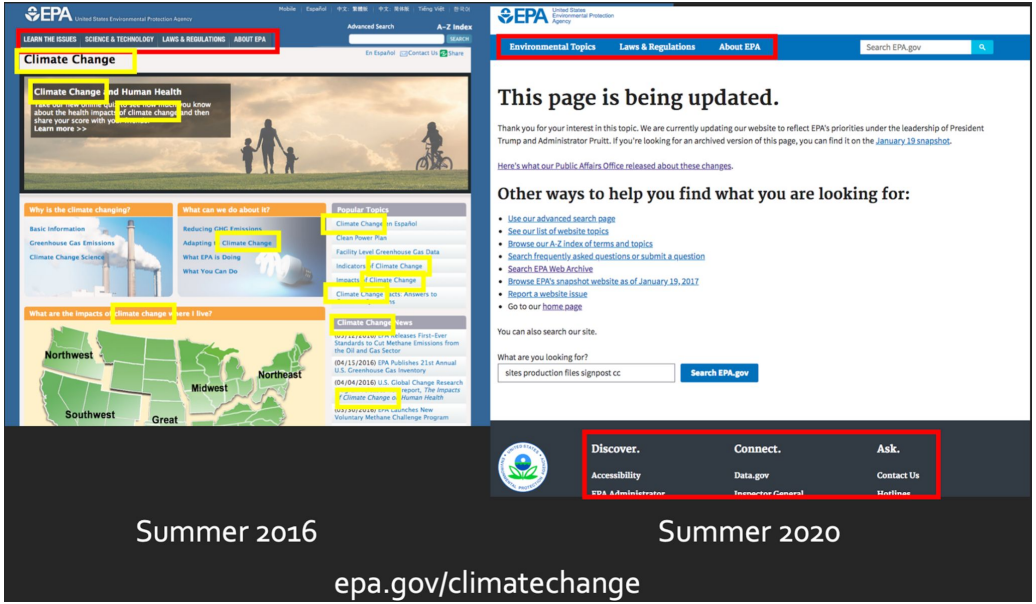


Frew, Nelson, and Weigle. *Work in Progress.*

How can we distinguish between pages where terms were deleted versus entire pages that were deleted?

Both edits and full page deletions have value!

But only pages that weren't deleted entirely will index well.

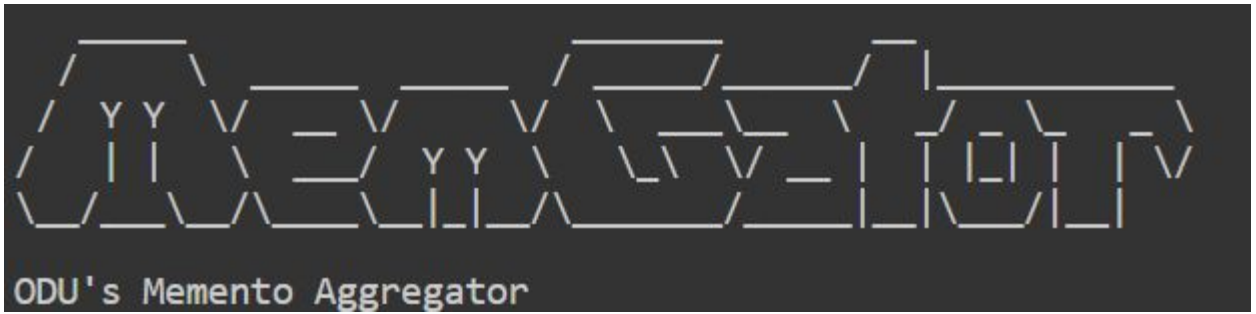


301 status code for this full page deletion

Nost et al., Visualizing changes to US federal environmental agency websites, Figure 1

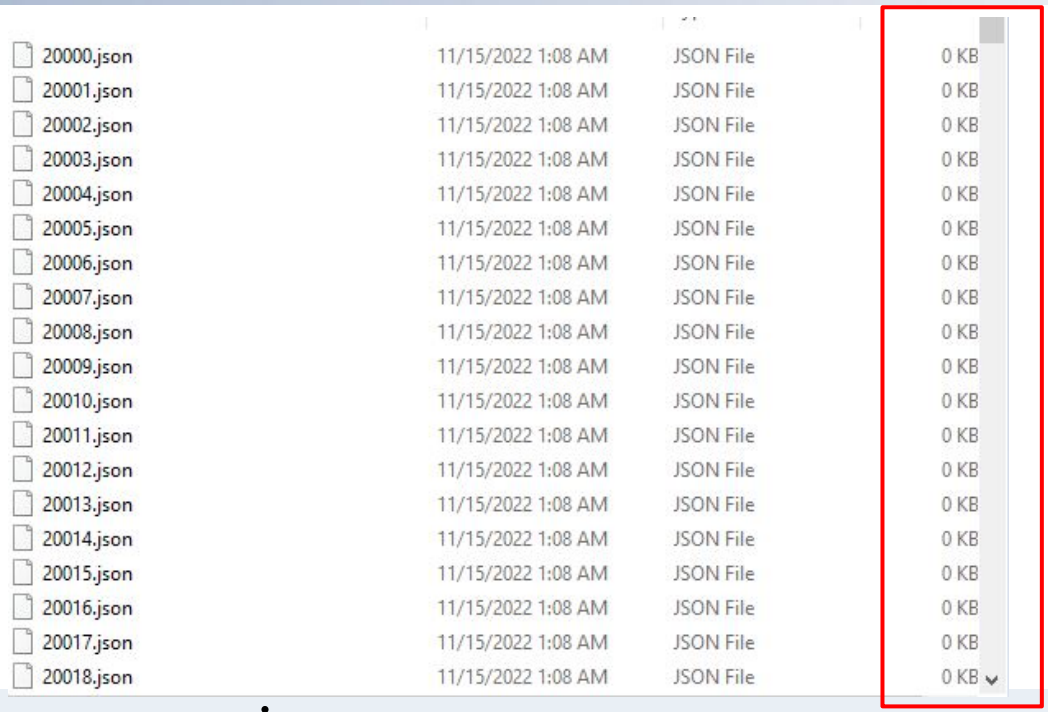
Investigation 1: Memgator can be used to search multiple web archives for additional paired mementos

- Downloaded the 30,000 time maps for the URI-Rs **without** paired mementos from the original data set
- Parsed the time maps to determine if paired mementos exist



https://github.com/oduwsdl/MemGator

Downloading 30,000 time maps requires a good internet connection



20000.json	11/15/2022 1:08 AM	JSON File	0 KB
20001.json	11/15/2022 1:08 AM	JSON File	0 KB
20002.json	11/15/2022 1:08 AM	JSON File	0 KB
20003.json	11/15/2022 1:08 AM	JSON File	0 KB
20004.json	11/15/2022 1:08 AM	JSON File	0 KB
20005.json	11/15/2022 1:08 AM	JSON File	0 KB
20006.json	11/15/2022 1:08 AM	JSON File	0 KB
20007.json	11/15/2022 1:08 AM	JSON File	0 KB
20008.json	11/15/2022 1:08 AM	JSON File	0 KB
20009.json	11/15/2022 1:08 AM	JSON File	0 KB
20010.json	11/15/2022 1:08 AM	JSON File	0 KB
20011.json	11/15/2022 1:08 AM	JSON File	0 KB
20012.json	11/15/2022 1:08 AM	JSON File	0 KB
20013.json	11/15/2022 1:08 AM	JSON File	0 KB
20014.json	11/15/2022 1:08 AM	JSON File	0 KB
20015.json	11/15/2022 1:08 AM	JSON File	0 KB
20016.json	11/15/2022 1:08 AM	JSON File	0 KB
20017.json	11/15/2022 1:08 AM	JSON File	0 KB
20018.json	11/15/2022 1:08 AM	JSON File	0 KB

How can you tell the difference between a timemap with no results, and a connection problem?

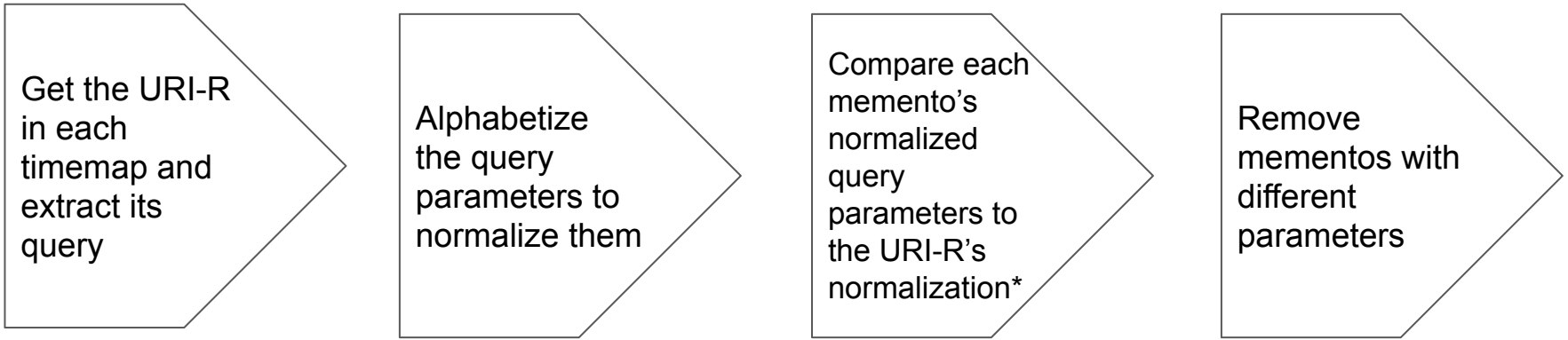
```
$ curl -I https://memgator.cs.odu.edu/timemap/json/http://energy.gov/atmospheric/
HTTP/1.1 404 Not Found
Server: nginx/1.18.0 (Ubuntu)
Date: Sat, 26 Nov 2022 15:58:21 GMT
Content-Type: text/plain; charset=utf-8
Content-Length: 19
Connection: keep-alive
Access-Control-Allow-Origin: *
Access-Control-Expose-Headers: Link, Location, X-Memento-Count, Server
X-Content-Type-Options: nosniff
X-Memento-Count: 0
```

Some web archives (those using PyWB) return extraneous mementos for URI-Rs containing queries


```
{
  "original_uri": "https://www.osha.gov/pls/oshaweb/owadisp.show_document?p_table=NEWS_RELEASES&p_id=33504",
  "self": "http://localhost:1208/timemap/json/https://www.osha.gov/pls/oshaweb/owadisp.show_document?p_table=NEWS_RELEASES&p_id=33504",
  "mementos": {
    "list": [
      {
        "datetime": "2016-05-17T12:09:39Z",
        "uri": "https://arquivo.pt/wayback/20160517120939mp_/https://www.osha.gov/pls/oshaweb/owadisp.show_document?p_table=standards&p_id=10051"
      },
      {
        "datetime": "2016-05-17T12:14:00Z",
        "uri": "https://arquivo.pt/wayback/20160517121400mp_/https://www.osha.gov/pls/oshaweb/owadisp.show_document?p_table=STANDARDS&p_id=10075"
      },
      {
        "datetime": "2016-05-17T13:22:49Z",
        "uri": "https://arquivo.pt/wayback/20160517132249mp_/https://www.osha.gov/pls/oshaweb/owadisp.show_document?p_id=10106&p_table=STANDARDS"
      },
```


https://pywb.readthedocs.io/en/latest/_modules/pywb/warcserver/index/fuzzymatcher.html

4,616 URI-Rs in the dataset contain queries, so these timemaps need to be filtered



Once these timemaps are filtered, all timemaps are ready to be processed.

*only need to compare mementos from web archives that implement (pywb) fuzzy query matching

Using timemaps, there are 8,500 additional paired mementos in the EDGI dataset

It's both surprising that and unclear why there is a paired memento from IA in the no-pair dataset


```
{ 'id': '00001' , 'original_uri': 'https://www3.epa.gov/' , 'urim_2016': {'datetime':
'2016-01-01T00:48:22Z', 'uri':
'https://web.archive.org/web/20160101004822/http://epa.gov/'} , 'urim_2020':
{'datetime': '2020-01-01T01:01:08Z', 'uri': 'https://wayback.archive-
it.org/all/20200101010108/https://www.epa.gov/'} },
{ 'id': '00002' , 'original_uri': 'https://www3.epa.gov/enviro/facts/ghg/search.html'
, 'urim_2016': {'datetime': '2016-03-18T13:49:01Z', 'uri':
'https://web.archive.org/web/20160318134901/http://www3.epa.gov:80/enviro/facts/ghg/se
arch.html'} , 'urim_2020': {'datetime': '2020-01-01T04:23:18Z', 'uri':
'https://web.archive.org/web/20200101042318/https://www3.epa.gov/enviro/facts/ghg/sear
ch.html'} },
{ 'id': '00003' , 'original_uri': 'https://www3.epa.gov/epafiles/usenotice.htm' ,
'urim_2016': {'datetime': '2016-02-01T10:31:18Z', 'uri':
'https://web.archive.org/web/20160201103118/http://www.epa.gov/epafiles/usenotice.htm'
} , 'urim_2020': {'datetime': '2020-01-01T04:23:31Z', 'uri':
'https://web.archive.org/web/20200101042331/https://www3.epa.gov/epafiles/usenotice.ht
m'} },
```

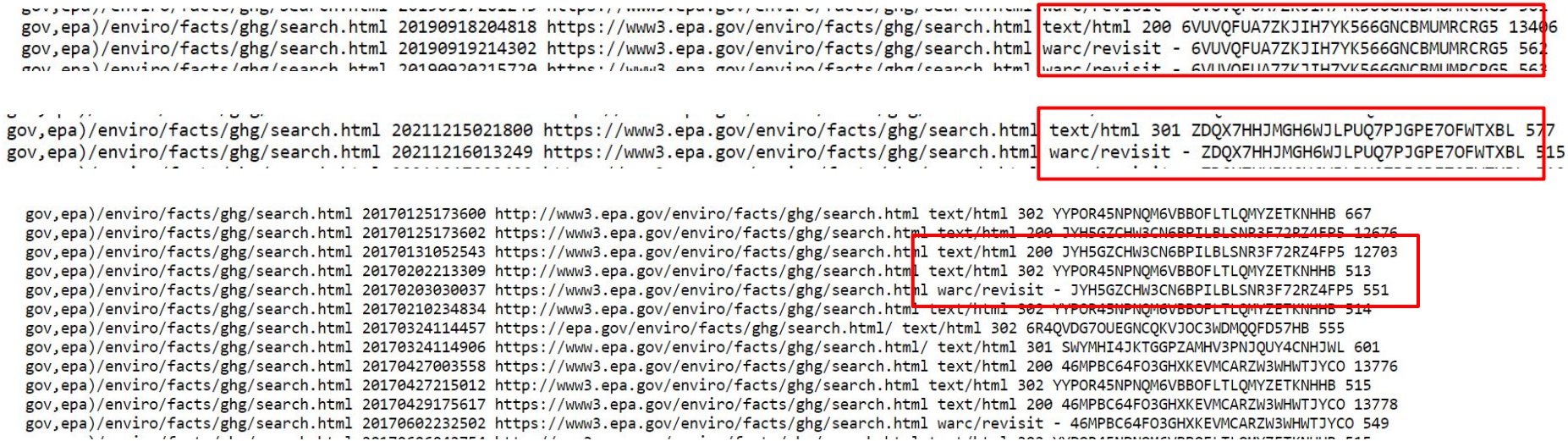
EDGI skipped some of the CDX mementos, but this code didn't do what they thought it did

```
with internetarchive.WaybackClient() as client:
    dump = client.list_versions(thisPage, from_date=datetime(dates[0], dates[1],dates[2]), to_date=datetime(dates[3], dates[4], dates[5])) #
    versions = reversed(list(dump))
    for version in versions: # For each version in all the snapshots
        if version.status_code == '200' or version.status_code == '-': # If the IA snapshot was viable...
            url=version.raw_url
            contents = requests.get(url, timeout=120).content.decode() # Decode the url's HTML # Handle the request so that it doesn't hang
```

warc/revisit (indicated by dash) doesn't guarantee a 200 status code

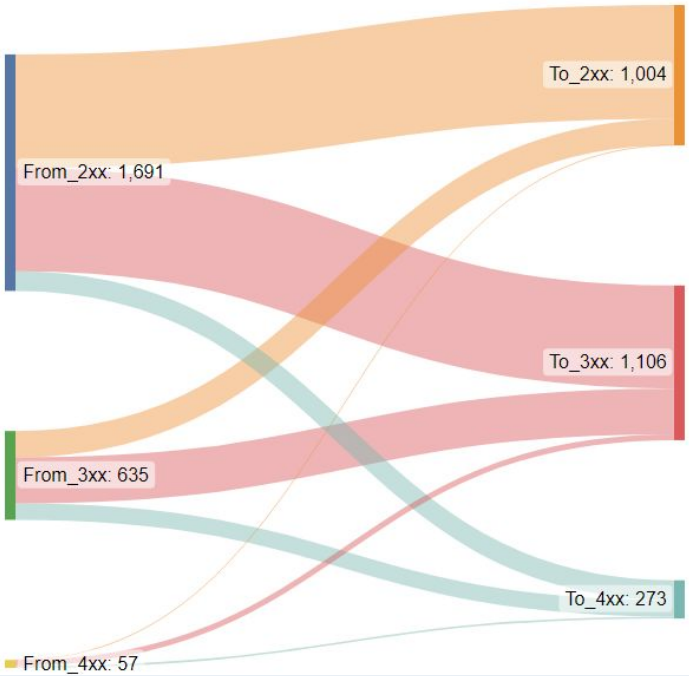
https://github.com/edgi-govdata-archiving/web_monitoring_research/blob/main/ctrl-f.py

warc-revisit doesn't guarantee a 200 status code



http://web.archive.org/cdx/search/cdx?url=https://www3.epa.gov/enviro/facts/ghg/search.html

There are at least 1,000 pairs at IA with 200 to 200 status codes in the new pair list



https://sankeymatic.com/build/

Perhaps WARCs were added to the Wayback Machine after the original study?

(are web archives' contents stable?)

Do any of the new pairs rely on archives besides the Internet Archive?

- Of the new pair URI-Rs, calculate when no memento in the first half of 2016 or 2020 is from the Internet Archive using the time maps, and keep only those pairs.


```
{
  "datetime": "2016-03-16T11:04:19Z",
  "uri": "https://arquivo.pt/wayback/20160316110419mp_/https://www.epa.gov/chemical-research/stochastic-human-exposure-and-dose-simulation-sheds-estimate-human-exposu
},
{
  "datetime": "2016-04-18T18:29:55Z",
  "uri": "https://wayback.archive-it.org/all/20160418182955/https://www.epa.gov/chemical-research/stochastic-human-exposure-and-dose-simulation-sheds-estimate-human-e
},
{
  "datetime": "2016-04-18T18:29:55Z",
  "uri": "https://web.archive.org/web/20160418182955/https://www.epa.gov/chemical-research/stochastic-human-exposure-and-dose-simulation-sheds-estimate-human-exposure
},
{
```

50% of the new paired mementos rely on archives besides the Internet Archive/Wayback Machine!

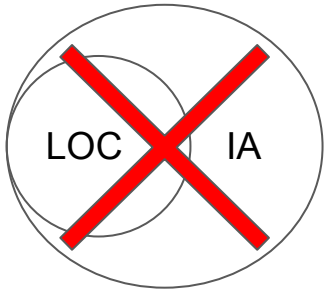
archive	reliance count
webarchive.loc.gov	4999
arquivo.pt	213
wayback.archive-it.org	31
archive.md	1
perma.cc	1
waext.banq.qc.ca	1

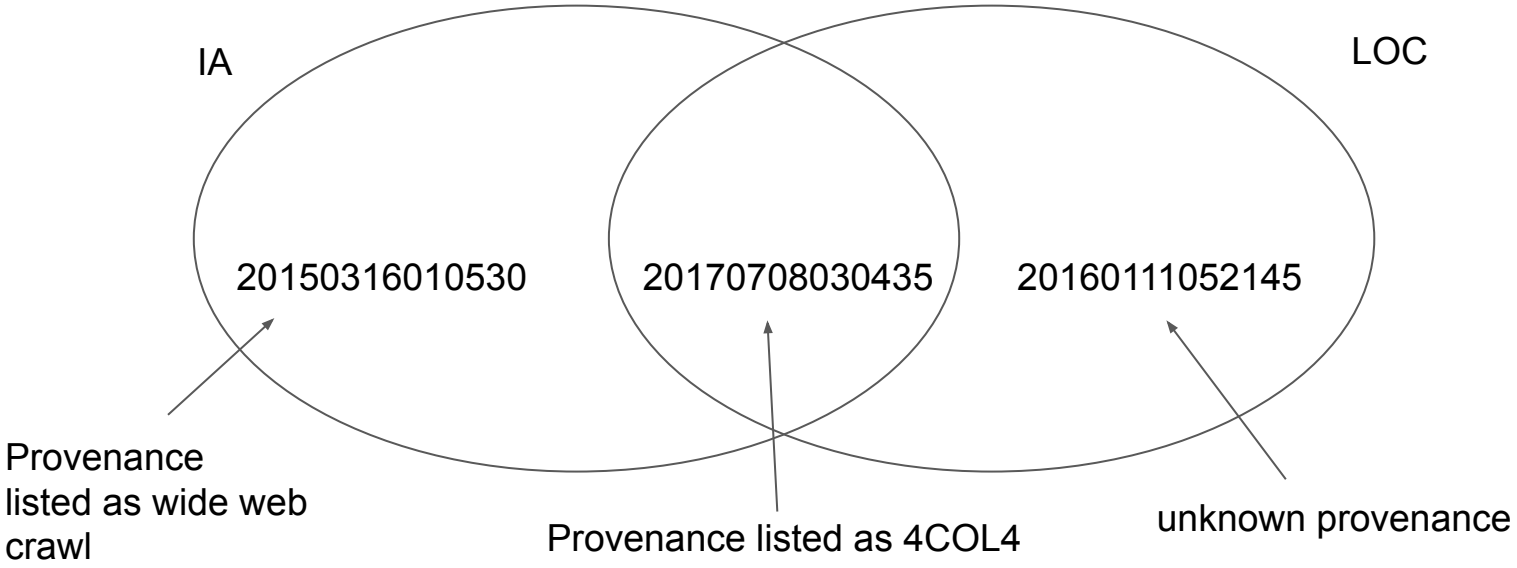
These all have a May 2016 timestamp, suggesting a crawl

This shows that LOC is a very important resource when researching government website mementos

If both mementos in a pair don't rely on IA, the pair will be double counted in the chart above

LOC and IA share some, but not all, captures



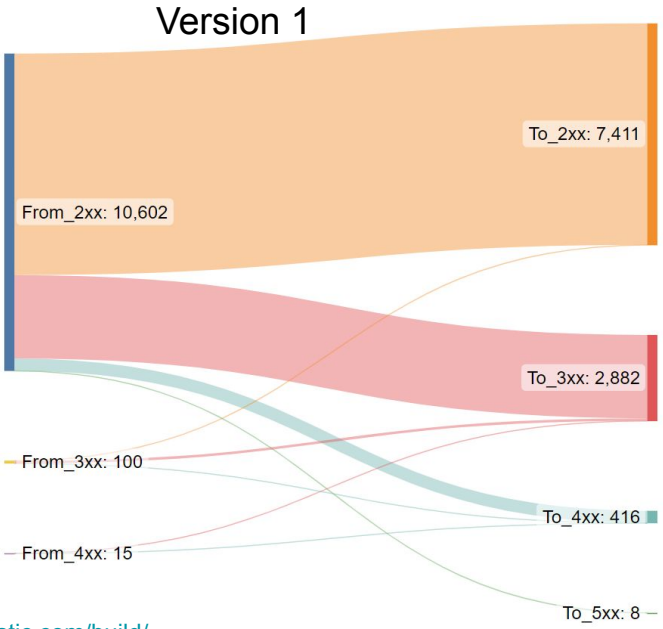


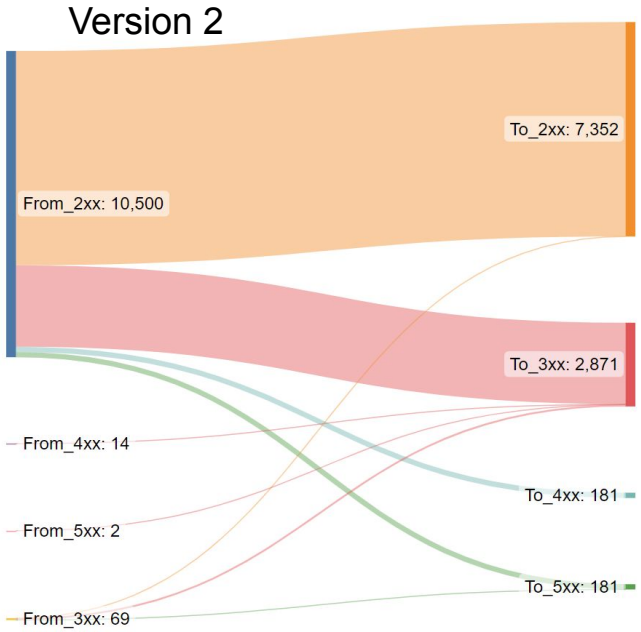
From timemap of http://www.osha.gov/dte/edcenters/

Investigation 2: What were the status codes of the original paired mementos?

- Script Version 1: calculate the status codes using the datetimes of the original paired mementos, including look-up of warc-revisits
- Script Version 2: Look through all mementos from the first half of 2016 and 2020, choosing a 200 if it exists, and filtering out warc-revisits

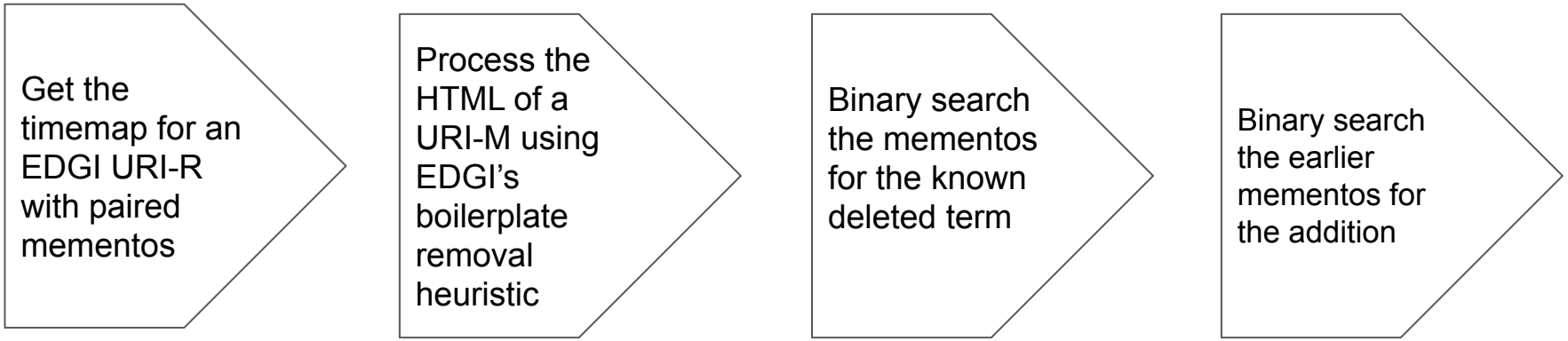
Both methods were comparable, though the second method found more 5xx status code mementos





https://sankeymatic.com/build/

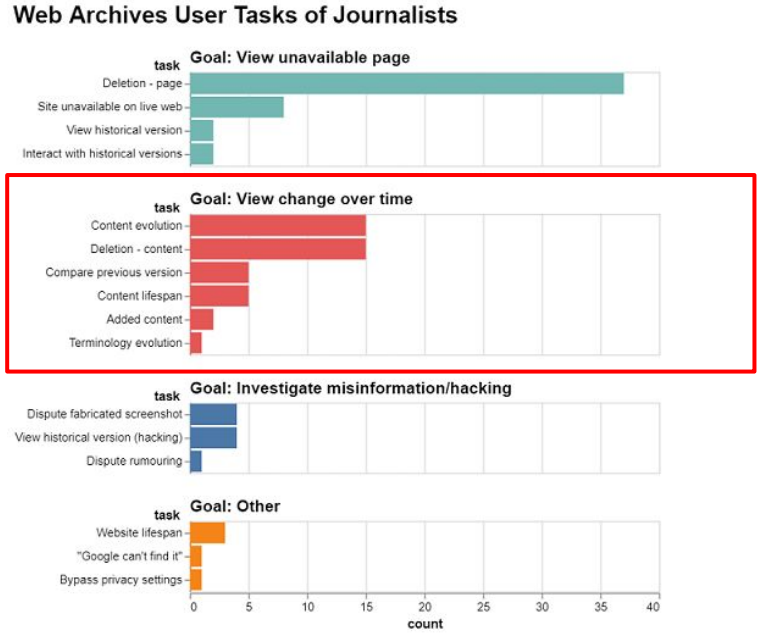
Investigation 3: Timemaps are essential in calculating finer granularity for the paired mementos



Workflow of granularity.py

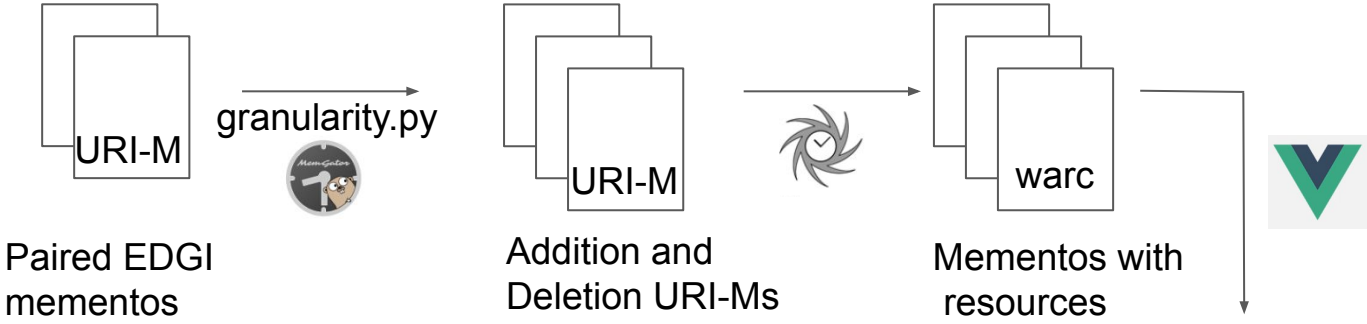
Web archive users viewing change over time are interested in change text and its temporal component

Accurate temporal calculations are important to these users!



https://ws-dl.blogspot.com/2022/08/2022-08-04-web-archiving-in-popular.html

Many tools contribute to the workflow of the search engine



- Memgator
- Hypercane
- SolrWayback
- Solarium

https://commons.wikimedia.org/wiki/File:Search_Engine_Results_Page_(SERP)_Graphic_Illustration.png

The temporal calculations on the search engine results page are more accurate

LOC memento →

title	ESRL Global Monitoring Division - Trinidad Head
url	https://www.esrl.noaa.gov/gmd/obop/thd/
pre-deletion	2017-02-03 15:02:50
post-deletion	2017-04-26 22:14:50

4 years → 3 months

diff

```
Differences
- Because of the characteristics of a relatively remote coastal location
  (insignificant anthropogenic influences and prevailing maritime airflow)
  the Trinidad Head site is an important location, providing and opportuni
  ty to observe and monitor both regional and global influences.
+ NOAA established an atmospheric baseline observatory at Trinidad Head in
  2002.
+ Because of its relatively remote coastal location and prevailing maritim
  e airflow, NOAA felt the site would provide scientists with an opportuni
  ty to observe and monitor both regional and global atmospheric condition
  s reasonably free from local influences.
```

addition	2009-10-19 15:02:25
content lifespan	2746 days

2016 (inaccurate) → 2009 (accurate)

4 years (inaccurate) → 7.5 years (accurate)

Frew, Nelson, and Weigle.
Work in Progress.

Each investigation has important future work still to do

- How many URI-Rs in the original dataset are truly unarchived?
  - Calculate which URI-Rs with empty time maps have 404 timemaps
- What is the baseline amount of change on a page with a term deletion?
  - Use CDX byte sizes and/or word counts to investigate soft 404s
- How can we increase the granularity of a page with a deleted phrase like “climate change”?
  - Extend the granularity script to work for phrases

Conclusion: aggregating multiple web archives is vital for showing change over time

- Aggregating multiple web archives increases the number of paired mementos
- 75% of the original paired mementos have 200 to 200 status codes
- Aggregating multiple web archives increases the granularity of the web page change text calculations