Where Did the Web Archive Go?

(Published in TPDL 2021)

Mohamed Aturban[1] Michael L. Nelson[2] Michele C. Weigle[2]

[1]Columbia College, Columbia MO
[2]Old Dominion University, Norfolk VA

Presented by – Yasith Jayawardana (@yasithdev)
Old Dominion University, Norfolk VA

CS895 Web Archiving Forensics, Fall 2022

ODU

Outline

Headers in the Memento Protocol

An HTTP protocol extension for **temporal** content negotiation of web resources[1]

Request Headers

Accept-Datetime: <approximate datetime of the memento>

[1] https://www.slideshare.net/hvdsomp/memento-101

ODU

Headers in the Memento Protocol

An HTTP protocol extension for **temporal** content negotiation of web resources[1]

Request Headers

Accept-Datetime: <approximate datetime of the memento>

Response Headers

Memento-Datetime: <exact datetime of the memento>
Vary: Accept-Datetime, ...
Link: <URL>; rel="<original/timegate/timemap/memento>", ...
Location: <URI-M of the memento>

[1] https://www.slideshare.net/hvdsomp/memento-101

ODU

Headers in the Memento Protocol

An HTTP protocol extension for **temporal** content negotiation of web resources[1]

Request Headers

Accept-Datetime: <approximate datetime of the memento>

Response Headers

Memento-Datetime: <exact datetime of the memento>
Vary: Accept-Datetime, ...
Link: <URL>; rel="<original/timegate/timemap/memento>", ...
Location: <URI-M of the memento>

- Datetime should comply with **RFC 9110** section 5.6.7
- For **Link** header entries
  - If rel="memento" ⇒ datetime="" is mandatory
  - If rel="timemap" ⇒ type="" is recommended

[1] https://www.slideshare.net/hvdsomp/memento-101

ODU

Using the Memento Protocol

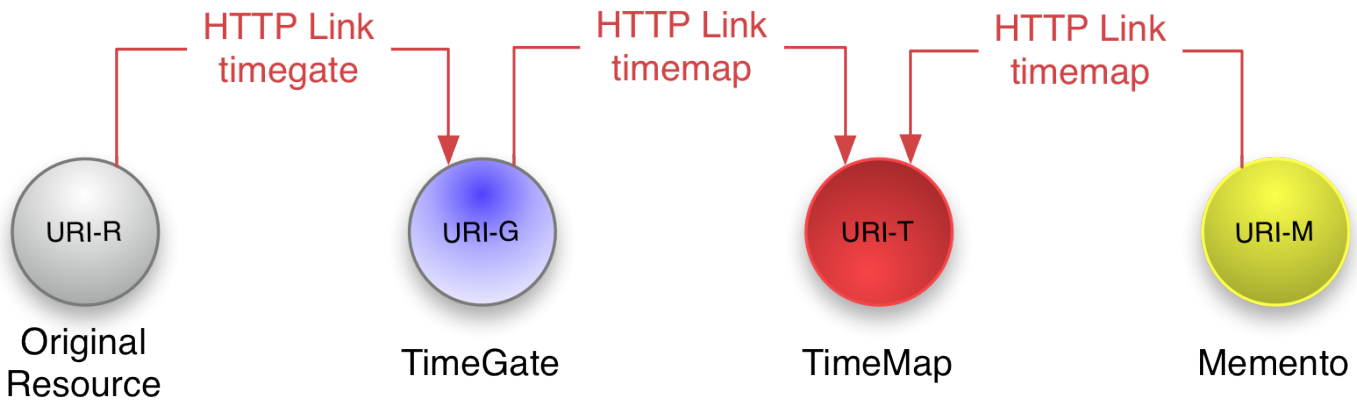


Figure: Components of the Memento Protocol (mementoweb.org)

Using the Memento Protocol

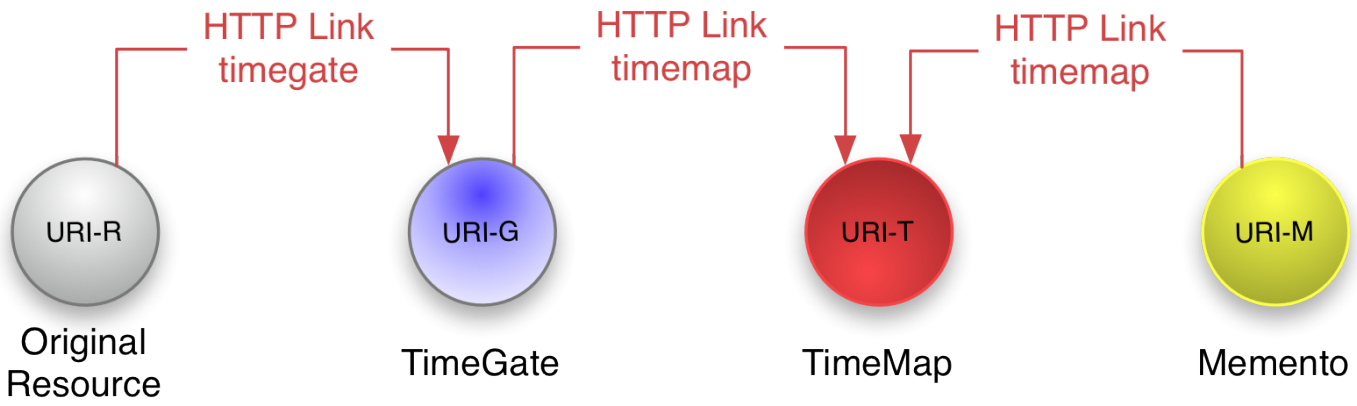


Figure: Components of the Memento Protocol (mementoweb.org)

Requesting with Accept-Datetime Headers

(req) **URI-R** → (res) Link: **URI-G**
(req) **URI-G** → (res) Link: **URI-R**, **URI-T**, Location: temporally closest **URI-M**
(req) **URI-T** → (res) list of **URI-M**
(req) **URI-M** → (res) Link: **URI-R**, **URI-T**, **URI-G**

Example URI-M

http://www.collectionscanada.gc.ca/webarchives/20060208075019/http://www.cdc.gov/

Outline

Not even archived resources are permanent [2]

The web archives might...

Go Defunct (funding issues, etc.)

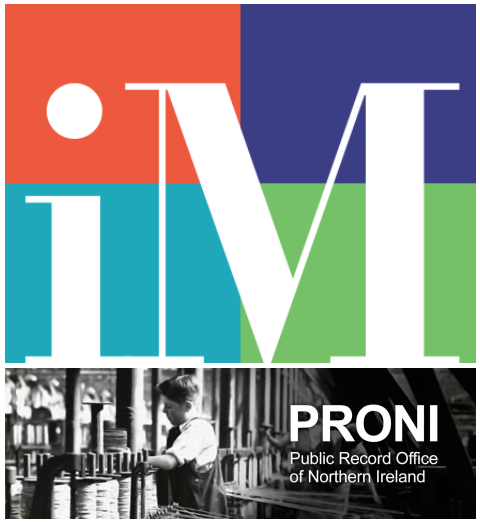


Figure: Now Defunct Web Archives

- Internet Memory Foundation (formerly European Archive Foundation)
- PRONI (now in archive-it.org)

[2] https://en.wikipedia.org/wiki/Wikipedia:List_of_web_archives_on_Wikipedia

ODU

Not even archived resources are permanent [2]

The web archives might...

Go Defunct (funding issues, etc.)

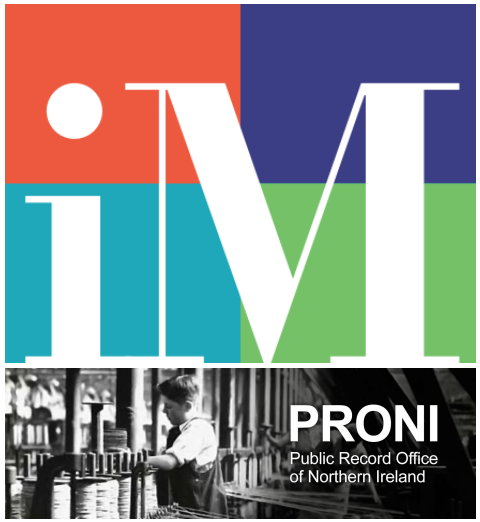


Figure: Now Defunct Web Archives

- Internet Memory Foundation (formerly European Archive Foundation)
- PRONI (now in archive-it.org)

Change Infrastructure

```
$ curl -i https://perma-archives.org
HTTP/2 301
date: Mon, 24 Oct 2022 03:03:52 GMT
location: https://perma.cc/
```

Figure: perma-archives.org is now perma.cc

- perma-archives.org
  → perma.cc
- collectionscanada.gc.ca
  → bac-lac.gc.ca
  → library-archives.canada.ca

[2] https://en.wikipedia.org/wiki/Wikipedia:List_of_web_archives_on_Wikipedia

ODU

16,627 Mementos, **3,698** URI-Rs, **17** Public Web Archives

Collected from 2017-11 to 2019-01

Table 1: A set of 17 public web archives.

Archive URI	Archive Name	Purpose
swap.stanford.edu	Stanford Web Archive Portal	General
web.archive.org	The Internet Archive	General and on-demand
archive.bibalex.org	Bibliotheca Alexandrina's Internet Archive	National
arquivo.pt	The Portuguese Web Archive (PWA)	National
collectionscanada.gc.ca	Library and Archives Canada	National
digar.ee	The Estonian Web Archive	National
nationalarchives.gov.uk	The National Archives	National
vefsafn.is	The Icelandic Web Archive	National
webarchive.loc.gov	Library of Congress Web Archives	National
webarchive.org.uk	The UK Web Archive (UKWA)	National
webarchive.proni.gov.uk	Public Record Office of Northern Ireland (PRONI)	National
webharvest.gov	Congressional & Federal Government Web Harvests	National
archive-it.org	Archive-It - Web Archiving Services for Libraries and Archives	On-demand
archive.is	Archive.is	On-demand
perma.cc	Perma.cc	On-demand
webcitation.org	WebCite	On-demand
europarchive.org	The European Archive	Organizational

Figure: Public web archives used in the study (source: "Collecting 16K archived web pages from 17 public web archives", Aturban et al., 2019)

ODU

Outline

4 Web Archives (**1,981** Mementos) Changed Base URIs

With no redirection to their new base URIs

- **Library and Archives Canada** (351 Mementos)
  49 changed, 2 missing
- **National Library of Ireland** (979 Mementos)
  192 changed, 0 missing
- **Public Record Office of Northern Ireland** (469 Mementos)
  114 changed, 0 missing
- **Perma.cc** (182 Mementos)
  164 changed, 18 missing

Overall Statistics

- **Total**: 1,981
- **Changed**: 537 (27.11%)
- **Missing**: 20 (1.01%)

ODU

Breakdown of Changed (**537**) and Missing (**20**) Mementos

Criteria – "Memento-Datetime" Header / HTTP Status / URI-R Changes

Table 1: Web archive changes based on how mementos changed. The number of missing mementos is shown in **bold**.

Original archive → New archive	**Same Memento-Datetimes?**	**Same status codes?**	**Same URI-Rs?**	**URI-Ms**
collectionscanada.gc.ca → bac-lac.gc.ca	Yes	Yes	Yes	302
	NO	Yes	Yes	**28**
	NO	Yes	NO	**18**
	NO	NO	Yes	**1**
	NO	NO	NO	**2**
europarchive.org/NLI → internetmemory.org/NLI	Yes	Yes	Yes	979
internetmemory.org/NLI → archive-it.org	Yes	Yes	Yes	787
	Yes	NO	Yes	**1**
	Yes	NO	NO	**2**
	NO	Yes	Yes	**184**
	NO	Yes	NO	**5**
proni.gov.uk → archive-it.org	Yes	Yes	Yes	355
	Yes	NO	Yes	**2**
	NO	Yes	Yes	**106**
	NO	Yes	NO	**6**
perma-archives.org → perma.cc	NO	Yes	Yes	**164**
	NO	NO	NO	**18**

Why did "Memento-Datetime" Headers Change?

The new archive may have post-processed the original WARC files

Library and Archives Canada

351 mementos, 49 changed, 2 missing

- Old mementos incorrectly redirected to the main webpage
- New mementos were accessible from webarchive.bac-lac.gc.ca
- Some new mementos didn't match old status codes / headers

```
http://www.collectionscanada.gc.ca/webarchives/20070220181041/http://www.
    berlin.gc.ca/ (302)
http://www.collectionscanada.gc.ca/webarchives/20070220181041/http://www.
    dfait-maeci.gc.ca/canadaeuropa/germany/ (302)
http://www.collectionscanada.gc.ca/webarchives/20070220181204/http://www.
    dfait-maeci.gc.ca/canadaeuropa/germany/ (302)
http://www.collectionscanada.gc.ca/webarchives/20070220181204/http://www.
    international.gc.ca/global/errors/404.asp?404%3Bhttp://www.dfait-
    maeci.gc.ca/canadaeuropa/germany/ (404)
```

Fig. 4: The HTTP status codes of the URI-M http://www.collectionscanada.gc.ca/webarchives/20070220181041/http://www.berlin.gc.ca/ from the original archive.

```
http://webarchive.bac-lac.gc.ca:8080/wayback/20070220181041/http://www.
    berlin.gc.ca/ (Redirect by JavaScript (JS))
http://webarchive.bac-lac.gc.ca:8080/wayback/20070220181041/http://www.
    dfait-maeci.gc.ca/canadaeuropa/germany/ (Redirect by JS)
http://webarchive.bac-lac.gc.ca:8080/wayback/20070220181204/http://www.
    international.gc.ca/global/errors/404.asp?404%3Bhttp://www.dfait-
    maeci.gc.ca/canadaeuropa/germany/ (Redirect by JS)
http://webarchive.bac-lac.gc.ca:8080/wayback/20071115025620/http://www.
    international.gc.ca/canada-europa/germany/ (302)
http://webarchive.bac-lac.gc.ca:8080/wayback/20071115023828/http://www.
    international.gc.ca/canada-europa/germany/ (200)
```

Fig. 5: The HTTP status codes of the URI-M http://webarchive.bac-lac.gc.ca:8080/wayback/20070220181041/http://www.berlin.gc.ca/ from the new archive.

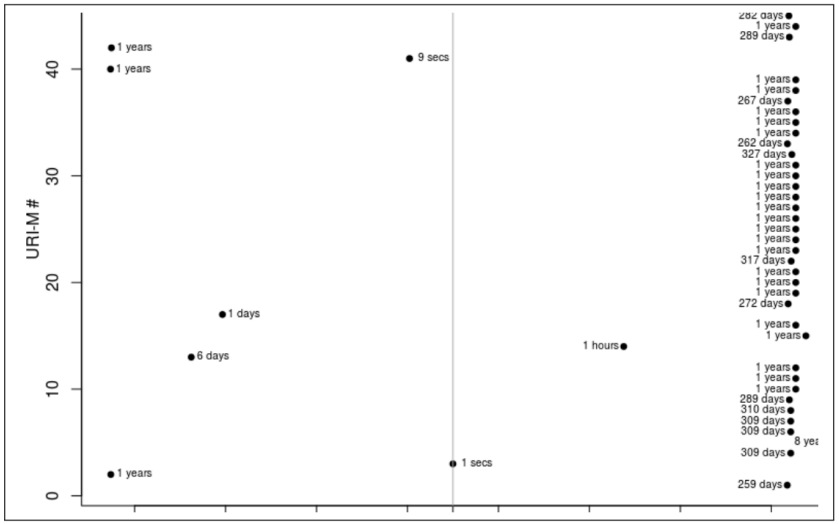


Fig. 3: Difference between the Memento-Datetimes for URI-Ms from www.collectionscanada.gc.ca and the corresponding URI-Ms from webarchive.bac-lac.gc.ca.

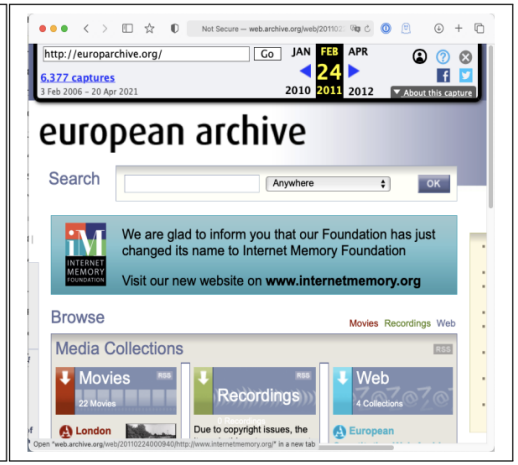
Suggested URL Rewrite

```
# With mod_rewrite
RewriteEngine on
RewriteRule   "^/webarchives/(\d{14})/(.+)"
              http://webarchive.bac-lac.gc.ca:8080/wayback/$1/$2
              [L,R=301]
```

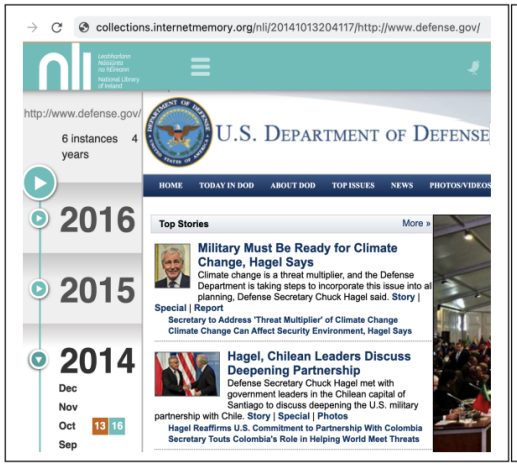
National Library of Ireland Changed Infrastructure Twice

(1) europarchive.org/NLI → internetmemory.org/NLI
(2) internetmemory.org/NLI → wayback.archive-it.org/10702

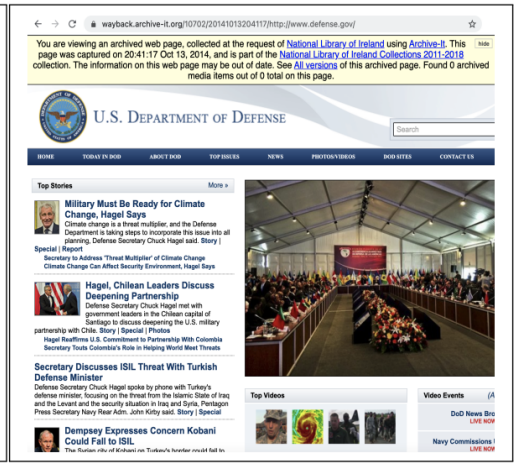
(a) In europarchive.org



(b) European Archive announcing their name change to IMF



(c) In internetmemory.org



(d) In archive-it.org

National Library of Ireland

979 mementos, 192 changed, 0 missing

- europarchive.org → internetmemory.org – **0 issues**
- internetmemory.org/NLI → wayback.archive-it.org/10702 – **192 issues**
- Some new mementos didn't match old status codes / headers

```
$ curl --head --location --silent http://wayback.archive-it.org
    /10702/20121221162201/http://bbc.co.uk/news/ | egrep -i "(HTTP/|^
    location:|^Memento-Datetime)"

HTTP/1.1 302 Found
Location: /10702/20121221163248/http://www.bbc.co.uk/news/
HTTP/1.1 200 OK
Memento-Datetime: Fri, 21 Dec 2012 16:32:48 GMT
```

Fig. 7: The HTTP status codes of the URI-M http://wayback.archive-it.org/10702/20121221162201/http://bbc.co.uk/news/ from the new archive archive-it.org.

```
$ curl --head --silent http://wayback.archive-it.org
    /10702/20121223031837/http://www2008.org/

HTTP/1.1 404 Not Found
Server: Apache-Coyote/1.1
Content-Security-Policy-Report-Only: default-src "self"
[...]
Content-Type: text/html;charset=utf-8
Content-Length: 4902
Date: Thu, 05 Sep 2019 08:28:27 GMT
```

Fig. 9: The HTTP status codes of the URI-M http://wayback.archive-it.org/10702/20121223031837/http://www2008.org/ from the new archive archive-it.org.

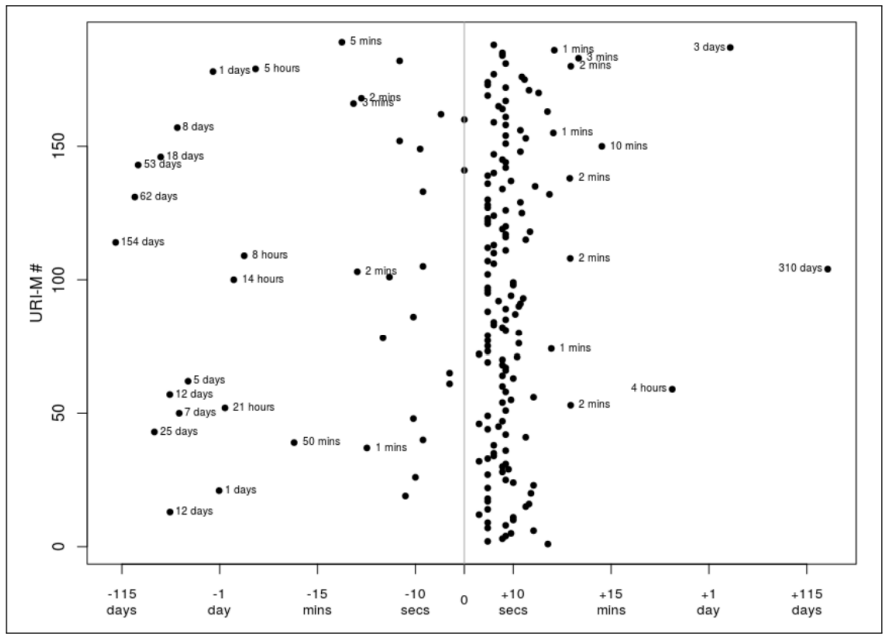


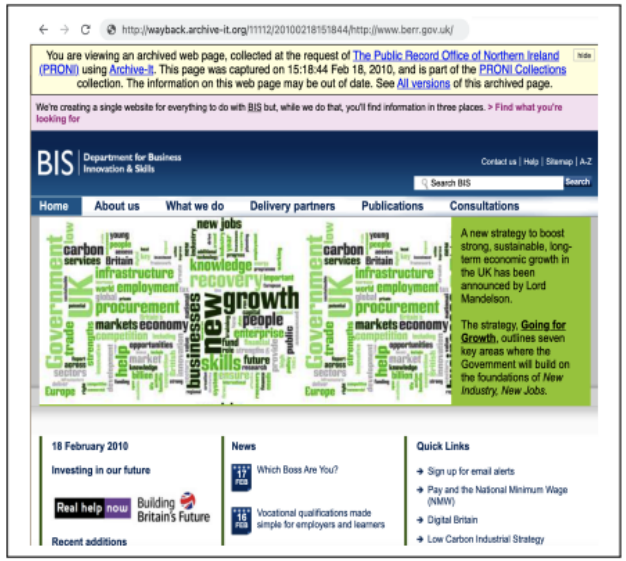
Fig. 8: Difference between the Memento-Datetimes for URI-Ms from internetmemory.org and the corresponding URI-Ms from archive-it.org.

Public Record Office of Northern Ireland (PRONI)

469 mementos, 114 changed, 0 missing

- webarchive.proni.gov.uk → archive-it.org/collections/11112

(a) In webarchive.proni.gov.uk



(b) In archive-it

```
$ curl --head --location http://webarchive.proni.gov.uk/20100218151844/
    http://www.berr.gov.uk/

HTTP/1.1 302 Found
Cache-Control: no-cache
Content-length: 0
Location: https://webarchive.proni.gov.uk/20100218151844/http://www.berr.
    gov.uk/
HTTP/2 404
date: Fri, 20 Sep 2019 08:13:45 GMT
server: Apache/2.4.18 (Ubuntu)
content-type: text/html; charset=iso-8859-1
```

Fig. 12: The HTTP status codes of the URI-M http://webarchive.proni.gov.uk/20100218151844/http://www.berr.gov.uk/ from the original archive webarchive.proni.gov.uk.

```
# With mod_rewrite
RewriteEngine on
RewriteRule "^/(\d{14})/(.+)"
            http://wayback.archive-it.org/11112/$1/$2 [L,R=301]
```

Fig. 14: The Apache `mod_rewrite` rules that can be used to handle redirects from webarchive.proni.gov.uk to archive-it.org.

ODU

Perma.cc

182 mementos, 164 changed, 18 missing

- perma-archives.org/warc/<datetime> → perma.cc/<id>
- Since 2020-02 perma.cc supports the Memento protocol
- New mementos don't have the 14-digit memento-datetime in the URI-M
  e.g., `https://perma.cc/T8U2-994F`

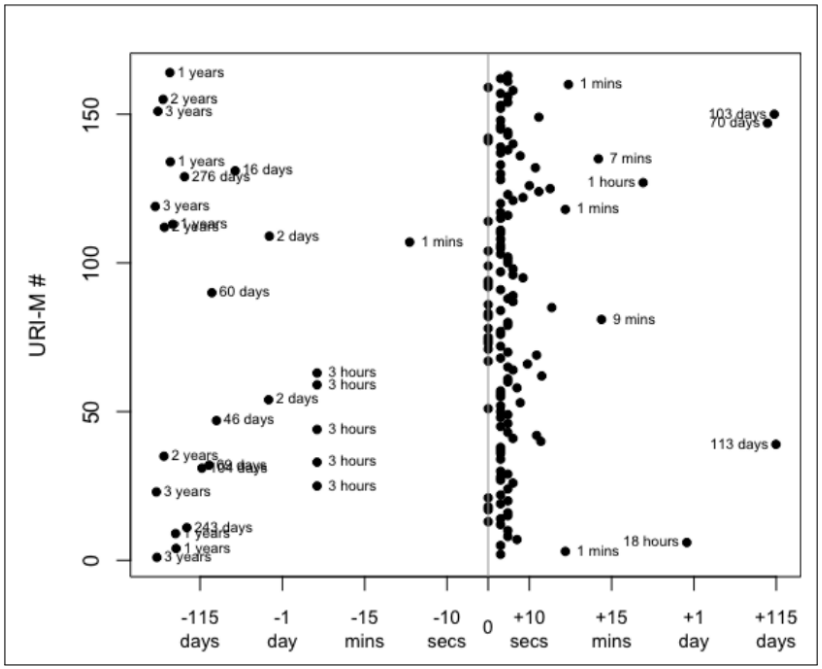


Fig. 16: Difference between the Memento-Datetimes for the long-form Perma.cc URI-Ms and the corresponding short-form URI-Ms.

Summary

Investigation

- Goal – detect variations and changes in replaying individual mementos
- Sample – 16,627 mementos between 2017-11-01 and 2019-01-01 from 17 public web archives

Findings

- 4 web archives changed their base URIs and left no *machine-readable* method to locate their new base URIs
- 537 of 1,981 mementos from these web archives were impacted
  - 517 rediscovered with changes to their Memento-Datetime, URI-R, or HTTP status code
  - 20 could not be found at all
- If a Web archive changes its base URL, setting rewrite rules may help to avoid broken URI-Ms

ODU