



# Aggregator Reuse and Extension for Richer Web Archive Interaction

Mat Kelly

**Drexel University** 

mkelly@drexel.edu

@machawk1

The 24th International Conference on Asia-Pacific Digital Libraries (ICADL 2022) November 30 - December 2, 2022

#### The Web as History

- Our history on the web
- The source(s) of archives dictate the narrative
- More sources  $\rightarrow$  More comprehensive narrative



#### The Webs as History

- Long tail of web archives
- Not all {public,private} content is fit for {public,private} consumption
- The "key" of a URI may need expansion for additional contexts





#### Time on the Web(s)



- Memento provides temporal negotiation, syntax, and semantics
- Memento aggregation allow for efficiently querying collective sources
- One endpoint solely for aggregation has additional functional potential

#### Questions to Consider

- RQ1: In what was can we leverage Memento aggregators for more sophisticated querying of web archives?
- RQ2: With newfound hierarchical use cases, what problems should we anticipate and how can they be mitigated?

#### Related Work – Time Travel



#### • De facto official Memento aggregation

• • • Time Travel × +	• • • Mementos for http://icadl.net a x +	•
$\leftarrow \rightarrow \mathbf{C}$ $\cong$ timetravel.mementoweb.org	← → C ▲ Not Secure   timetravel.mementoweb.org/list/20040512170041/https://icadl.net	🖈 🖈 😸 Incognito :
Time Travel	Time Travel	About API Privacy Terms
time travel	http://icadl.net       2004-05-12     17:00:41     Find     Reconstruct     12 May	y 2004 17:00:41 GMT
http://	Mementos closest to the requested date 12 May 2004 17:00:41 GMT	
2014-10-20 23:25:34 Find Reconstruct	Arquivo.pt Memento, 1 hour before Embed https://arquivo.pt/wayback/20040512170041mp_/http://icadl.net/ 12 May 2004 17:00:41 GMT [ <1 hour from requested date ]	experience web time travel
Find Mementos in Internet Archive, Archive-It, British Library, archive.today, GitHub and many more!	Previous Memento data not provided data not provided	install Memento for Chrome
	First Memento data not provided data not provided	enable web time travel
		[[]]]
	Internet Archive Memento, 13 years 358 days after Embed https://web.archive.org/web/20180503103914/http://icadi.net/ 03 May 2018 10:39/14 GMT [+1:3 years 358 days from requested date ]	MEDIAWIKI install Memento for MediaWiki
https://mementoweb.org/	Previous Memento Next Memento   data not provided 15 Aug 2020 05:03:20 GMT   First Memento [+16 years 98 days ]   data not provided 29 Nov 2020 17:03:49 GMT   [+16 years 205 days ]	say no to "404 Not Found" Robust Links use links that refuse to die

Aggregator Reuse and Extension for Richer Web Archive Interaction @machawk1

#### Related Work – MemGator



- Open source Memento aggregator
- Ripe for:
  - Distribution, Decentralization
  - Personal/Private use
  - Modification & Adaptation
- Used extensively in web archive research, less so an introspecting the aggregation process and component

Alam & Nelson. MemGator - A Portable Concurrent Memento Aggregator: Cross-Platform CLI and Server Binaries in Go. JCDL 2016, pp. 243-244

#### Related Work – Mementity Framework

- Aggregators as means of combining captures from both private & public web archives
- Introduced "Memento Meta Aggregator"
- Outstanding issues on MMA



Kelly, Nelson, and Weigle, "A Framework for Aggregating Private and Public Web Archives," In Proc. of JCDL 2018, pp. 273-282.

#### Related Work – Prefer

- HTTP Prefer for expression & consumption of clients' preferences
- Prior work re: web archiving deals w/ custom set of archives to-be-aggregated



Kelly, Alam, Nelson, and Weigle, "Client-Assisted Memento Aggregation Using the Prefer Header," Presented at the Web Archiving and Digital Libraries Workshop (WADL 2018)



Aggregator Reuse and Extension for Richer Web Archive Interaction @machawk1

#### **URI** Templating

- t<sub>0</sub>: {scheme & hostname}/{resource type}/{format}/{URI-R}
- t<sub>1</sub>: https://myarchive.org/timemap/link/http://example.com
- $m_0: \{ scheme \& hostname \} / \{ datetime \} / \{ URI-R \} \}$
- m1: http://archive.md/20210619183508/https://icadl.net/icadl2021/
- m<sub>2</sub>: https://archive.ph/eoQRZ
- t<sub>0</sub>: Templated URL of a capture of the past web (memento)
  - Replicated with t<sub>1</sub>
- m<sub>0</sub>: A URI-M template
  - m<sub>1</sub>: datetime and URI-R apparent
  - m<sub>2</sub>: URI-R need not be apparent, available in HTTP Link header (RFC)

#### Base Querying Models

- Proxy-style querying
- Conventional querying
- Aggregator Chaining



- Client relays request to a single archive
  - Base case for client-aggregator interaction

#### Base Querying Models

- Proxy-style querying
- Conventional querying
- Aggregator Chaining
- Client requests TimeMap containing all URI-Ms for a provided URI-R



## Base Querying Models

- Proxy-style querying
- Conventional querying
- Aggregator Chaining
- Allows for scoped archival knowledge by each aggregator
- Decentralized, facilitates inclusion of more sources w/o human intervention
- Useful, but potentially problematic...



- When a Tree Becomes a Graph
- Self-reference
- Duplication of Sources



- When a Tree Becomes a Graph
- Self-reference
- Duplication of Sources



- When a Tree Becomes a Graph
- Self-reference
- Duplication of Sources



 Reverse proxies may not expose service's own access points to service

- When a Tree Becomes a Graph
- Self-reference
- Duplication of Sources



- User-defined set of archives
- Cycle detection
- Rescoping the aggregator for client-side execution
- Preliminary results streaming

- User-defined set of archives
- Cycle detection
- Rescoping the aggregator for client-side execution
- Preliminary results streaming

- User-defined set of archives
- Cycle detection
- Rescoping the aggregator for client-side execution
- Preliminary results streaming

Nonce passed as query parameter independent of URI-R

- Allows short-circuiting
- Provide requestor indication that requestee was requestor in chain
- Potentially violates HTTP's statelessness.

web archive

web archive

archive

memento

web archive

- User-defined set of archives
- Cycle detection
- Rescoping the aggregator for client-side execution
- Preliminary results streaming



See github.com/machawk1/mink

- User-defined set of archives
- Cycle detection
- Rescoping the aggregator for client-side execution
- Preliminary results streaming



web archive



Aggregator Reuse and Extension for Richer Web Archive Interaction @machawk1

# Asynchronous Processing and Streaming Response



Aggregator Reuse and Extension for Richer Web Archive Interaction @machawk1

#### IN CLOSING...

#### More Capable Aggregator→More Usable Web Archives

- Formally defined client-aggregator querying models
- Aggregator chaining facilitates archival aggregation but is problematic
- Much functional potential remains untapped for open source Memento Aggregators

Potential next steps:

- Client-side aggregators
- Re-consider functional cohesion

