2018

Client-Assisted Memento Aggregation Using The Prefer Header

Mat Kelly
Old Dominion University

Sawood Alam
Old Dominion University

Michael L. Nelson
Old Dominion University

Michele C. Weigle
Old Dominion University

Original Publication Citation

Kelly, M., Alam, S., Nelson, M. L., & Weigle, M. C. (2018) *Client-assisted memento aggregation using the prefer header* [Conference Paper]. Web Archiving & Digital Libraries Workshop, Fort Worth, Texas. http://hdl.handle.net/10919/97988

Client-Assisted Memento Aggregation Using the Prefer Header

Mat Kelly, Sawood Alam, Michael L. Nelson, and Michele C. Weigle
Old Dominion University
Department of Computer Science
Norfolk, Virginia, USA
{mkelly,salam,mln,mweigle}@cs.odu.edu

1 INTRODUCTION

Preservation of the Web ensures that future generations have a picture of how the Web was. Web archives like Internet Archive's Wayback Machine[1], WebCite[2], and archive.is[3] allow individuals to submit URIs to be archived, but the captures they preserve then reside at the archives. Traversing these captures in time as preserved by multiple archive sources (using Memento [8]) provides a more comprehensive picture of the past Web than relying on a single archive. Some content on the Web, such as content behind authentication, may be unsuitable or inaccessible for preservation by these organizations. Furthermore, this content may be inappropriate for the organizations to preserve due to reasons of privacy or exposure of personally identifiable information [4]. However, preserving this content would ensure an even-more comprehensive picture of the Web and may be useful for future historians who wish to analyze content beyond the capability or suitability of archives created to preserve the public Web.

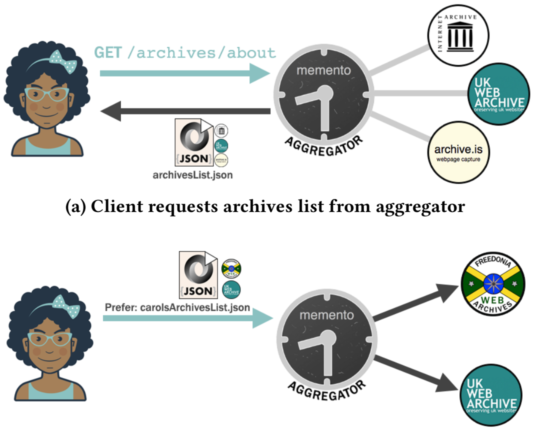
State-of-the-art Memento aggregators relay requests to a "static" set of archives. Thus, a client requesting an aggregated TimeMap has no say in which Web archives are used as the sources ($\{A_0\}$). By leveraging our previous work [4] of supplementing the capability of Memento aggregators (e.g., adding query precedence, aggregation short-circuiting, and multi-dimensional content negotiation of TimeMaps), we reuse this functionality for a more standards-based approach. This approach provides the novel contribution of involving the client's request in the Memento aggregation process beyond the specification of a URI-R and datetime.

More sophisticated aggregation may require filtering on a memento-level (e.g., only source mementos from archives with a certain quality of capture) or on a TimeMap-level. For instance, a user may wish to provide a previous unaggregated public archive (e.g., the "Freedonia Web Archive" in Figure 1b) or a private/personal Web archive as an additional source for aggregation. A conventional Memento aggregator may be required to provide additional parameters or communication flows to obtain mementos for a URI-R from private Web archives (as we discuss more in-depth in our preceding work [4]). In the current operation, a Memento aggregator assumes that all archives in a set are willing to provide a TimeMap in all instances. This may not be the case for a client's personal archive or a public Web archive that is not currently included in the aggregated set.

This submission represents a preliminary investigation in allowing the clients of Memento aggregators to be involved in determining the set of archives aggregated. In this work, we leverage the HTTP Prefer header [7]. Previous discussions have revolved around using Prefer for memento-level negotiation [5, 9]. This work considers using Prefer for TimeMap-level aggregation, particularly for the set of archives via archive specification instead of the representation of an individual memento.

2 BACKGROUND AND RELATED WORK

MemGator [2], the open-source Memento aggregator, provides conventional Memento aggregation with extended features including additional support for TimeMap formats beyond Link [6] and customization of the set of archives on startup of the aggregator software. CDXJ [1] is one such TimeMap format that is leveraged by the TimeMap endpoints in MemGator. Originally created as a replacement for CDX[4] files that act as an index to WARC [3] files, the CDXJ format allows for additional attributes about mementos to be specified within a JSON block. This capability allows for CDXJ-formatted TimeMaps to be much richer than Link-formatted TimeMaps due to the extensible semantics.



(a) Client requests archives list from aggregator

(b) Client supplies own list, potentially with custom attributes

Figure 1: A client first requests a list of aggregated archives from a Memento aggregator then modifies the response, encodes it, and supplies the encoded archive specification using the Prefer header for the aggregator to process.

Previously [4], we introduced the "Memento Meta-Aggregator" (MMA) concept to supplement functionality to conventional Memento aggregators using a hierarchical approach. There, we also introduced a rudimentary approach for a client to specify additional

[1] https://web.archive.org/
[2] http://www.webcitation.org/
[3] http://archive.is/
[4] https://iipc.github.io/warc-specifications/specifications/cdx-format/cdx-2015/

WADL'18, June 2018, Fort Worth, Texas USA
2018. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

```
> GET /timemap/link/http://fox.cs.vt.edu/wadl2017.html HTTP/1.1
> Host: mma.cs.odu.edu
> Prefer: archives="data:application/json;charset=utf-8;base64,Ww0KICB7...NCn0="

< HTTP/1.1 200
< content-type: application/link-format
< vary: prefer
< preference-applied: return=representation; archives="data:application/json;charset=utf-8;base64,Ww0KICB7...NCn0="
< content-location: /timemap/link/5bd...8e9/http://fox.cs.vt.edu/wadl2017.html
```

Figure 2: Client-side specification of a set of archives via encoded JSON using HTTP Prefer. The Memento aggregator responds with a location of a TimeMap for the URI-R at a URI-T representative of the set.

archives to an MMA using an ad hoc X-Archives HTTP request header. We also explored utilizing the Prefer [7] HTTP header to accomplish negotiation of mementos in dimensions beyond time, as may be facilitated with the usage of CDXJ TimeMaps.

Van de Sompel et al. [9] described using the Prefer header to distinguish mementos that have been rewritten when replaying Web archives to those with an untouched response body. By using Prefer header values like original-content and original-headers, a client may request that the representation return not be transformed by the Web archive.

Various presentations exist for an aggregator to use as the defining a set of archives to be aggregated, inclusive of definitions by MementoWeb.org[5], Webrecorder.io, and MemGator[6].

3 ARCHIVE SET SPECIFICATION WITH PREFER

An objective of this work is to allow a client of a Memento aggregator to be able to specify a custom set of archives ($\{A_f\}$) to be aggregated using standard syntax and semantics. We anticipate a 3-step process for a client to specify the archive set: **(1)** Client requests the set of archives to be aggregated by default from a Prefer-aware Memento aggregator (Figure 1a). **(2)** The aggregator returns the set of archives, e.g., as a JSON (per MemGator) or an XML (per mementoweb.org) file (Figure 1a), represented as $\{A_0\}$. **(3)** Once a response is received from the aggregator (e.g., https://git.io/archives), a client may manipulate the contents to be either an identical set ($\{A_f\} = \{A_0\}$), subset ($\{A_f\} \subset \{A_0\}$), supplementary set ($\{A_f\} \supset \{A_0\}$), or disjoint set ($\{A_f\} \dot{\bigcup} \{A_0\}$) (Figure 1b) and submit back to the aggregator for subsequent queries (Figure 2).

A client may also manipulate an existing archive's specification in the response received. For instance, a profiling probability (a value already defined in the MemGator specification) may be manipulated or a value of query precedence or short-circuiting may be modified, both of which we discussed in previous work [4].

Given that no Memento aggregator yet supports the client-side archive specification, we extend this idea with the assumption that a JSON response is received (like MemGator and Webrecorder's aggregator). A client may perform step 3 using the HTTP Prefer request header. After potentially manipulating the JSON response, a client would encode the JSON as a base64-encoded data URI (or supply some other URI for specification-by-reference) and submit a request with the Prefer header and a URI-R (Figure 2).

Archive supplementation may be accomplished using a hierarchical MMA approach (Figure 3), as we described in previous work [4]. This approach is necessary to adapt the capability of conventional Memento aggregators while still allowing them to be functionally cohesive. However, this hierarchical approach is insufficient if the a client would rather that subsequent queries to the aggregator after step 3 not be sent to certain archives supported by the base Memento aggregator. With the disclosure of the aggregated archives from a conventional aggregator (which is not conventionally exposed), an MMA could configure the default set from the conventional aggregator as the default to be queried and subsume the functions of the conventional aggregator.

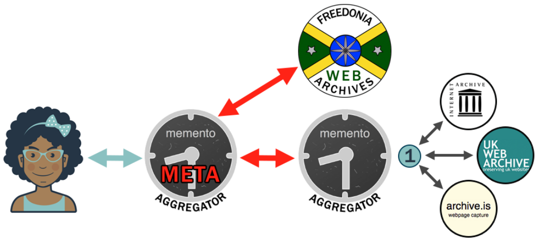


Figure 3: Using a hierarchical MMA approach, a previously unaggregated public Web archive may be aggregated with the results for a URI-R from a conventional Memento aggregator.

4 FUTURE WORK

In future work we will explore this approach's interoperability with using Prefer on mementos, which is ongoing research. We will also look to additional approaches like Cookies, and usage of CoRE's *well-known* syntax for archive specification.

REFERENCES

[1] Sawood Alam. 2015. CDXJ: An Object Resource Stream Serialization Format. http://ws-dl.blogspot.com/2015/09/2015-09-10-cdxj-object-resource-stream.html. (September 2015).

[2] Sawood Alam and Michael L. Nelson. 2016. MemGator - A Portable Concurrent Memento Aggregator: Cross-Platform CLI and Server Binaries in Go. In *Proceedings of JCDL*. 243–244.

[3] ISO 28500. 2009. WARC (Web ARChive) file format. http://www.digitalpreservation.gov/formats/fdd/fdd000236.shtml. (August 2009).

[4] Mat Kelly, Michael L. Nelson, and Michele C. Weigle. 2018. A Framework for Aggregating Private and Public Web Archives. In *Proceedings of JCDL*. Accepted for Publication.

[5] Ilya Kreymer. 2018. Feedback on new implementation of Prefer header in pywb. https://github.com/mementoweb/rfc-extensions/issues/7. (March 2018).

[6] M. Nottingham. 2017. Web Linking. IETF RFC 8288. (October 2017).

[7] J. Snell. 2014. Prefer Header for HTTP. IETF RFC 7240. (June 2014).

[8] Herbert Van de Sompel, Michael Nelson, and Robert Sanderson. 2013. HTTP Framework for Time-Based Access to Resource States – Memento. IETF RFC 7089. (December 2013).

[9] Herbert Van de Sompel, Michael L. Nelson, Lyudmila Balakireva, Martin Klein, Shawn M. Jones, and Harihar Shankar. 2016. Mementos In the Raw, Take Two. http://ws-dl.blogspot.com/2016/08/2016-08-15-mementos-in-raw-take-two.html. (August 2016).

[5] http://labs.mementoweb.org/aggregator_config/archivelist.xml

[6] https://git.io/archives