Interoperability for Accessing Versions of Web Resources with the Memento Protocol

Shawn M. Jones, Martin Klein, Herbert Van de Sompel, Michael L. Nelson, and Michele C. Weigle

Abstract The Internet Archive pioneered web archiving and remains the largest publicly accessible web archive hosting archived copies of webpages (Mementos) going back as far as early 1996. Its holdings have grown steadily since, and it hosts more than 475 billion URIs as of September 2019. However, the landscape of web archiving has changed significantly over the last two decades. There are more than 20 web archives around the world. This diversity contributes to the preservation of archived content that documents the past Web but requires standards to enable interoperability among them. The Memento Protocol is one of the main enablers of interoperability among web archives. We describe this protocol and present a variety of tools and services that leverage the broad adoption of the Memento Protocol and discuss a selection of research efforts made possible by these interoperability standards. In addition, we outline examples of technical specifications that enhance machines to access resource versions on the Web in an automatic, standardised and interoperable manner.

1 Introduction

The Internet Archive pioneered web archiving and remains the largest publicly accessible web archive preserving copies of pages from the past Web (Mementos) going back as far as early 1996. Its holdings have grown steadily since, and it hosts more than 475 billion URIs as of September 2019. However, the landscape of web

S. M. Jones (✉) · M. Klein
Los Alamos National Laboratory, Los Alamos, NM, USA
e-mail: smjones@lanl.gov; mklein@lanl.gov

H. Van de Sompel
Data Archiving and Networked Services (DANS), Den Haag, The Netherlands
e-mail: herbert.van.de.sompel@dans.knaw.nl

M. L. Nelson · M. C. Weigle
Old Dominion University, Norfolk, VA, USA
e-mail: mln@cs.odu.edu; mweigle@cs.odu.edu

D. Gomes et al. (eds.), *The Past Web*,
https://doi.org/10.1007/978-3-030-63291-5_9

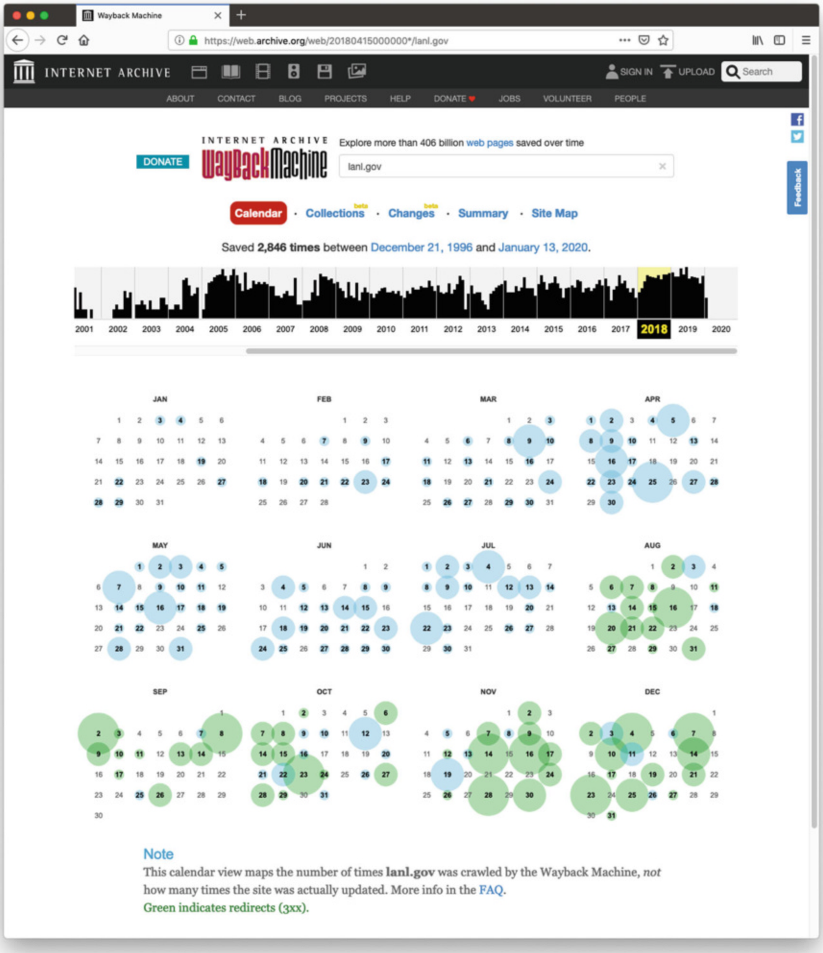
archiving has changed significantly over the last two decades. Today we can freely access Mementos from more than 20 web archives around the world, operated by for-profit and nonprofit organisations, national libraries and academic institutions, as well as individuals. The resulting diversity improves the odds of the survival of archived records but also requires technical standards to ensure interoperability between archival systems. To date, the Memento Protocol and the WARC file format are the main enablers of interoperability between web archives. Before the Memento Protocol (Van de Sompel et al. 2013; Nelson and Van de Sompel 2019), there was no standard way to answer some basic questions about archived web resources, such as:

- Is this web resource, identified by a specific Uniform Resource Identifier (URI), an archived web resource (Memento)?
- From what URI was this Memento created?
- When was this Memento captured?
- Given the URI for a web resource and a datetime, how do we find a Memento from that datetime?
- Given the URI for a web resource, how do we locate all of the Mementos for that URI at an archive?
- What about across all archives?

1.1 Answering Questions Without the Memento Protocol

This section outlines several scenarios that illustrate the difficulties in answering the above questions without the Memento Protocol. Consider the process that a human user would follow to acquire a Memento from a web archive running OpenWayback (Ko 2019), the most common web archival replay software. The web archive presents the user with a search page. The users enter their desired URI, and the system delivers them a page consisting of a series of calendars containing highlighted dates, as seen in Fig. 1. These dates represent the capture datetimes for that URI (Gomes and Costa 2014). To view the Memento, the user clicks on one of these capture datetimes. This tedious process requires that the user manually click through calendars of dates to find the datetime of interest before finding their Memento.

Behind the scenes, the Memento is stored as a record in a Web ARChive (WARC) file (ISO 2009), the standard web archive file format. Web archives routinely generate one or more Crawl Index (CDX) (International Internet Preservation Coalition 2006) files to help the playback system locate Mementos. When the user clicks on a date, the archive's playback system searches CDX files for the corresponding WARC file and the position of the Memento record in that file. Based on this information, the system then locates the WARC record and presents the Memento to the user, that is, the content of a page from the past as it was in the datetime when it was captured. This process provides data for the user interface, but how would a machine client acquire a Memento?



Explore more than 406 billion web pages saved over time

lanl.gov

Saved 2,846 times between December 21, 1996 and January 13, 2020.

Note

This calendar view maps the number of times **lanl.gov** was crawled by the Wayback Machine, *not* how many times the site was actually updated. More info in the FAQ.

Green indicates redirects (3xx).

Fig. 1 Links to Mementos of http://www.lanl.gov for 2018 exist in the calendar interface employed by the Internet Archive

Given a URI, how does a software executed by a client machine know that the resource is a Memento? How does it determine the capture datetime of this resource given that the capture datetime is not necessarily the same as the last modification date (Nelson 2010), which datetime is available in the resource's archived HTTP header? How does the client machine determine the URI that the web archive captured? The home page of the Internet Archive's Wayback Machine, https://web.archive.org, is not a Memento, nor are the various images, scripts and stylesheets used to build the web archive's page banners (e.g. https://web.archive.org/_static/js/toolbar.js). However, the page at https://web.archive.org/web/20191210234446/

https://www.lanl.gov/ is a Memento. A client would need to identify the pattern applied to generate this URI to know that it found a Memento. Once a client can parse this pattern, it can screen for Memento URIs, identify the capture datetime of the Memento (i.e. 20191210234446) and the URI that the archive had captured (i.e. https://www.lanl.gov/). However, if we consider a different archive that does not use such URI patterns, such as Perma.cc, then this approach is not feasible. Perma.cc stores a Memento for that same URI from that same date at https://perma.cc/Q938-LPMG. How does a client know that this is a Memento and https://perma.cc/about is not? How does a client find the captured URI or its datetime?

If the client software knows the captured URI and the desired datetime, how does it find the temporally closest Memento in an archive for that datetime? It could employ web scraping to mimic the tedious process executed by the human at the beginning of this section. Then it could create a list of all captures and search the list for the closest Memento. Because web archives have different user interfaces, the developer of the client software would need to update this scraper for all changes to any archive. Web archives may not share CDX files for technical or legal reasons, but if the archive does share its CDX files, then a machine client could parse each CDX file to find the location of the Memento at the archive. CDX files list captured URIs and not the URIs of Mementos. So even with the CDX record, how does a client software discover the URI of the Memento? If a web archive uses a known URI pattern to reference its preserved contents, then a client might be able to construct the Memento URI, but how does it work when an archive, such as Perma.cc, did not base the URI structure on the CDX data? What if the client software wants to search across archives? A client would need to be able to perform the aforementioned URI parsing, CDX parsing and screen scraping for several archives and combine the results. This approach would be time consuming and expensive to maintain at large scale.

What if web archives could keep their own customised user interfaces and URI structures, control access to their CDX files, but still be able to answer these questions? What if the machine client could obtain responses effectively by simply executing HTTP requests? The Memento Protocol provides a standard for answering questions about individual Mementos and the means for interoperability across archives.

1.2 Answering These Questions with the Memento Protocol

In order to understand how Memento answers these questions, we first describe how the Memento Protocol works. To provide a standard and consistent method of access across web archives, the Memento Protocol introduces the following components:

- **Original Resource**: A web resource as it exists or used to exist on the Web. A **URI-R** (e.g., http://lanl.gov) denotes the URI of an Original Resource.

- **Memento**: A capture of an Original Resource from a specific point in time. A **URI-M** (e.g., https://webarchive.loc.gov/all/19990125090547/http://lanl.gov/) denotes the URI of a Memento.
- **Memento-Datetime**: The capture datetime (e.g., Mon, 25 Jan 1999 09:05:47 GMT) of the Original Resource for the given Memento.
- **TimeGate**: A web resource that helps clients find the temporally closest Memento for an Original Resource given a specific datetime using datetime negotiation, a process described below. A **URI-G** (e.g., https://webarchive.loc.gov/all/http://lanl.gov/) denotes the URI of a TimeGate.
- **TimeMap**: A resource that contains a list of the URI-Ms for an Original Resource. A **URI-T** (e.g., https://webarchive.loc.gov/all/http://lanl.gov/) denotes the URI of a TimeMap.

The remainder of this section details how these components answer the questions posed in the introduction. Table 1 summarises these questions and the components that support answering them.

The Memento Protocol conforms to REST (Fielding 2000) and thus uses HTTP (Fielding et al. 2014a,b; Fielding and Reschke 2014a,b,c,d) to help clients answer questions about Mementos. Figure 2 displays the HTTP response headers for a single Memento from the United States Library of Congress web archive. The presence of the `Memento-Datetime` header indicates that this resource is a

Table 1 The questions addressed by the Memento Protocol

Question	Memento feature addressing question
Is this resource a Memento?	`Memento-Datetime` header in the Memento's HTTP response header
From what URI was this Memento created?	`original` relation in the Memento's HTTP `Link` response header
When was this Memento captured?	`Memento-Datetime` in the Memento's HTTP response header
How do we locate all of the Mementos for a URI at an archive?	TimeMap
How do we locate all of the Mementos for a URI across all archives?	TimeMap aggregator (see Sect. 2.2)
How do we find an Original Resource's TimeGate?	`timegate` relation in the Original Resource's HTTP `Link` response header (if available)
How do we find the temporally closest Memento for a given datetime and URI?	TimeGate
How do we find the temporally closest Memento across all web archives for a given datetime and URI?	TimeGate aggregator (see Sect. 2.2)
How do we find a Memento's TimeGate?	`timegate` relation in the Memento's HTTP `Link` response header
How do we find a Memento's TimeMap?	`timemap` relation in the Memento's HTTP `Link` response header

```
HTTP/1.1 200 OK
Date: Thu, 19 Dec 2019 20:38:53 GMT
Server: Apache-Coyote/1.1
Memento-Datetime: Mon, 25 Jan 1999 09:05:47 GMT
Link: <http://lanl.gov/>
    ; rel="original",
  <https://webarchive.loc.gov/all/timemap/link/http://lanl.gov/>
    ; rel="timemap"; type="application/link-format",
  <https://webarchive.loc.gov/all/http://lanl.gov/>
    ; rel="timegate",
  <https://webarchive.loc.gov/all/19961221031231/http://lanl.gov/>
    ; rel="first memento"; datetime="Sat, 21 Dec 1996 03:12:31 GMT",
  <https://webarchive.loc.gov/all/19990117083819/http://lanl.gov/>
    ; rel="prev memento"; datetime="Sun, 17 Jan 1999 08:38:19 GMT",
  <https://webarchive.loc.gov/all/19990125090547/http://lanl.gov/>
    ; rel="memento"; datetime="Mon, 25 Jan 1999 09:05:47 GMT",
  <https://webarchive.loc.gov/all/19990208005037/http://lanl.gov/>
    ; rel="next memento"; datetime="Mon, 08 Feb 1999 00:50:37 GMT",
  <https://webarchive.loc.gov/all/20180305142008/http://lanl.gov/>
    ; rel="last memento"; datetime="Mon, 05 Mar 2018 14:20:08 GMT"
Content-Type: text/html;charset=utf-8
Transfer-Encoding: chunked
Set-Cookie: JSESSIONID=6C40C0CA8C02BF49519DDA6A551632DA; Path=/
```

Fig. 2 The HTTP response headers for a Memento provided by the United States Library of Congress at URI-M https://webarchive.loc.gov/all/19990125090547/http://lanl.gov/. Link headers have been reformatted to ease reading. We have omitted the content of the Memento for brevity

Memento and conveys the capture datetime for this Memento. The `Link` header contains several relations that help clients find resources to answer other questions about this Memento and its Original Resource. The `original` relation tells the client that this Memento's Original Resource (URI-R) is http://lanl.gov/. The `timemap` and `timegate` relations help clients find this Memento's TimeMap (URI-T) and TimeGate (URI-G), respectively. The HTTP response headers shown in Fig. 2 allow a client to determine that this is a Memento, that the archive captured it on 25 January 1999, at 09:05:47 GMT, that its Original Resource URI-R is http://lanl.gov/, and how to find more Mementos for this resource via TimeGates and TimeMaps. Figure 2 also includes some optional relations. Using these relations, a client can locate the `next` and `previous` Mementos relative to this one as well as the `first` and `last` Mementos for this Original Resource (URI-R), known to the archive.

TimeMaps provide a full list of all Mementos for an Original Resource, but what if we know the datetime and the Original Resource and want the temporally closest Memento? TimeGates provide this functionality. Figure 3 demonstrates the process of datetime negotiation. Memento-compliant Original Resources make their TimeGate discoverable, and thus, a Memento-aware HTTP client (e.g. a

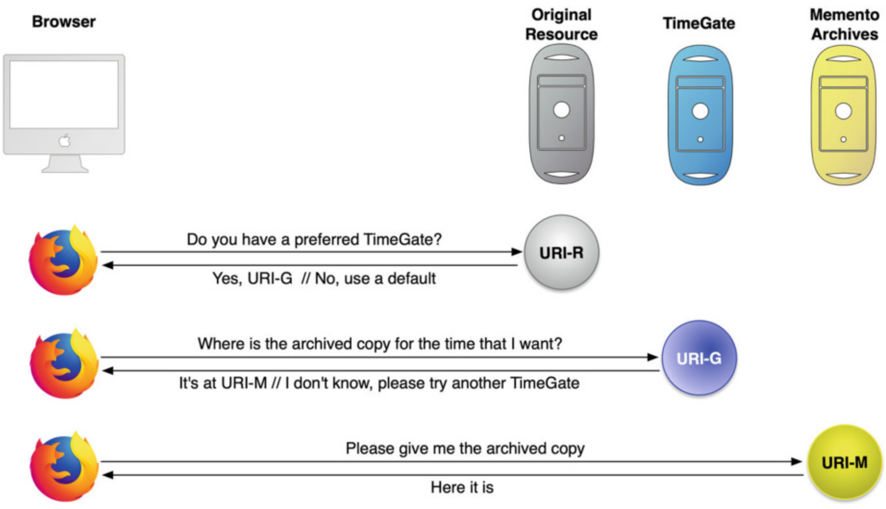

Browser
Original Resource
TimeGate
Memento Archives
Do you have a preferred TimeGate?
Yes, URI-G // No, use a default
URI-R
Where is the archived copy for the time that I want?
It's at URI-M // I don't know, please try another TimeGate
URI-G
Please give me the archived copy
Here it is
URI-M

Fig. 3 Datetime negotiation with a Memento TimeGate

browser) first asks the Original Resource for its preferred TimeGate. If the Original Resource does not express a preference, then the client will use a default TimeGate instead. The client then uses the `accept-datetime` request header to supply the desired datetime to the TimeGate. The TimeGate then redirects the client to the Memento (URI-M) that best matches these criteria. Using `accept-datetime` and TimeGates, a Memento client holds the desired datetime steady, allowing the user to seamlessly browse the Web as if it were a date in the past.

2 Architecture of the Memento Protocol

2.1 How Memento Enables Interoperability

The `Link` relations in Memento provide most of its capability. In Fig. 2, the `Link` header provides relations with URIs to other resources related to this Memento. This same method is applied with Original Resources, TimeGates and Mementos to allow software clients to “follow their nose” and find other resources. Figure 4 demonstrates how a client can discover each resource from another via their `Link` headers. From a Memento or TimeGate, a client can find the TimeMap and gain access to the other Mementos for the same Original Resource. From a Memento, a client can find the TimeGate with which to perform datetime negotiation to gain access to other Mementos for the same Original Resource. From a Memento or TimeGate, a client can find the URI-R of the Original Resource, helping them visit the present state of the resource.

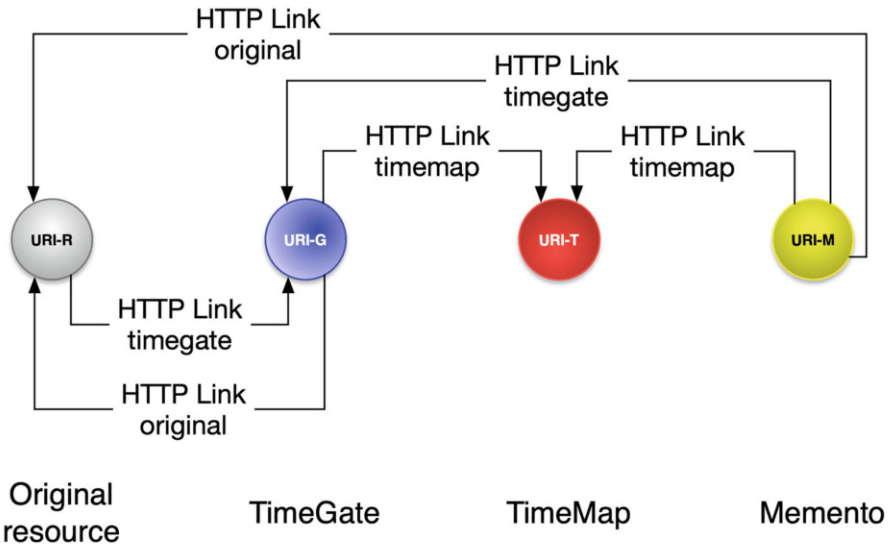


Fig. 4 Workflow of Memento protocol. The arrows show how a client can find additional Memento components through the URIs in Link relations

By using relations the same way, the content of a TimeMap contains links to all other resources. Figure 5 displays an example of a TimeMap in link format (Shelby 2012). The TimeMap informs clients of the Original Resource (URI-R) via the `original` relation. The `self` relation confirms the URI-T of this TimeMap and can optionally provide its date range via `from` and `until` attributes. The `timegate` relation delivers this Memento's corresponding TimeGate URI-G. Each subsequent entry has a relation containing the type `memento`. These entries list each Memento (URI-M) and its Memento-Datetime. The `first` and `last` relations allow a client to easily reach the first and last Mementos, respectively.

2.2 *Memento-Compliant Infrastructure and Standardised Access*

The Wayback Machine hosted by the Internet Archive is Memento-compliant, but the protocol is also supported across all popular web archiving platforms such as OpenWayback or pywb (Kreymer 2019). More than 20 web archives support Memento,[1] including Arquivo.pt (Melo et al. 2016) and archive.today (Nelson 2013). The Memento Protocol does not require that the URIs listed in a TimeMap

[1] http://mementoweb.org/depot/.

```
<http://www.lanl.gov/>; rel="original",
<https://wayback.archive-it.org/all/timemap/link/http://www.lanl.gov/>
  ; rel="self"; type="application/link-format"
  ; from="Fri, 17 Feb 2006 02:53:16 GMT"
  ; until="Tue, 17 Dec 2019 08:38:43 GMT",
<https://wayback.archive-it.org/all/http://www.lanl.gov/>
  ; rel="timegate",
<https://wayback.archive-it.org/all/20060217025316/http://www.lanl.gov/>
  ; rel="first memento"; datetime="Fri, 17 Feb 2006 02:53:16 GMT",
<https://wayback.archive-it.org/all/20060518103556/http://www.lanl.gov/>
  ; rel="memento"; datetime="Thu, 18 May 2006 10:35:56 GMT",
<https://wayback.archive-it.org/all/20060617084403/http://www.lanl.gov/>
  ; rel="memento"; datetime="Sat, 17 Jun 2006 08:44:03 GMT",
<https://wayback.archive-it.org/all/20060617090337/http://www.lanl.gov/>
  ; rel="memento"; datetime="Sat, 17 Jun 2006 09:03:37 GMT",
<https://wayback.archive-it.org/all/20060705231634/http://www.lanl.gov/>
  ; rel="memento"; datetime="Wed, 05 Jul 2006 23:16:34 GMT",
<https://wayback.archive-it.org/all/20060717071357/http://www.lanl.gov/>
  ; rel="memento"; datetime="Mon, 17 Jul 2006 07:13:57 GMT",
<https://wayback.archive-it.org/all/20060717073942/http://www.lanl.gov/>
  ; rel="memento"; datetime="Mon, 17 Jul 2006 07:39:42 GMT",

  ... truncated for brevity ...

<https://wayback.archive-it.org/all/20191217083843/https://www.lanl.gov/>
  ; rel="last memento"; datetime="Tue, 17 Dec 2019 08:38:43 GMT"
```

Fig. 5 The first 10 lines and last line of a TimeMap provided by Archive-It at URI-T https://wayback.archive-it.org/all/timemap/link/http://www.lanl.gov/. Spaces and line breaks have been inserted for readability

come from the same archive. A client can also produce a combined TimeMap containing the results of a query across multiple archives. For multiple archives supporting the Memento Protocol, a client can request each TimeMap in turn and produce a new TimeMap listing all Mementos across those web archives for a single Original Resource. Similarly, a client can consult TimeGates from multiple archives to simultaneously find the temporally closest Memento for the desired datetime. *Aggregators* provide this functionality across web archives, and they do so by supporting the very concepts of the Memento Protocol. In 2013, AlSum et al. (2014) demonstrated that the overlap between the Internet Archive and other web archives is high. By making use of the content across archives, aggregators ensure that clients can discover content even if one or more web archives is unavailable. A client processes an aggregated TimeMap like any other TimeMap, and a software client performs datetime negotiation with a TimeGate aggregator like with any other TimeGate.

The Prototyping Team in the Research Library at Los Alamos National Laboratory (LANL) developed the first Memento aggregator that allows users to find Mementos from not just a single archive, but across archives. For example, if we want a Memento for http://www.cs.odu.edu from 24 April 2010, a TimeGate aggregator would help us to find the temporally closest Memento across more than 20 Memento-compliant archives. The LANL Memento Aggregator has become a popular resource, receiving roughly a million requests per month (Klein et al. 2019a,b). The TimeTravel service at LANL[2] provides this aggregation capability for the public. LANL also offers additional APIs[3] that provide services built upon this aggregator, such as the Time Travel Reconstruct service that reconstructs a Memento out of content from multiple web archives (Costa et al. 2017). If users want to run their own aggregator, they can instal MemGator (Alam and Nelson 2016), developed by Old Dominion University. Research by Kelly et al. extends aggregators to query both public and private web archives (Kelly et al. 2018), allowing users to seamlessly transition between archives at different levels of access. Without the standardised interfaces offered by the Memento Protocol, such aggregators would need to apply an assortment of hacks specific to each web archive, like those mentioned in Sect. 1.1.

Any system that provides access to versioned resources can support the Memento Protocol. MediaWiki is the platform that supports Wikipedia. The Memento MediaWiki extension (Jones et al. 2014) provides a complete implementation of the Memento Protocol for MediaWiki. We have demonstrated how the Memento Protocol allows a user to temporally navigate across web archives and Wikipedia.[4] We have continually engaged with the MediaWiki community[5] in an effort to add Memento support to Wikipedia,[6] in order to provide this time travel capability could become available to its users and machines alike, unfortunately to no avail to date. Likewise, Memento for Wordpress (Welsh 2019) provides standardised access to different versions of blog posts. Much like the transactional archives provided by tools like SiteStory (Brunelle et al. 2013), these systems are web archives in their own right. When we begin to consider the implications of seamless temporal browsing between web archives and other resource versioning systems, we begin to see the power and capability that Memento delivers for all kinds of research and analysis.

3 Tools That Leverage the Memento Protocol

The interoperability specified by Memento has led to a rich ecosystem of tools to engage with web archives. Prior to Memento, tools were often tied to a specific web archive because the cost was too high to engineer solutions that collected data across archives.

[2] http://timetravel.mementoweb.org/.

[3] http://timetravel.mementoweb.org/guide/api/.

[4] https://www.youtube.com/watch?v=WtZHKeFwjzk.

[5] https://phabricator.wikimedia.org/T164654.

[6] https://en.wikipedia.org/wiki/Wikipedia:Requests_for_comment/Memento.

3.1 *Browsing the Past Web*

The Prototyping Team in the Research Library at LANL developed browser extensions so users could browse the Web as if it were a date in the past. The Memento Time Travel extensions for Chrome[7] or Firefox[8] allow users to set a datetime in the past. From there, they can right-click on any page or link and visit a Memento of the selected resource at their chosen datetime. Figure 6 demonstrates how the extension makes its decisions about URIs. For every URI the browser visits, the extension tries to determine what kind of resource it has encountered: an Original Resource, a TimeGate or a Memento. This knowledge helps the extension determine its next step in terms of datetime negotiation, as depicted in Fig. 3. Kelly et al. (2014) developed Mink, a Google Chrome extension, that does not use datetime negotiation but helps users find other versions of their current URI via TimeMaps. Mink also allows them to save a page they are visiting to multiple web archives. These extensions offload decisions about discovering Mementos to the Memento Protocol infrastructure at web archives. In the past, such tools would have been far more complex, requiring special logic to handle different archives, assuming the tool could find all of the information needed for this kind of exploration.

Developers have improved the existing platforms by adding links to Memento aggregators. Warclight (Ruest et al. 2019) provides an interface for users to explore

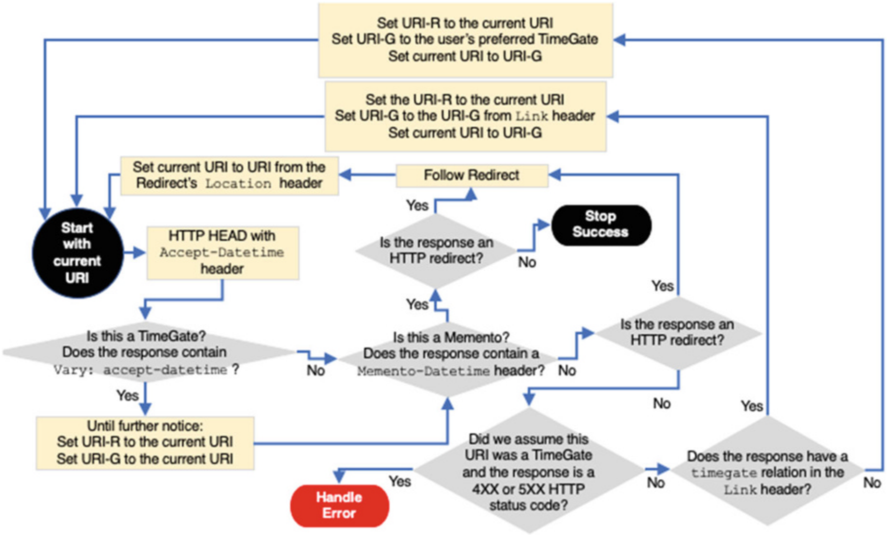

Set URI-R to the current URI
Set URI-G to the user's preferred TimeGate
Set current URI to URI-G
Set the URI-R to the current URI
Set URI-G to the URI-G from Link header
Set current URI to URI-G
Set current URI to URI from the Redirect's Location header
Follow Redirect
Yes
Stop Success
Start with current URI
HTTP HEAD with Accept-Datetime header
Is the response an HTTP redirect?
No
Yes
Yes
Is the response an HTTP redirect?
Is this a TimeGate? Does the response contain Vary: accept-datetime ?
No
Is this a Memento? Does the response contain a Memento-Datetime header?
No
Yes
No
Yes
Until further notice: Set URI-R to the current URI Set URI-G to the current URI
Did we assume this URI was a TimeGate and the response is a 4XX or 5XX HTTP status code?
Yes
Handle Error
No
Does the response have a timegate relation in the Link header?
No

Fig. 6 The algorithm employed by the Memento browser extensions for datetime negotiation

[7] http://bit.ly/memento-for-chrome.

[8] https://addons.mozilla.org/en-US/firefox/addon/memento-timetravel/.

WARC files via facets and other search engine features. Because WARC files contain URIs of the Original Resources (URI-Rs), Warclight can query the LANL Memento aggregator to help users find other Mementos for an indexed resource. The New York Art Resources Consortium (NYARC) started a web archiving initiative for online art and art resources, creating 10 collections in the process. NYARC (Duncan 2017) integrates the LANL Time Travel service into their search results. This integration allows users to view the content supplied by NYARC and the surrounding context provided by other web archives. Arquivo.pt links to LANL's Time Travel Reconstruct API (Arquivo.pt 2016), augmenting the holdings of one archive with others. In the past, such improvements providing easy access to a multitude of web archives did not exist.

3.2 Summarising Web Archive Collections

TimeMaps provide us with a roadmap of observations that we can use to summarise and visualise web resources over time. AlNoamany et al. (2016), and later Jones et al. (2018b), incorporated TimeMaps into algorithms that detect off-topic content in web archive collections, allowing users to exclude these Mementos from summarisation efforts. AlSum and Nelson (2014) developed solutions for visualising Memento content across a TimeMap via browser thumbnails (screenshots). Thumbnails must be stored, and the authors' goal was to save disk space by only generating thumbnails when Memento content was significantly different. TMVis (Weigle 2017) is a tool that implements this work. *What Did It Look Like?* (Nwala 2015) uses a Memento aggregator to poll multiple web archives to generate animated thumbnails demonstrating the change of a webpage over time. The consistent interface provided by Memento allowed all of these researchers to consider the TimeMap a building block to an algorithm that applies to any web archive rather than a specific one.

The existing webpage summarisation tools, such as Embed.ly[9] and microlink,[10] misattribute a Memento to the wrong source and fail to separate web archive branding from Memento content, so Jones et al. (2019) created MementoEmbed. MementoEmbed makes use of Memento headers and TimeMaps to find information about Mementos and provides added value over other services. MementoEmbed summaries also link to the LANL Memento Aggregator to provide access to other Mementos for the same Original Resource. Users can aggregate such summaries for social media storytelling in order to summarise entire web archive collections. AlNoamany et al. (2017) provided an algorithm that makes use of TimeMaps for selecting the best Mementos representing a collection so users can then visualise them via this storytelling method. Because of the Memento Protocol, researchers

[9] https://embed.ly/.

[10] https://microlink.io/.

can focus on the problems of visualisation and summarisation rather than the problems of integrating many disparate components from different web archives.

3.3 Locating and Dating Content

Researchers have developed tools to help others locate lost content. Klein et al. (2011) developed Synchronicity, a Firefox browser extension, that uses TimeMaps to help users discover missing content. Klein and Nelson (2014) also employed TimeMaps to discover lexical signatures for archived pages in an attempt to locate their content on the live Web. The Tempas system, by Holzmann and Anand (2016), uses Delicious tags to help users find Original Resources and then employs TimeGates to help them discover past versions of that content. SalahEldeen and Nelson (2013a) created Carbon Date, which queries aggregated TimeMaps and social media sources to help users estimate when a web resource first came into existence. To combat link rot and content drift, Robust Links (Klein and Van de Sompel 2015) combines HTML link decoration, JavaScript and Memento aggregator queries to provide long-lasting links to the Original Resource and the Mementos thereof, thus helping readers find the Memento of a referenced web resource that is temporally closest to the time the resource was referenced. Without the Memento Protocol, these tools could not benefit from the coverage supplied by multiple archives.

4 Research Enabled by the Memento Protocol

Web archives offer unparalleled research opportunities with data from the past. Sociologists, such as Curty and Zhang (2011), have conducted longitudinal studies with Mementos that were crawled incidentally by the Internet Archive. Journalists such as Hafner and Palmer (2017) have exposed questionable practices by studying the Mementos of businesses' home pages. Milligan (2019) has noted that web archive data is key to historians studying our contemporary times, allowing them to determine not only what social movements existed, but also how people of our era lived their lives. Such studies require that a researcher painstakingly acquires Mementos from web archives before manual review. A familiar aphorism is that a researcher spends 80–90% of their time acquiring the data needed for their work. The Memento Protocol reduces this time by helping researchers find the Mementos that help support their analysis.

In this section, we highlight work that has applied Memento to different problems. Many of the following studies employed TimeMaps and TimeGates, but some also used tools such as Carbon Date and Memento aggregators that developers had built upon the Memento infrastructure.

4.1 Analysis of Social Media

Central to the social media studies in this section were two capabilities. Some needed to find all Mementos for a captured URI. Others needed to find the temporally closest Memento for a URI at a given datetime across all web archives. Without the Memento Protocol, the researchers would have had to request and process CDX files from multiple web archives, and CDX files are often not available for technical or legal reasons. If these files were available, the transfer, storage, update and analysis of CDX files would then be a new problem that researchers would need to solve before engaging in this work. Alternatively, they could have manually visited the multiple web archives and recorded their results, something that would likely have taken years to complete.

Social media allows users to share links to resources. In some cases, web archives capture a given post, but the resource referenced in the post is no longer available. In other cases, the referenced resource no longer contains the content mentioned in the post. These disconnects were discussed by Klein and Nelson (2011) in their analysis of the TimeMaps of pages tagged by Delicious users. SalahEldeen and Nelson (2012) queried the LANL Memento aggregator to explore the nature of this loss of context concerning the 2011 Egyptian Revolution, an event for which social media played a key role. In the subsequent work (SalahEldeen and Nelson 2013b), the authors highlighted a new issue. If a user creates a social media post about a news event and links to the front page of a news service, then it is likely that a reader at a later date will not see the version of the news service that the user intended. However, if the user creates a post and links to a specific article, then their intention may be more clear. The authors dubbed this new problem *temporal intention*. Thanks to Memento aggregators, they were able to determine if an archive had created a Memento of the referenced resource close to the time of the social media post. They were then able to apply these Mementos in a user study to better understand the context of the post and whether the current linked content still reflected the original poster's intention.

How similar are social media collections about events to web archive collections on the same topic? Nwala et al. (2018a) analysed the text from webpages shared on Twitter, Storify, Reddit and the web archiving platform Archive-It. Their results show that web archive collections about events are very similar to social media collections. With this result, they defended their position that archivists can mine social media collections to bootstrap web archive collections about events. More popular resources tend to be archived more often than less popular ones (Ainsworth et al. 2011); thus, we can use the archived status of a resource as a proxy for its popularity. As part of their evaluation, Nwala et al. queried MemGator to determine if one or more web archives had captured an Original Resource referenced from social media. Without the capability of aggregators, they would not have been able to leverage this metric in their analysis.

Aside from analysing the behaviour of social media users, what can historians learn from social media in a historiographical fashion? Helmond and van der

Vlist (2019) brought MemGator to bear to analyse not only the posts of social media users but also the social media platforms as entities themselves. By utilising Memento, they quantified how well the user documentation and business plans of these platforms are archived. This insight will allow future historians to connect historical events to changes in social media platforms and better understand changes in human behaviour over time.

4.2 Age and Availability of Resources

Memento has also played a pivotal role in helping researchers estimate archiving rates for various types of resources. Ainsworth et al. (2011) investigated how much of the Web was archived and estimated that 35–90% of the existing web resources have at least one Memento. Alkwai et al. estimated the archive coverage of Arabic websites (Alkwai et al. 2015) and later conducted an additional study (Alkwai et al. 2017) to compare the archiving rates of English-, Arabic-, Danish- and Korean-language webpages. Alkwai showed that English has a higher archiving rate than Arabic, which in turn has a higher archiving rate than Danish or Korean. Nwala et al. (2017) established that local news sources have lower archiving rates than non-local sources. Bicho and Gomes (2016) evaluated how well research and development websites were archived, and used this evaluation to construct heuristics and algorithms for proactively identifying and preserving them. Because they used Memento aggregators, each of these studies achieved higher coverage of web history in far less time than if they had developed bespoke cross-archive solutions.

When an event takes place, how many resources that pertain to it can we archive after the fact? Ben-David (2019) analysed web resources that were present during the 2014 Gaza War. While building a web archive for the 2014 event in 2018, she observed that 38.16% of the resources were still available on the Web, 40.63% were inaccessible and 21.21% were redirected or moved. Thus, the Memento infrastructure allowed her to generate the evidence needed to defend the idea that incidental web archiving is insufficient for documenting events. Ben-David's study relied upon Carbon Date to establish which content authors had published during the 2014 Gaza War. Carbon Date relies upon the Memento infrastructure to supply the earliest Memento across multiple web archives. She also gathered Mementos via MemGator for analysis. Each of these tools saved the author time and effort so that she could focus on her research questions rather than the complexity of solving interoperability issues between archives.

Can we build event collections from web archives? Even if resources about events remain on the live Web, Nwala et al. (2018b) detailed how they become more challenging to discover via search engine results as we get farther from the event. Topical focused crawling (Chakrabarti et al. 1999) provides resources whose terms closely match the terms of a desired topic, such as an event, and these crawlers stop when the matching score for new content is too low. Building on the work by Gossen et al. (2017, 2018), Klein et al. (2018) designed a framework to build collections about unexpected events (e.g. shootings) by employing focused crawling of web archives. Focused crawling of web archives is different from focused crawling of the live Web because Mementos have a temporal dimension to consider; thus, the authors' crawler had to score each resource's relevance in terms of similarity to the desired topic and distance from the datetime of the event. The authors considered the canonical Wikipedia article about an event to be a good source of authoritative information for their focused crawls. Because Wikipedia articles are created in response to events and then updated by many contributors thereafter, it takes some time for the article to stabilise into an adequate description of the event. The authors used the article revision corresponding to the datetime of this change point as the source of terms relevant to the event. They also recorded the Original Resource (URI-R) references in this revision as the seeds for their focused crawl. The authors employed Memento TimeGates with these Original Resources (URI-Rs) and the datetime of the event to find candidate Mementos for their event collection. To ensure that they only included resources created after the event, the authors also acquired publication dates from Original Resources (URI-Rs) or HTML metadata. They estimated publication dates with Carbon Date if those methods failed. They found that collections built from the live Web for recent events are more relevant than those built from web archives, but events from the distant past produced more relevant documents using their archive-based method. Without the Memento Protocol, such crawling would be far more costly to implement for a single archive. With the Memento Protocol, Klein et al. were able to create collections from Mementos spanning many web archives.

4.3 Analysing and Addressing Lost Context

Context is key to understanding many documents. Thus, in many cases, merely preserving a single document is insufficient to truly understand the author's intention. Various studies have tried to explore the availability of documents that supply this context.

Because references from scholarly papers provide evidence and justification for arguments, the decay of scholarly references has been of particular concern. Even though there are systems such as LOCKSS (Reich and Rosenthal 2001) for preserving referenced papers, no concerted long-term effort exists for preserving referenced web resources. Sanderson et al. (2011) observed that 45% of the web resource references from papers on arxiv.org still exist but are not archived. Klein et al. (2014) expanded the dataset to include data from Elsevier and PubMed. They observed the rate of decay varied depending on the paper's year of publication, getting worse with the paper's age. Jones et al. (2016) discerned from the same dataset that not only were many referenced resources missing, but, for those references for which the Original Resource still existed on the live Web, 80% had content that had drifted when compared to the time the paper referenced it. The authors demonstrated that the older the publication date of the paper, the greater the number of references that are no longer accessible or have highly drifted from their original content. Zhou et al. (2015) analysed how incidental web archiving misses many of these resources and proposed machine learning techniques to predict which resources might decay faster so that crawlers would archive them sooner. If these studies had queried a single web archive, then the representativeness of their results would have been more easily challenged—however, all employed aggregated TimeMaps and Mementos to get a better view of the archived resources across the world. Memento also provides a solution to the problem identified by the authors. Tools such as the browser extensions or Robust Links can employ aggregated TimeGates to help connect users to an appropriate Memento for a referenced resource (Sanderson and Van de Sompel 2010; Van de Sompel and Davis 2015) and thus restore this context.

Annotations developed via tools such as Hypothes.is provide similar context to documents. If a future researcher has access to the document, but not the annotations, crucial insight is lost. But having access to an annotation but not the annotated resource is probably even worse. Sanderson and Van de Sompel (2010) formulated a theoretical solution and demonstrated how the Memento Protocol can be leveraged both to attach an annotation to the appropriate temporal version of an annotated resource, assuming it was archived, and to find temporal annotations for a version of a resource given a Memento for it. Via the notion of Time Selectors, their insights eventually found their way into the W3C Web Annotation Data Model (Sanderson et al. 2017).

Aturban et al. (2015) analysed the issue of orphaned annotations. By analysing TimeMaps, they determined that 27% of the annotations were no longer available, and 61% were in danger of becoming disconnected from their referenced document. Much like the other lost context studies mentioned above, Aturban's study serves as a call to action to archive annotations and annotated resources so that Sanderson and Van de Sompel's solution can effectively solve this problem.

4.4 Improving Web Archiving

The quality of Mementos can vary depending on a variety of factors. Sometimes webpages are missing images, stylesheets and other embedded resources because the web archive was unable to capture them. Brunelle et al. (2015) developed the metric of *Memento damage* for measuring this notion of archival quality, which compares well with later studies (Kiesel et al. 2018) that employed neural networks for this purpose. They found that 54% of all Mementos were missing at least one embedded resource. Brunelle et al. (2016) were able to identify how much JavaScript contributed to this problem. They later developed a better crawling technique (Brunelle et al. 2017), which captured 15 times more embedded resources than before. Alam et al. (2017) further improved upon this by combining browser technologies and the Memento Aggregator to reconstruct pages from embedded resources spread across multiple web archives.

Missing resources are one thing, but what if the embedded resource in a Memento is the wrong one? Ainsworth et al. (2015) demonstrated that sometimes web archive playback engines choose versions of images from the future or the distant past. The wrong embedded resources can profoundly change the meaning of a page and reproduce a webpage that never existed. They estimated that web archives accurately reproduced only 20% of the Mementos. They introduced the concept of the *composite Memento*, a resource that includes not only the page that the web archive had captured but all embedded resources as well. Aturban et al. (2017) analysed the effectiveness of applying cryptographic hashes to detect when Mementos do not represent the pages from which they were captured. To alleviate issues identified by Ainsworth, Aturban argued that archives should compute such hashes on the composite Memento rather than just the individual resources. They also cited work by Brunelle and Nelson (2013) that indicates that the Mementos listed in TimeMaps are not always the same, affecting the ability to reliably compute these hashes.

Improving aggregators is an active area of research. Early aggregators tried to combine CDX files from different web archives (Sanderson 2012) in order to provide TimeGate responses from the best web archive. As Memento adoption increased, newer aggregators directly queried the TimeGates or TimeMaps from each web archive, and new approaches to optimise these queries emerged. To help aggregators avoid unnecessary queries, several studies tried to predict which web archives to query for a given Original Resource (URI-R). Alam et al. demonstrated that profiling web archives via domain name and language (AlSum et al. 2014) would allow aggregators to route queries more efficiently. Brunelle et al. evaluated the changes in web archive holdings and proposed caching TimeMaps for 15 days (Brunelle and Nelson 2013) to improve response times. Alam et al. (2016a) further demonstrated improvements to query routing via CDX summarisation. Alam et al. (2016b) later demonstrated that queries could be routed based on the results of textual queries sent to the web archive's search engine. Unfortunately, many archives do not provide access to their CDX files nor have search engine capability, so Bornand et al. (2016) trained binary classifiers using data from the LANL Memento Aggregator cache in order to determine whether or not a query should be sent to a given archive. This approach led to a 42% improvement in the overall response times, but Klein et al. (2019b) noted that classifiers need to be retrained regularly to maintain this level of performance.

To analyse these problems and develop solutions, all of these authors queried publicly available TimeMaps and TimeGates. Before the Memento Protocol, this work would have required close collaboration with web archives, screen scraping or manual work to gather the necessary data.

4.5 Temporal Access to Data

Because Memento provides access to previous resource versions via HTTP, the ability to acquire past web resources is independent of the file format of the content. Early on in the effort to formally specify the Memento Protocol, Van de Sompel et al. described and demonstrated (Van de Sompel et al. 2010) its applicability to interact with versions of Linked Data sets. LANL deployed an archive of various versions of DBpedia[11] that provided subject URI access. The associated current versions of those subject URIs in DBpedia pointed at an associated TimeGate exposed by the archive, as per the Memento Protocol. To illustrate Memento's potential, the authors leveraged this infrastructure to automatically plot a time series analysis of the evolving Gross Domestic Product per capita of various countries by obtaining current and past versions of the DBpedia descriptions for each of these countries, merely using Memento's "follow your nose" approach, and extracting the pertinent attribute from each description. In an attempt to provide access to Linked Data sets in a manner that is more sustainable than full-blown SPARQL endpoints and more expressive than dataset dumps or subject URI access, Taelman et al. (2017) suggested Triple Pattern Fragments (TPF) endpoints, which support *?subject ?predicate ?object* query functionality. Vander Sande et al. (2018) argued that the combination of the binary Header Dictionary Triple (HDT)[12] approach to store Linked Data set and TPF to query them offers a sweet spot between functionality and sustainability for Linked Data archives. They refurbished the LANL DBpedia archive using this approach and extended the Linked Data Fragment server software[13] with Memento capabilities, as such providing support for temporal *?subject ?predicate ?object* queries (Van de Sompel and Vander Sande 2016). Since Verborgh et al. (2016) showed that SPARQL queries can be broken down into consecutive TPF queries, in essence, a client can issue temporal SPARQL queries against Linked Data archives modelled in this manner.

Many others have leveraged the Memento Protocol in Linked Data architectures. Coppens et al. (2011) detailed how multiple projects in Belgium disseminate temporal versions of datasets as Linked Open Data and used Memento TimeGates to provide version access. Mannens et al. (2012) provided TimeGates over aggregated Linked Data to provide provenance for news stories. Meinhardt et al. (2015) developed TailR, a system that archives Linked Data resources so that researchers have access to their dataset history via TimeGates and TimeMaps. Neumaier et al. (2017) republished the dataset descriptions of 261 data portals using the standard Linked Data vocabularies such as DCAT[14] and Schema.org[15] and provided

[11] http://wikidata.dbpedia.org/.

[12] http://www.w3.org/Submission/HDT/.

[13] https://github.com/LinkedDataFragments/Server.js/.

[14] https://www.w3.org/TR/vocab-dcat-2/.

[15] http://schema.org/.

access to different versions of this data via Memento TimeGates. Van de Vyvere et al. (2019) applied Linked Data Fragments and Memento to create a system that allowed users to explore historical automobile traffic data. Fafalios et al. (2017) semantically annotated Mementos that reside in web archives, for example, connecting entities mentioned in archived webpages to their associated DBpedia URI and as such created enhanced search capabilities. When enriching archived webpages with information from DBpedia, they used the Memento Protocol and the LANL DBpedia archive to obtain DBpedia entity descriptions contemporary of the archived webpage. Conceptually, Sanderson and Van de Sompel (2012) noted that Linked Data descriptions heavily rely on links to external resources, which a client software may need to retrieve. When doing so, most likely, the client requires a description of the linked resource that was the current version at the time the description containing the link was published. In cases where versions of Linked Data sets are maintained, clients can use the Memento Protocol to access temporally matching descriptions of linked resources. Powell et al. (2011) mentioned how one could use the Memento Protocol to analyse the evolution of knowledge graphs over time.

4.6 Other Work Using Memento

Can we use Memento to avoid information? Why would we want to do this? Fans of TV shows and sport may not be able to experience an event as it is broadcast and may watch a recorded version afterwards. Jones et al. (2018a) explored the use of the Memento Protocol to avoid spoilers for TV shows. Fans of television shows update wikis very quickly after an episode has aired. Thus, to avoid spoilers, a fan can just visit the wiki page version that existed immediately prior to the first airing of an episode or event. The authors discovered that wikis, having access to all versions, do a better job at helping fans avoid spoilers than the Internet Archive. For this use case, the authors quantified how temporally closest is not always the best TimeGate heuristic to use when we have access to all revisions of a document because it can lead a user to a Memento captured after the spoiler was revealed. Their results have implications for television shows, sporting events and information security. TimeMaps provided the data necessary to analyse the different behaviours between wikis and web archives. Alternatively, the authors could have requested CDX files from the archive. Even if filtered, these files would have likely contained superfluous information not core to their study (e.g. CDX records not containing wiki page content).

Researchers have applied Memento to analyse web author behaviour. Cocciolo (2015) utilised the Memento Protocol to find the Mementos of prominent American webpages. He discovered that the percentage of text in webpages peaked in 2005 and has been declining ever since. These results confirm Cocciolo's anecdotal observation that webpages have been getting shorter, and this work may lead to further analysis to understand the causes of this behaviour. Hashmi et al. (2019)

performed a longitudinal analysis of the evolution of ad-blocking blacklists by using the Memento Protocol to query the Internet Archive at specific time intervals for advertisement services. They then contrasted this with the Mementos of the blacklists that try to keep up with them.

5 Summary

The Memento Protocol provides client software with an ability to seamlessly navigate between the current Web and the archived Web using the omnipresent HTTP protocol. It separates the concept of an Original Resource from a Memento. An Original Resource exists or used to exist on the Web. A Memento is a capture of that resource at a specific point in time. TimeMaps provide lists of all Mementos for an Original Resource. TimeGates allow a client to request a specific Memento given an Original Resource URI and the desired datetime. With these resources in place, the community has constructed aggregators, allowing for queries across web archives. These components have given rise to a new tool ecosystem for web archives.

Researchers have accessed these tools to conduct studies that otherwise would be too costly, or even impossible to conduct, in terms of time and effort. They have been taken advantage of the Memento Protocol to analyse social media, determine the age and availability of resources, address lost context, improve web archives or obtain temporal access to data.

The Memento Protocol is relatively simple and has its roots in HTTP. What if Memento support existed as a native browser feature rather than as an extension? A user could have some indication that they have reached a Memento rather than a live page. The browser could help them browse the Web of the past via datetime negotiation across all web archives and versioning systems. What if all web archives and versioning systems supported the Memento Protocol? At their fingertips, everyone would truly be able to time travel through the past Web.

References

Ainsworth SG, Alsum A, SalahEldeen H, Weigle MC, Nelson ML (2011) How much of the web is archived? In: ACM/IEEE joint conference on digital libraries, pp 133–136. https://doi.org/10.1145/1998076.1998100

Ainsworth SG, Nelson ML, Van de Sompel H (2015) Only one out of five archived web pages existed as presented. In: ACM conference on hypertext and social media, pp 257–266. https://doi.org/10.1145/2700171.2791044

Alam S, Nelson ML (2016) MemGator – A portable concurrent memento aggregator: cross-platform CLI and server binaries in go. In: ACM/IEEE joint conference on digital libraries, pp 243–244. https://doi.org/10.1145/2910896.2925452

Alam S, Nelson ML, Balakireva LL, Shankar H, Rosenthal DSH (2016a) Web archive profiling through CDX summarization. Int J Digital Libraries 17(3):223–238. https://doi.org/10.1007/s00799-016-0184-4

Alam S, Nelson ML, Van de Sompel H, Rosenthal DSH (2016b) Web archive profiling through fulltext search. In: International conference on theory and practice of digital libraries (TPDL), vol 9819, pp 121–132. https://doi.org/10.1007/978-3-319-43997-6_10

Alam S, Kelly M, Weigle MC, Nelson ML (2017) Client-side reconstruction of composite mementos using ServiceWorker. In: ACM/IEEE joint conference on digital libraries, pp 1–4. https://doi.org/10.1109/JCDL.2017.7991579

Alkwai LM, Nelson ML, Weigle MC (2015) How well are arabic websites archived? In: ACM/IEEE joint conference on digital libraries, pp 223–232. https://doi.org/10.1145/2756406.2756912

Alkwai LM, Nelson ML, Weigle MC (2017) Comparing the archival rate of Arabic, English, Danish, and Korean Language web pages. ACM Trans Inf Syst 36(1):1–34. https://doi.org/10.1145/3041656

AlNoamany Y, Weigle MC, Nelson ML (2016) Detecting off-topic pages within TimeMaps in web archives. Int J Digital Libraries 17(3):203–221. https://doi.org/10.1007/s00799-016-0183-5

AlNoamany Y, Weigle MC, Nelson ML (2017) Generating stories from archived collections. In: ACM conference on web science, pp 309–318. https://doi.org/10.1145/3091478.3091508

AlSum A, Nelson ML (2014) Thumbnail summarization techniques for web archives. In: European conference on information retrieval (ECIR), vol 8416, pp 299–310. https://doi.org/10.1007/978-3-319-06028-6_25

AlSum A, Weigle MC, Nelson ML, Van de Sompel H (2014) Profiling web archive coverage for top-level domain and content language. Int J Digital Libraries 14(3–4):149–166. https://doi.org/10.1007/s00799-014-0118-y

Arquivopt (2016) Arquivo.pt – new version. https://sobre.arquivo.pt/en/arquivo-pt-new-version-2/

Aturban M, Nelson ML, Weigle MC (2015) Quantifying orphaned annotations in hypothes.is. In: International conference on theory and practice of digital libraries (TPDL), vol 9316, pp 15–27. https://doi.org/10.1007/978-3-319-24592-8_2

Aturban M, Nelson ML, Weigle MC (2017) Difficulties of timestamping archived web pages. Technical Report. arXiv:1712.03140. http://arxiv.org/abs/1712.03140

Ben-David A (2019) 2014 not found: a cross-platform approach to retrospective web archiving. Internet Histories 3(3–4):316–342. https://doi.org/10.1080/24701475.2019.1654290

Bicho D, Gomes D (2016) Automatic identification and preservation of R&D websites. Technical report, Arquivo.pt - The Portuguese Web Archive. https://sobre.arquivo.pt/wp-content/uploads/automatic-identification-and-preservation-of-r-d.pdf

Bornand NJ, Balakireva L, Van de Sompel H (2016) Routing memento requests using binary classifiers. In: ACM/IEEE joint conference on digital libraries, pp 63–72. https://doi.org/10.1145/2910896.2910899

Brunelle JF, Nelson ML (2013) An evaluation of caching policies for memento TimeMaps. In: ACM/IEEE joint conference on digital libraries, pp 267–276. https://doi.org/10.1145/2467696.2467717

Brunelle JF, Nelson ML, Balakireva L, Sanderson R, Van de Sompel H (2013) Evaluating the SiteStory transactional web archive with the ApacheBench Tool. In: International conference on theory and practice of digital libraries (TPDL), vol 8092, pp 204–215. https://doi.org/10.1007/978-3-642-40501-3_20

Brunelle JF, Kelly M, SalahEldeen H, Weigle MC, Nelson ML (2015) Not all mementos are created equal: measuring the impact of missing resources. Int J Digital Libraries 16(3–4):283–301. https://doi.org/10.1007/s00799-015-0150-6

Brunelle JF, Kelly M, Weigle MC, Nelson ML (2016) The impact of JavaScript on archivability. Int J Digital Libraries 17(2):95–117. https://doi.org/10.1007/s00799-015-0140-8

Brunelle JF, Weigle MC, Nelson ML (2017) Archival crawlers and JavaScript: discover more stuff but crawl more slowly. In: ACM/IEEE joint conference on digital libraries, pp 1–10. https://doi.org/10.1109/JCDL.2017.7991554

Chakrabarti S, Van den Berg M, Dom B (1999) Focused crawling: a new approach to topic-specific web resource discovery. Comput Netw 31(11–16):1623–1640. https://doi.org/10.1016/S1389-1286(99)00052-3

Cocciolo A (2015) The rise and fall of text on the web: a quantitative study of web archives. Inf Res Int Electron J 20(3):1–11. https://eric.ed.gov/?id=EJ1077827

Coppens S, Mannens E, Deursen DV (2011) Publishing provenance information on the web using the memento datetime content negotiation. In: Linked data on the web workshop, pp 1–10. http://events.linkeddata.org/ldow2011/papers/ldow2011-paper02-coppens.pdf

Costa M, Gomes D, Silva MJ (2017) The evolution of web archiving. Int J Digital Libraries 18(3):191–205. https://doi.org/10.1007/s00799-016-0171-9

Curty RG, Zhang P (2011) Social commerce: looking back and forward. Proc Am Soc Inf Sci Technol 48(1):1–10. https://doi.org/10.1002/meet.2011.14504801096

Duncan S (2017) Web archiving at the New York art resources consortium (NYARC): Collaboration to preserve specialist born-digital art resources. In: Digital humanities. opportunities and risks. connecting libraries and research. https://hal.archives-ouvertes.fr/hal-01636124

Fafalios P, Holzmann H, Kasturia V, Nejdl W (2017) Building and querying semantic layers for web archives. In: ACM/IEEE joint conference on digital libraries, pp 1–10. https://doi.org/10.1109/JCDL.2017.7991555

Fielding RT (2000) REST: Architectural styles and the design of network-based software architectures. Doctoral dissertation, University of California, Irvine. https://www.ics.uci.edu/~fielding/pubs/dissertation/top.htm

Fielding R, Reschke J (2014a) RFC 7230 - hypertext transfer protocol (HTTP/1.1): message syntax and routing. https://tools.ietf.org/html/rfc7230

Fielding R, Reschke J (2014b) RFC 7231 - hypertext transfer protocol (HTTP/1.1): semantics and content. https://tools.ietf.org/html/rfc7231

Fielding R, Reschke J (2014c) RFC 7232 - hypertext transfer protocol (HTTP/1.1): conditional requests. https://tools.ietf.org/html/rfc7232

Fielding R, Reschke J (2014d) RFC 7235 - hypertext transfer protocol (HTTP/1.1): authentication. https://tools.ietf.org/html/rfc7235

Fielding R, Lafon Y, Reschke J (2014a) RFC 7233 - hypertext transfer protocol (HTTP/1.1): range requests. https://tools.ietf.org/html/rfc7233

Fielding R, Nottingham M, Reschke J (2014b) RFC 7234 - hypertext transfer protocol (HTTP/1.1): caching. https://tools.ietf.org/html/rfc7234

Gomes D, Costa M (2014) The importance of web archives for humanities. Int J Human Arts Comput 8(1):106–123. https://doi.org/10.3366/ijhac.2014.0122

Gossen G, Demidova E, Risse T (2017) Extracting event-centric document collections from large-scale web archives. In: International conference on theory and practice of digital libraries (TPDL), vol 10450, pp 116–127. https://doi.org/10.1007/978-3-319-67008-9_10

Gossen G, Risse T, Demidova E (2018) Towards extracting event-centric collections from web archives. Int J Digital Libraries. https://doi.org/10.1007/s00799-018-0258-6

Hafner K, Palmer G (2017) Skin cancers rise, along with questionable treatments. The New York Times. https://www.nytimes.com/2017/11/20/health/dermatology-skin-cancer.html

Hashmi SS, Ikram M, Kaafar MA (2019) A longitudinal analysis of online ad-blocking blacklists. Technical Report. arXiv:1906.00166. https://arxiv.org/abs/1906.00166

Helmond A, van der Vlist FN (2019) Social media and platform historiography: challenges and opportunities. J Media History 22(1):6–34. https://www.tmgonline.nl/articles/434/

Holzmann H, Anand A (2016) Tempas: temporal archive search based on tags. In: International world wide web conference, pp 207–210. https://doi.org/10.1145/2872518.2890555

International Internet Preservation Coalition (2006) The CDX file format. https://iipc.github.io/warc-specifications/specifications/cdx-format/cdx-2006/

International Organization for Standardization (ISO) (2009) 28500: 2009 Information and documentation-WARC file format. International Organization for Standardization

Jones SM, Nelson ML, Shankar H, Van de Sompel H (2014) Bringing web time travel to MediaWiki: an assessment of the memento MediaWiki extension. Technical Report. arXiv:1406.3876. http://arxiv.org/abs/1406.3876

Jones SM, Van de Sompel H, Shankar H, Klein M, Tobin R, Grover C (2016) Scholarly context Adrift: three out of four URI references lead to changed content. PLoS One 11(12):e0167475. https://doi.org/10.1371/journal.pone.0167475

Jones SM, Nelson ML, Van de Sompel H (2018a) Avoiding spoilers: wiki time travel with Sheldon Cooper. Int J Digital Libraries 19(1):77–93. https://doi.org/10.1007/s00799-016-0200-8

Jones SM, Weigle MC, Nelson ML (2018b) The off-topic memento toolkit. In: International conference on digital preservation, pp 1–10. https://doi.org/10.17605/OSF.IO/UBW87

Jones SM, Weigle MC, Nelson ML (2019) Social cards probably provide for better understanding of web archive collections. In: ACM international conference on information and knowledge management, pp 2023–2032. https://doi.org/10.1145/3357384.3358039

Kelly M, Nelson ML, Weigle MC (2014) Mink: integrating the live and archived web viewing experience using web browsers and memento. In: ACM/IEEE joint conference on digital libraries, pp 469–470. https://doi.org/10.1109/JCDL.2014.6970229

Kelly M, Nelson ML, Weigle MC (2018) A framework for aggregating private and public web archives. In: ACM/IEEE joint conference on digital libraries, pp 273–282. https://doi.org/10.1145/3197026.3197045

Kiesel J, Kneist F, Alshomary M, Stein B, Hagen M, Potthast M (2018) Reproducible web corpora: interactive archiving with automatic quality assessment. J Data Inf Qual 10(4):1–25. https://doi.org/10.1145/3239574

Klein M, Nelson ML (2011) Find, new, copy, web, page – tagging for the (re-)discovery of web pages. In: International conference on theory and practice of digital libraries (TPDL), vol 6966, pp 27–39. https://doi.org/10.1007/978-3-642-24469-8_5

Klein M, Nelson ML (2014) Moved but not gone: an evaluation of real-time methods for discovering replacement web pages. Int J Digital Libraries 14(1–2):17–38. https://doi.org/10.1007/s00799-014-0108-0

Klein M, Van de Sompel H (2015) Reference rot in web-based scholarly communication and link decoration as a path to mitigation. https://blogs.lse.ac.uk/impactofsocialsciences/2015/02/05/reference-rot-in-web-based-scholarly-communication/

Klein M, Aly M, Nelson ML (2011) Synchronicity: automatically rediscover missing web pages in real time. In: ACM/IEEE joint conference on digital libraries, p 475. https://doi.org/10.1145/1998076.1998193

Klein M, Van de Sompel H, Sanderson R, Shankar H, Balakireva L, Zhou K, Tobin R (2014) Scholarly context not found: one in five articles suffers from reference rot. PLoS One 9(12):e115253. https://doi.org/10.1371/journal.pone.0115253

Klein M, Balakireva L, Van de Sompel H (2018) Focused crawl of web archives to build event collections. In: ACM conference on web science, pp 333–342. https://doi.org/10.1145/3201064.3201085

Klein M, Balakireva L, Shankar H (2019a) Evaluating memento service optimizations. In: ACM/IEEE joint conference on digital libraries, pp 182–185. https://doi.org/10.1109/JCDL.2019.00034

Klein M, Balakireva L, Shankar H (2019b) Evaluating memento service optimizations. Technical Report. arXiv:1906.00058. https://arxiv.org/abs/1906.00058

Ko L (2019) OpenWayback - IIPC. http://netpreserve.org/web-archiving/openwayback/

Kreymer I (2019) GitHub – webrecorder/pywb – Core Python Web Archiving Toolkit for replay and recording of web archives. https://github.com/webrecorder/pywb

Mannens E, Coppens S, Verborgh R, Hauttekeete L, Van Deursen D, Van de Walle R (2012) Automated trust estimation in developing open news stories: combining memento & provenance. In: IEEE annual computer software and applications conference workshops, pp 122–127. https://doi.org/10.1109/COMPSACW.2012.32

Meinhardt P, Knuth M, Sack H (2015) TailR: a platform for preserving history on the web of data. In: International conference on semantic systems, pp 57–64. https://doi.org/10.1145/2814864.2814875

Melo F, Viana H, Gomes D, Costa M (2016) Architecture of the Portuguese web archive search system version 2. Technical report, Arquivo.pt - The Portuguese Web Archive. https://sobre.arquivo.pt/wp-content/uploads/architecture-of-the-portuguese-web-archive-search-1.pdf

Milligan I (2019) History in the age of abundance: how the web is transforming historical research. McGill-Queen's University Press, Montreal

Nelson ML (2010) Memento-datetime is not last-modified. https://ws-dl.blogspot.com/2010/11/2010-11-05-memento-datetime-is-not-last.html

Nelson ML (2013) Archive.is supports memento. https://ws-dl.blogspot.com/2013/07/2013-07-09-archiveis-supports-memento.html

Nelson ML, Van de Sompel H (2019) Adding the dimension of time to HTTP. In: SAGE handbook of web history. SAGE Publishing, Philadelphia, pp 189–214

Neumaier S, Umbrich J, Polleres A (2017) Lifting data portals to the web of data. In: Linked data on the web workshop, pp 1–10. http://ceur-ws.org/Vol-1809/article-03.pdf

Nwala AC (2015) What did it look like? https://ws-dl.blogspot.com/2015/01/2015-02-05-what-did-it-look-like.html

Nwala AC, Weigle MC, Ziegler AB, Aizman A, Nelson ML (2017) Local memory project: providing tools to build collections of stories for local events from local sources. In: ACM/IEEE joint conference on digital libraries, pp 1–10. https://doi.org/10.1109/JCDL.2017.7991576

Nwala AC, Weigle MC, Nelson ML (2018a) Bootstrapping web archive collections from social media. In: ACM conference on hypertext and social media, pp 64–72. https://doi.org/10.1145/3209542.3209560

Nwala AC, Weigle MC, Nelson ML (2018b) Scraping SERPs for archival seeds: it matters when you start. In: ACM/IEEE joint conference on digital libraries, pp 263–272. https://doi.org/10.1145/3197026.3197056

Powell JE, Alcazar DA, Hopkins M, McMahon TM, Wu A, Collins L, Olendorf R (2011) Graphs in libraries: a primer. Inf Technol Libraries 30(4):157. https://doi.org/10.6017/ital.v30i4.1867

Reich V, Rosenthal DSH (2001) LOCKSS: a permanent web publishing and access system. D-Lib Mag 7(6). http://dlib.org/dlib/june01/reich/06reich.html

Ruest N, Milligan I, Lin J (2019) Warclight: a rails engine for web archive discovery. In: ACM/IEEE joint conference on digital libraries, pp 442–443. https://doi.org/10.1109/JCDL.2019.00110

SalahEldeen HM, Nelson ML (2012) Losing my revolution: how many resources shared on social media have been lost? In: International conference on theory and practice of digital libraries (TPDL), vol 7489, pp 125–137. https://doi.org/10.1007/978-3-642-33290-6_14

SalahEldeen HM, Nelson ML (2013a) Carbon dating the web: estimating the age of web resources. In: International world wide web conference, pp 1075–1082. https://doi.org/10.1145/2487788.2488121

SalahEldeen HM, Nelson ML (2013b) Reading the correct history?: Modeling temporal intention in resource sharing. In: ACM/IEEE joint conference on digital libraries, pp 257–266. https://doi.org/10.1145/2467696.2467721

Sanderson R (2012) Global web archive integration with memento. In: ACM/IEEE joint conference on digital libraries, p 379. https://doi.org/10.1145/2232817.2232900

Sanderson R, Van de Sompel H (2010) Making web annotations persistent over time. In: ACM/IEEE joint conference on digital libraries, pp 1–10. https://doi.org/10.1145/1816123.1816125

Sanderson R, Van de Sompel H (2012) Cool URIs and dynamic data. IEEE Internet Comput 16(4):76–79. https://doi.org/10.1109/MIC.2012.78

Sanderson R, Phillips M, Van de Sompel H (2011) Analyzing the persistence of referenced web resources with memento. Technical Report. arXiv:1105.3459. https://arxiv.org/abs/1105.3459

Sanderson R, Ciccarese P, Young B (2017) Web annotation data model. https://www.w3.org/TR/annotation-model

Shelby Z (2012) RFC 6690 – Constrained RESTful Environments (CoRE) link format. https://tools.ietf.org/html/rfc6690

Taelman R, Verborgh R, Mannens E (2017) Exposing RDF archives using triple pattern fragments. In: Knowledge engineering and knowledge management (EKAW), pp 188–192. https://doi.org/10.1007/978-3-319-58694-6_29

Van de Sompel H, Davis S (2015) From a system of journals to a web of objects. Serials Librarian 68(1–4):51–63. https://doi.org/10.1080/0361526X.2015.1026748

Van de Sompel H, Vander Sande M (2016) DBpedia archive using memento, triple pattern fragments, and HDT. In: CNI spring meeting. https://www.slideshare.net/hvdsomp/dbpedia-archive-using-memento-triple-pattern-fragments-and-hdt

Van de Sompel H, Sanderson R, Nelson ML (2010) An HTTP-based versioning mechanism for linked data. In: Linked data on the web workshop, pp 1–10. http://events.linkeddata.org/ldow2010/papers/ldow2010_paper13.pdf

Van de Sompel H, Nelson M, Sanderson R (2013) RFC 7089 - HTTP framework for time-based access to resource states – memento. https://tools.ietf.org/html/rfc7089

Van de Vyvere B, Colpaert P, Mannens E, Verborgh R (2019) Open traffic lights: a strategy for publishing and preserving traffic lights data. In: International world wide web conference, pp 966–971. https://doi.org/10.1145/3308560.3316520

Vander Sande M, Verborgh R, Hochstenbach P, Van de Sompel H (2018) Toward sustainable publishing and querying of distributed linked data archives. J Doc 74(1):195–222. https://doi.org/10.1108/JD-03-2017-0040

Verborgh R, Vander Sande M, Hartig O, Van Herwegen J, De Vocht L, De Meester B, Haesendonck G, Colpaert P (2016) Triple pattern fragments: a low-cost knowledge graph interface for the Web. J Web Semant 37–38:184–206. https://doi.org/10.1016/j.websem.2016.03.003

Weigle MC (2017) Visualizing webpage changes over time - new NEH digital humanities advancement grant. https://ws-dl.blogspot.com/2017/10/2017-10-16-visualizing-webpage-changes.html

Welsh B (2019) Memento for Wordpress. http://pastpages.github.io/wordpress-memento-plugin/

Zhou K, Grover C, Klein M, Tobin R (2015) No more 404s: predicting referenced link rot in scholarly articles for pro-active archiving. In: ACM/IEEE joint conference on digital libraries, pp 233–236. https://doi.org/10.1145/2756406.2756940