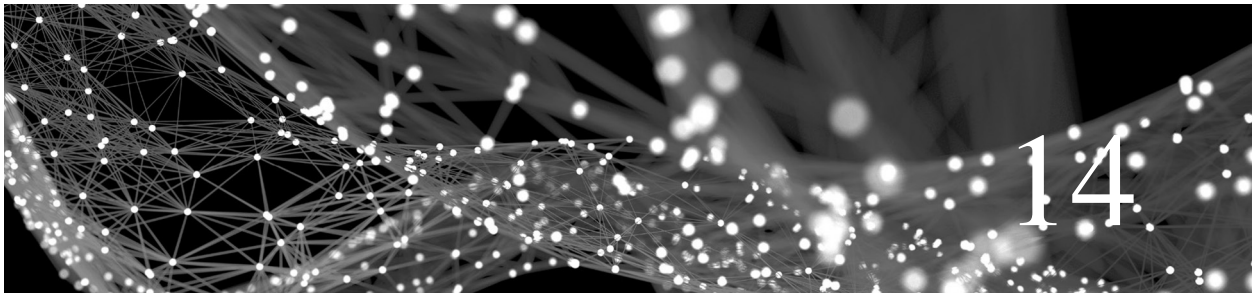

14

Adding the Dimension of Time to HTTP

Michael L. Nelson and Herbert Van de Sompel

INTRODUCTION

While the web is distributed, most web archives are centralized silos that do not cooperate with each other. This is partially because the technology that is necessary to replay the archived content and keep it from being influenced by material on the live web also makes it difficult for web archives to cooperate. The Memento Protocol (which we played a central role in defining) addresses this problem by defining an extension to the Hypertext Transfer Protocol (HTTP) that allows for standardized, machine-readable integration of both the past web and the present web. The Memento Protocol extends the concept of HTTP content negotiation to include not only well-known dimensions such as Multipurpose Internet Mail Extensions (MIME) types (e.g., JPEG vs. PNG) and file encodings (e.g., gzip vs. compress), but also the dimension of Coordinated Universal Time (UTC) as a universal versioning system. The protocol can be supported by all systems that hold temporal resource versions, including conventional web archives, as well as resource versioning systems such as wikis.

The Memento Protocol introduces some standard terminology with which to discuss web archiving, the most fundamental of which are: original resource (the resource on the live web), Memento (an archived version of an Original Resource, frozen in time), TimeGate (a resource capable of datetime content negotiation to discover a temporally appropriate Memento), and TimeMap (a machine-readable list of all Mementos for an Original Resource). Furthermore, the Memento Protocol is the first web archiving API, enabling aggregation of access to disparate web archives. Web archiving has been dominated by the Internet Archive's Wayback Machine, but via the Memento Protocol it is possible to leverage the more than a dozen publicly accessible web archives throughout the world for increased completeness, consistency, verifiability, resilience, and availability.

This chapter begins by introducing the history of Unix and HTTP and how they continue to influence the design of web archives today, then reviews the Memento terminology and concepts that allow for standardized discussion of the basic mechanics of web archiving. Next, we review how different web archives can be aggregated and how user agents (i.e., web browsers) can interact with Memento-enabled archives, and then conclude with a brief review of our activities in areas of open research that result when there is an integrated, distributed network of web archives.

HISTORY OF HTTP AND UNIX

HTTP is the protocol (Fielding et al., 1999) that underpins what we know colloquially as 'the web'. A simple version, HTTP version 0.9, was documented in 1991 (Berners-Lee, 1991), and in 1996 a version similar to the version still in use today was defined (Berners-Lee et al., 1996). Unix, in various vendor-specific formats, was the predominant operating system during this timeframe for workstations and mainframes, so the development of HTTP, while technically independent of Unix, is ultimately deeply intertwined with the operating system that made it possible. As such, the history of HTTP is one of incrementally implementing the original vision of a fully featured, distributed filesystem. Initially conceived with read-write and versioning capabilities, early implementations fell short of the original vision due in part to the tight integration with the Unix filesystem. Advances in encryption and authentication have made the read-write capability more widespread, but it was not until the development of the Memento Protocol that the versioning capability was added for HTTP.

Despite versioning (described in terms of generic vs. specific resources) being part of an early design document for the Web (Berners-Lee, 1996), the historical coupling of HTTP and Unix filesystems stood in the way of embracing a time dimension for the web. Since most of the early HTTP servers were implemented on Unix workstations, the Unix filesystem and HTTP were co-deployed in almost all cases. Metadata about files in the Unix filesystem is stored in 'inodes' (a contraction of 'index node'), and the original description of the Unix filesystem defined three notions of time to be stored in an inode: file creation, last use, and last modification (Ritchie and Thompson, 1974). However, at some early point the storage of the file creation time in the inode was replaced with the last modification time of the inode itself. The result was that the last modification and access times of a file are defined, but the creation time, a crucial part of establishing provenance, cannot be stored in a standard Unix filesystem. Thus, files stored on a Unix filesystem, which are subsequently served through HTTP, inherit the semantic limitations of the filesystem, and as a result HTTP has only two notions of time in server responses: the 'Date' header, which provides the date of the response, and the 'Last-Modified' header, which is inherited from the inode.

To better understand how the Memento Protocol integrates with HTTP, many of the figures listed below provide raw, actual HTTP requests and responses. The web browsers we use everyday (e.g., Firefox, Chrome) hide these detailed and verbose HTTP requests and responses from us, but in this chapter we choose to surface these details because they are integral to understanding the web archiving infrastructure that the Memento Protocol enables. We recognize the HTTP responses can be intimidating to those not used to reading them, but with a short primer they quickly become invaluable. We will walk through how to read the HTTP session in Figure 14.1.

First, for every HTTP response, we include the curl request, with all the appropriate arguments, that generated the response. 'Curl' is a command line web browser that does not render web pages (like Firefox, Chrome, etc.), but rather shows the raw HTTP response from the web server. The '$' symbol in the

figures is the command line prompt, and what follows is what the user types. If you are in a terminal program, you should be able to copy and paste everything after the '$' and get a similar response (keep in mind that responses will change over time). Not all curl options and other command line arguments are fully explained in this chapter, but in the interest of reproducibility they are provided as executed. For example, the following line uses curl to return just the metadata (via the '-I' argument) for an image at lanl.gov:

```
$ curl -I http://www.lanl.gov/_assets/
   images/lanl-logo-footer.png
```

The next line is the response from the server and has three components: 'HTTP/1.1' is an acknowledgment from the server that it is supporting the 1.1 version of HTTP (new and older versions exist, but 1.1 is still currently the most commonly deployed). Next, '200' is the numerical code that provides the semantics of the response; in this case '200' means the request was understood, processed, and there were no errors. The next portion is a human-readable phrase, in this case 'OK', which is explanatory for the '200' response code. The phrase is for human readability and web browsers only process the numeric code. In Figure 14.1, the full response is:

```
HTTP/1.1 200 OK
```

There are many other HTTP response codes defined, but in this chapter we will focus on '200' and '302'. The following 302 responses are equivalent, since the text phrase is for humans and only the code '302' is processed by browsers:

```
HTTP/1.1 302 Found
```

Or:

```
HTTP/1.1 302 Moved Temporarily
```

The 302 response is a redirection from the requested URI to another URI. The response provides metadata, but also instructs the browser to issue a new request for a different URI (which in turn may also issue a redirection, until the process stops with a '200' response). Redirections frequently occur, but regular web browsers hide them and users are typically unaware of this fundamental HTTP event. The argument '-L' will cause curl to automatically follow the redirection.

The final part of an HTTP response is a series of lines of metadata (which can appear in any order), arranged in a 'key: value' format, followed by a carriage return. For example, these four lines indicate the date of the response (2017-02-07), when the resource was last modified (2014-10-28), that the returned representation is an image in 'PNG' format, and that the image is 8,719 bytes long:

```
Date: Tue, 07 Feb 2017 00:08:10 GMT
Last-Modified: Tue, 28 Oct 2014
   22:12:02 GMT
Content-Length: 8719
Content-Type: image/png
```

Now we can examine the HTTP response in Figure 14.1 and see that the Los Alamos National Laboratory (LANL) logo PNG file was last modified in 2014, over two years ago from the request in 2017. Thus, if the file were stored in a cache after the Last-Modified date, it does not need to be downloaded again.

On the other hand, pages that are dynamically generated typically do not set the 'Last-Modified' header, in part because it is not stored by default in the filesystem and thus is extra work to track and compute. The result is that the web is losing expressiveness about time, not gaining it. For example:

```
$ curl -I http://www.lanl.gov/_assets/
   images/lanl-logo-footer.png
HTTP/1.1 200 OK
Date: Tue, 07 Feb 2017 00:08:10 GMT
Last-Modified: Tue, 28 Oct 2014
   22:12:02 GMT
Content-Length: 8719
Content-Type: image/png
```

Figure 14.1 The Last-Modified response header often exists for images, pdfs, and other typically static files.

Figure 14.2 is for the HTML home page that embeds (among other resources) the PNG from Figure 14.1. The HTML home page has surely changed from 2014, but it might not have changed from yesterday or even last week. Unfortunately, without the Last-Modified header we cannot be sure and we are likely to have unnecessary downloads if we visit the page often.

A careful inspection of Figure 14.2 shows the 'Vary' response header, indicating that the server can perform content negotiation on this resource; in this case the server is capable of transmitting a compressed version of this HTML page for quicker download times, and the users' browser will uncompress it on their behalf. Note that while the 'Vary' header indicates that content negotiation is possible, it is the presence of the 'Content-Encoding' response header that indicates that compression has actually occurred (Figure 14.3).

Tim Berners-Lee's original design document about 'generic' and 'specific' resources that formed the basis for content negotiation did not anticipate character sets or encodings (Berners-Lee, 1996), but it did describe 'target

```
$ curl -I http://www.lanl.gov/
HTTP/1.1 200 OK
Date: Tue, 07 Feb 2017 00:08:46 GMT
Vary: Accept-Encoding
Content-Type: text/html; charset=UTF-8
```

Figure 14.2 The Last-Modified response header is typically absent from resources with dynamically constructed representations (i.e., almost all HTML files).

```
$ curl -I -H "Accept-Encoding: gzip"
   http://www.lanl.gov/
HTTP/1.1 200 OK
Date: Tue, 07 Feb 2017 05:05:31 GMT
Vary: Accept-Encoding
Content-Encoding: gzip
Content-Length: 20
Content-Type: text/html; charset=UTF-8
```

Figure 14.3 Based on the 'Accept-Encoding' request header, the server responds with a gzipped HTML page, as declared in the 'Content-Encoding' response header.

mediums' (or 'features'), which have now been replaced with cascading style sheets (CSS) functionality, and time, which the Memento Protocol made possible nearly 20 years later.

MEMENTO TERMINOLOGY AND CONCEPTS

To better understand the Memento Protocol (Van de Sompel et al., 2009, 2010, 2013), we must clarify some terminology that is often used imprecisely even in technical writing. Figure 14.4 is from the seminal Web Architecture document (Jacobs and Walsh, 2004), and shows: 1) Uniform Resource Identifiers (URI), 2) Resources, and 3) Representations.

URIs, which are a superset of the more commonly known Uniform Resource Locators (URLs), identify *resources*. At any given moment, resources exist in a certain state, and that state can be serialized into a *representation* of the resource. It is this representation that is transmitted via HTTP, rendered by the browser, etc. – the resource itself is not transferred, and indeed can be a real-world object (i.e., non-digital). When a URI is dereferenced (most commonly with HTTP, though many other URI schemes (i.e., protocols) are defined), the representation of that resource is returned. As shown in Figures 14.2 and 14.3, the representation can vary based on input from the client (e.g., compressed or not), which means there can simultaneously be different representations that capture the current state of the resource.

The Memento Protocol specifies how to retrieve a representation of a prior, not current, state of a resource by allowing a client to express the datetime of the prior state it is interested in. Building on the Web Architecture, the Memento Protocol introduces standard terminology with which to discuss the mechanics of versioning on the web (Van de Sompel et al., 2013):

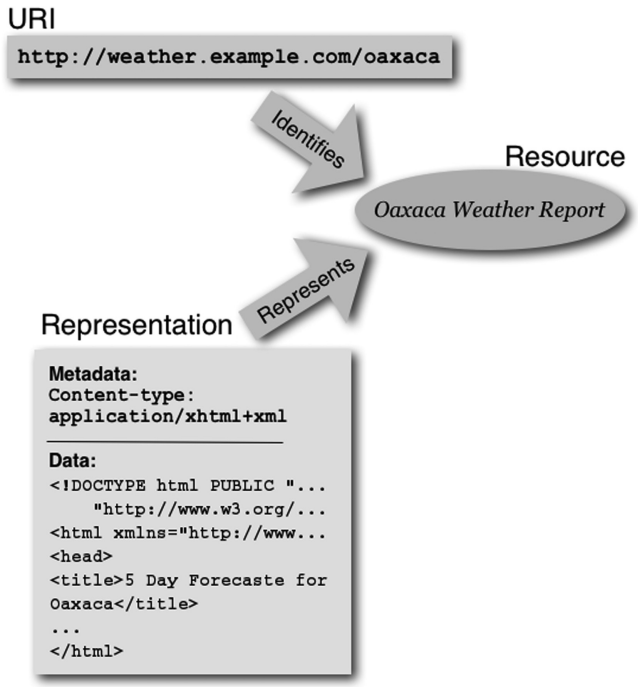

Figure 14.4 URIs, resources, and representations (Jacobs and Walsh, 2004).

- Original Resource: an Original Resource (identified by URI-R) is a resource that exists or used to exist, and for which access to one of its prior states may be required.
- Memento: a Memento (identified by URI-M) for an Original Resource is a resource that encapsulates a prior state of the Original Resource. A Memento for an Original Resource as it existed at time T is a resource that encapsulates the state the Original Resource had at time T.
- TimeGate: a TimeGate (identified by URI-G) for an Original Resource is a resource that is capable of datetime negotiation to support access to prior states of the Original Resource.
- TimeMap: a TimeMap (identified by URI-T) for an Original Resource is a resource from which a list of URIs of Mementos of the Original Resource is available.

When an archive crawls an Original Resource, it saves the state of the resource at the time it was crawled. This state now becomes its own frozen resource, a Memento, and the time at which it was crawled is stored in the resource's Memento-Datetime response header. Note that the value for Memento-Datetime can be different from the Last-Modified header sent from the archive, since the archives can continually update banners, archive logos, and other extra information injected into the Memento when it is replayed (Nelson, 2011). For example, the banner at the top of Figure 14.18 is supplied by the web archive itself and the contents of the banner are updated over time, necessitating updated values for Last-Modified even though the Memento-Datetime value does not change.

For the Original Resource http://www.lanl.gov/, the first known Memento is in the Internet Archive's Wayback Machine (the first (operational in 1996) and by far the largest public web archive (Negulescu, 2010)) with a URI-M of http://web.archive.org/web/19961221031231/http://lanl.gov/. We can examine its HTTP response in Figure 14.5.

```
$ curl -I http://web.archive.org/web/
   19961221031231/http://lanl.gov/
HTTP/1.1 200 OK
Server: Tengine/2.1.0
Date: Wed, 08 Feb 2017 02:26:53 GMT
Content-Type: text/
html;charset=utf-8
Content-Length: 9237
Memento-Datetime: Sat, 21 Dec 1996
   03:12:31 GMT
Link:       <http://lanl.gov/>;
   rel="original",    <http://web.
   archive.org/web/timemap/link/
   http://lanl.gov/>; rel="timemap";
   type="application/link-
   format",    <http://web.archive.
   org/web/http://lanl.gov/>;
   rel="timegate",    <http://web.
   archive.org/web/19961221031231/
   http://lanl.gov/>;    rel="first
   memento"; datetime="Sat, 21 Dec
   1996 03:12:31 GMT", <http://web.
   archive.org/web/19981212015212/
   http://lanl.gov/>;    rel="next
   memento"; datetime="Sat, 12 Dec
   1998 01:52:12 GMT", <http://web.
   archive.org/web/20170201114455/
   http://lanl.gov/>;    rel="last
   memento"; datetime="Wed, 01 Feb
   2017 11:44:55 GMT"
```

Figure 14.5 HTTP response for a Memento from the Internet Archive.

In this case, the Original Resource and Memento-Datetime are extractable from the URI-M as www.lanl.gov and 19961221031231, respectively. However, not all archives follow this convention and extracting strings from URIs violates the practice of URI opacity as recommended by the Web Architecture (Jacobs and Walsh, 2004). In short, URI opacity means that one should not rely on extracting semantics from substrings in a URI. For example, the string '19961221031231' might not always indicate the Memento-Datetime that it appears to indicate. Furthermore, not all archives use URIs with apparent semantics. For example, the WebCite web archive has archived www.lanl.gov on 2011-06-28, and that page is available at both of these URIs, neither of which indicate the URI of the Original Resource nor the Memento-Datetime:

```
http://webcitation.org/query?
   id=1309246026894208
http://webcitation.org/5zm0eNcVU
```

The standardized, machine-readable method is to return the Memento-Datetime in the response header, with the datetime in a format borrowed from Email (Crocker, 1982), as with the Date and Last-Modified headers:

```
Memento-Datetime: Sat, 21 Dec 1996
   03:12:31 GMT
```

Inside the Link response header, although the syntax is challenging, a careful reading will reveal the unambiguous statements for the Original URI, first, last, and next Mementos, and the TimeGate and TimeMap (note that even though the line wraps, it is a single logical 'key: value' line, where the value has multiple, comma-separated sub-values):

```
Link:   <http://lanl.gov/>;   rel=
   "original", <http://web.archive.
   org/web/timemap/link/http://
   lanl.gov/>;     rel="timemap";
   type="application/link-
   format",    <http://web.archive.
   org/web/http://lanl.gov/>;
   rel="timegate",    <http://web.
   archive.org/web/19961221031231/
   http://lanl.gov/>;    rel="first
   memento"; datetime="Sat, 21 Dec
   1996 03:12:31 GMT", <http://web.
   archive.org/web/19981212015212/
   http://lanl.gov/>;    rel="next
   memento"; datetime="Sat, 12 Dec
   1998 01:52:12 GMT", <http://web.
   archive.org/web/20170201114455/
   http://lanl.gov/>;    rel="last
   memento"; datetime="Wed, 01 Feb
   2017 11:44:55 GMT"
```

In Figure 14.6 we see the HTTP response for the URI-M http://archive.is/OYfTd, and from just this URI-M we cannot extract the URI-R and Memento-Datetime (similar to the WebCite example above); thus the Link

```
$ curl -I http://archive.is/OYfTd
HTTP/1.1 200 OK
Date: Wed, 08 Feb 2017 02:52:25 GMT
Content-Type: text/
  html;charset=utf-8
Memento-Datetime: Wed, 17 Dec 2014
  21:16:53 GMT
Link: <http://www.lanl.gov/>;
  rel="original", <http://archive.
  is/timegate/http://www.lanl.
  gov/>; rel="timegate", <http://
  archive.is/timemap/http://www.
  lanl.gov/>; rel="timemap";
  type="application/link-format";
  from="Sat, 15 Oct 2011 08:20:59
  GMT"; until="Wed, 17 Dec 2014
  21:16:53 GMT", <http://archive.
  is/20141106023554/http://www.
  lanl.gov/>; rel="prev memento";
  datetime="Thu, 06 Nov 2014
  02:35:54 GMT", <http://archive.
  is/20111015082059/http://www.
  lanl.gov/>; rel="first memento";
  datetime="Sat, 15 Oct 2011
  08:20:59 GMT", <http://archive.
  is/20141217211653/http://www.
  lanl.gov/>; rel="last memento";
  datetime="Wed, 17 Dec 2014
  21:16:53 GMT"
```

Figure 14.6 HTTP response for a Memento from archive.is.

and Memento-Datetime response headers are crucial.

We established that the Memento in Figure 14.5 is the first Memento available for http://www.lanl.gov/. There are two ways to verify this with the Internet Archive. The first is to download the TimeMap for http://www.lanl.gov/ and check the first value. Figure 14.7 only shows the first few entries since the TimeMap is quite large, with over 1,800 Mementos.

Some TimeMaps are already larger than 100,000 Mementos, so even though they are sorted by datetime, it can still be a lot to download and parse. To address this need, TimeGates perform datetime content negotiation and will issue an HTTP redirect to the most appropriate Memento. For conventional web archives that are unaware of the change rate of the Original Resource, the TimeGate simply chooses the closest Memento to the requested datetime (as expressed in the Accept-Datetime request header). This algorithm is known as 'mindist' for 'Minimum Distance'. Figure 14.8 shows the client asking the Internet Archive for a Memento closest to October 16, 2013.

```
$ curl --silent http://web.archive.org/web/timemap/link/http://lanl.gov/ |
  head -10
<http://lanl.gov/>; rel="original",
<http://web.archive.org/web/timemap/link/http://lanl.gov/>; rel="self";
  type="application/link-format"; from="Sat, 21 Dec 1996 03:12:31 GMT";
  until="Wed, 01 Feb 2017 11:44:55 GMT",
<http://web.archive.org/web/http://lanl.gov/>; rel="timegate",
<http://web.archive.org/web/19961221031231/http://lanl.gov/>; rel="first
  memento"; datetime="Sat, 21 Dec 1996 03:12:31 GMT",
<http://web.archive.org/web/19981206235030/http://lanl.gov/>; rel="memento";
  datetime="Sun, 06 Dec 1998 23:50:30 GMT",
<http://web.archive.org/web/19981212015212/http://lanl.gov/>; rel="memento";
  datetime="Sat, 12 Dec 1998 01:52:12 GMT",
<http://web.archive.org/web/19981212030449/http://www.lanl.gov/>;
  rel="memento"; datetime="Sat, 12 Dec 1998 03:04:49 GMT",
<http://web.archive.org/web/19990117014439/http://lanl.gov/>; rel="memento";
  datetime="Sun, 17 Jan 1999 01:44:39 GMT",
<http://web.archive.org/web/19990117083819/http://lanl.gov/>; rel="memento";
  datetime="Sun, 17 Jan 1999 08:38:19 GMT",
<http://web.archive.org/web/19990125090547/http://lanl.gov/>; rel="memento";
  datetime="Mon, 25 Jan 1999 09:05:47 GMT",
```

Figure 14.7 The first ten lines of the TimeMap for http://www.lanl.gov/.

```
$ curl -I -H "Accept-Datetime: Wed,
   16 Oct 2013 22:59:48 GMT" http://
   web.archive.org/web/http://lanl.
   gov/
HTTP/1.1 302 Found
Server: Tengine/2.1.0
Date: Wed, 08 Feb 2017 03:38:20 GMT
Content-Type: text/html
Content-Length: 0
Location:    /web/20131019030442/
   http://www.lanl.gov/
Vary: accept-datetime
Link:       <http://lanl.gov/>;
   rel="original",    <http://web.
   archive.org/web/timemap/link/
   http://lanl.gov/>; rel="timemap";
   type="application/link-format",
   <http://web.archive.org/
   web/19961221031231/http://lanl.
   gov/>;   rel="first   memento";
   datetime="Sat,   21   Dec   1996
   03:12:31   GMT",   <http://web.
   archive.org/web/20131014072106/
   http://lanl.gov/>;    rel="prev
   memento"; datetime="Mon, 14 Oct
   2013 07:21:06 GMT", <http://web.
   archive.org/web/20131019030442/
   http://lanl.gov/>; rel="memento";
   datetime="Sat,   19   Oct   2013
   03:04:42   GMT",   <http://web.
   archive.org/web/20131020082626/
   http://lanl.gov/>;    rel="next
   memento"; datetime="Sun, 20 Oct
   2013 08:26:26 GMT", <http://web.
   archive.org/web/20170201114455/
   http://lanl.gov/>;    rel="last
   memento"; datetime="Wed, 01 Feb
   2017 11:44:55 GMT"
```

Figure 14.8 Negotiating with a TimeGate for a Memento of http://www.lanl.gov/ close to October 16, 2013.

In Figure 14.8, the TimeGate issues a redirection to URI-M http://web.archive.org/web/20131019030442/http://lanl.gov/, which although off by three days (October 19, 2013) is the closest available Memento to October 16, 2013 that the Internet Archive has.

The Memento Protocol also applies to transactional archives, such as LANL's SiteStory (Brunelle et al., 2013), and content management systems that act as their own archives, such as wikis. Figure 14.9 shows the HTTP response from the World Wide Web Consortium (W3C, a standards body for the web) wiki which implements our Memento for MediaWiki extension (Jones et al., 2014), and the server is specifying its own TimeGate for the URI-R https://www.w3.org/wiki/SpecProd/Restyle. One could consult the Internet Archive to discover Mementos for this URI-R, but since this is a MediaWiki, the wiki is its own archive and is authoritative on which versions existed at different points in time (see Jones et al., 2016a for an analysis of missed updates and redundant Mementos when web archives interact with wikis). Figure 14.10 shows datetime negotiation requesting a Memento one second before the Memento-Datetime for the last Memento in the wiki. If this were a conventional web archive that did not know the complete version history, it would use mindist and choose the Memento that was one second into the future. But since this is a content management system with a complete version history, it selects the closest Memento with a Memento-Datetime value less than or equal to the Accept-Datetime value because that is the version that was the current one at the requested datetime. This algorithm is known as 'minpast' for 'Minimum Past' and in Figure 14.10 it results in the redirection to a Memento two years in the past instead of one second in the future, which the mindist algorithm would have chosen.

The entire Memento framework comes together as shown in Figure 14.11, where ideally the Original Resource links to the TimeGate (URI-G); if this link is not provided the user's browser can be configured to know the location of a suitable TimeGate. It is the TimeGate's job to handle the datetime negotiation and redirect the user to the closest available Memento. The Mementos in the web archive provide machine-readable links to each other, as well as back to the Original Resource, allowing for seamless navigation between the current and past webs. Because the Memento Protocol is an

```
$ curl -I https://www.w3.org/wiki/
   SpecProd/Restyle
HTTP/1.1 200 OK
Link: <https://www.w3.org/wiki/
   SpecProd/Restyle>; rel="original
   latest-version",<https://www.
   w3.org/wiki/Special:TimeGate/
   SpecProd/Restyle>;
   rel="timegate",<https://www.
   w3.org/wiki/Special:TimeMap/
   SpecProd/Restyle>; rel="timemap";
   type="application/link-format";
   from="Thu, 08 Dec 2011 20:09:41
   GMT"; until="Fri, 11 Mar 2016
   16:02:35 GMT",<https://www.w3.org/
   wiki/index.php?title=SpecProd/
   Restyle&oldid=55833>; rel="first
   memento"; datetime="Thu,
   08 Dec 2011 20:09:41
   GMT",<https://www.w3.org/
   wiki/index.php?title=SpecProd/
   Restyle&oldid=97718>; rel="last
   memento"; datetime="Fri, 11 Mar
   2016 16:02:35 GMT"
Content-language: en
Last-Modified: Mon, 13 Feb 2017
   09:44:38 GMT
Content-Type: text/html;
   charset=UTF-8
Content-Length: 22688
Date: Mon, 13 Feb 2017 17:08:03 GMT
```

Figure 14.9 HTTP response with Memento headers from the W3C MediaWiki.

extension of HTTP and content negotiation, it retains HTTP's compliance with the architectural principles of Representational State Transfer (REST) and Hypermedia as the Engine of Application State (HATEOAS) (Fielding, 2000). In its simplest form, this means that clients interact with the state of resources using only the methods defined in HTTP (i.e., REST) and 'follow their nose' and interact with self-describing resources to discover navigable links to related, typed resources, such as other Mementos, TimeGates, and TimeMaps (i.e., HATEOAS). In summary, the Memento Protocol is not a service separate and apart from a web archive, it is embedded within the normal HTTP operations of a web archive.

MEMENTO AND AGGREGATING MULTIPLE ARCHIVES

In Figure 14.8, the Internet Archive did not have a Memento for the exact date of October 16, 2013 and redirected to one of October 19 instead. For many applications, this three-day difference is inconsequential, but what if it was essential to view the state of that URI-R on October 16, 2013? Users could query different web archives separately, but keeping track of the publicly accessible web archives that natively support the Memento Protocol would require more knowledge than most users possess. For example, the earliest Memento for the Smithsonian Institute's home page (http://www.si.edu/) is in the Portuguese Web Archive and not the Internet Archive (October, 1996 vs. May, 1997) (Fuhrig, 2014). Too much time has passed to definitely say why the Portuguese Web Archive has the earliest page, but the Memento Protocol facilitated the discovery of the Memento in an archive that many people in the United States may not have known.

Fortunately, the Memento Protocol makes it easy to combine various web archives into *aggregators*, which provide a single TimeGate and a single TimeMap for all the web archives that it aggregates (Sanderson, 2012). For example, Los Alamos National Laboratory runs an aggregator available at: http://timetravel.mementoweb.org. To revisit the query in Figure 14.8, finding a Memento of http://www.lanl.gov/ close to October 16, 2013, Figure 14.12 shows the response from the TimeTravel aggregator.

Since the aggregator at timetravel.mementoweb.org currently knows of approximately 30 publicly accessible web archives, it can redirect the client to exactly the desired Memento for http://www.lanl.gov/ on October 16, 2013. These examples are from the publicly accessible aggregator at timetravel.mementoweb.org, but there is no requirement to use this aggregator. Old Dominion University (ODU) runs a

```
$ curl -I -L -H "Accept-Datetime: Fri, 11 Mar 2016 16:02:34 GMT" https://www.
   w3.org/wiki/Special:TimeGate/SpecProd/Restyle
HTTP/1.1 302 Found
Vary: Accept-Encoding,Accept-Datetime
Location:       https://www.w3.org/wiki/index.php?title=SpecProd/
   Restyle&oldid=77827
Link:     <https://www.w3.org/wiki/Special:TimeMap/SpecProd/Restyle>;
   rel="timemap"; type="application/link-format"; from="Thu, 08 Dec 2011
   20:09:41 GMT"; until="Fri, 11 Mar 2016 16:02:35 GMT",<https://www.w3.org/
   wiki/index.php?title=SpecProd/Restyle&oldid=55833>; rel="first memento";
   datetime="Thu, 08 Dec 2011 20:09:41 GMT",<https://www.w3.org/wiki/index.
   php?title=SpecProd/Restyle&oldid=97718>; rel="last memento"; datetime="Fri,
   11 Mar 2016 16:02:35 GMT",<https://www.w3.org/wiki/SpecProd/Restyle>;
   rel="original latest-version"
Content-Type: text/html; charset=UTF-8
Content-Length: 0
Date: Mon, 13 Feb 2017 17:17:07 GMT

HTTP/1.1 200 OK
Memento-Datetime: Fri, 10 Oct 2014 04:07:37 GMT
Link:  <https://www.w3.org/wiki/SpecProd/Restyle>; rel="original latest-
   version",<https://www.w3.org/wiki/Special:TimeGate/SpecProd/Restyle>;
   rel="timegate",<https://www.w3.org/wiki/Special:TimeMap/SpecProd/Restyle>;
   rel="timemap"; type="application/link-format"; from="Thu, 08 Dec 2011
   20:09:41 GMT"; until="Fri, 11 Mar 2016 16:02:35 GMT",<https://www.w3.org/
   wiki/index.php?title=SpecProd/Restyle&oldid=55833>; rel="first memento";
   datetime="Thu, 08 Dec 2011 20:09:41 GMT",<https://www.w3.org/wiki/index.
   php?title=SpecProd/Restyle&oldid=97718>; rel="last memento"; datetime="Fri,
   11 Mar 2016 16:02:35 GMT"
Content-language: en
Vary: Accept-Encoding,Cookie
Content-Type: text/html; charset=UTF-8
Content-Length: 21235
Date: Mon, 13 Feb 2017 17:17:07 GMT
```

Figure 14.10 Datetime negotiation with a MediaWiki TimeGate for one second before the latest Memento; MediaWiki uses the minpast algorithm instead of mindist.

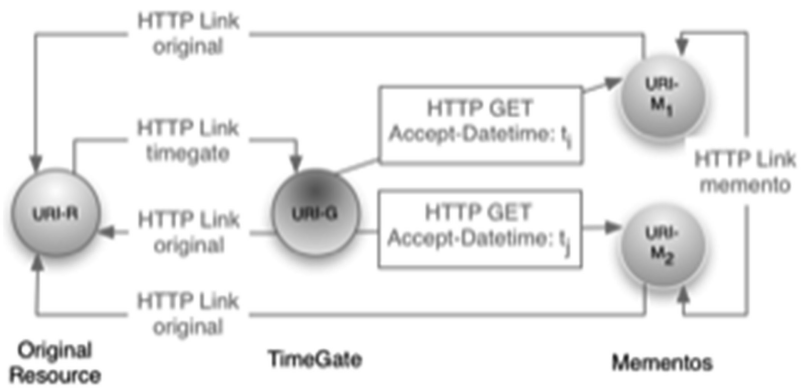

Figure 14.11 Architectural overview of how the Memento framework allows a representation of a prior state of a resource to be accessed.

```
$ curl -I -H "Accept-Datetime:
  Wed, 16 Oct 2013 22:59:48 GMT"
  http://timetravel.mementoweb.org/
  timegate/http://www.lanl.gov/
HTTP/1.1 302 Moved Temporarily
Server: nginx/1.10.1
Date: Thu, 09 Feb 2017 17:50:37 GMT
Content-Type: text/plain;
  charset=iso-8859-1
Content-Length: 0
Location: http://archive.
  is/20131016225948/http://www.
  lanl.gov/
Vary: Accept-Datetime
Last-Modified: Wed, 08 Feb 2017
  04:27:12 GMT
Link: <http://www.lanl.
  gov/>;rel="original", <http://
  timetravel.mementoweb.org/
  timemap/link/http://www.
  lanl.gov/>;rel="timemap";
  type="application/link-
  format",<http://archive.
  is/20131016225948/http://www.
  lanl.gov/>;rel="memento";
  datetime="Wed, 16 Oct 2013 22:59:48
  GMT",<http://web.archive.bibalex.
  org:80/web/19961221031231/
  http://lanl.gov/>;rel="memento
  first"; datetime="Sat, 21 Dec
  1996 03:12:31 GMT",<http://web.
  archive.org/web/20170201114455/
  http://lanl.gov/>;rel="memento
  last"; datetime="Wed, 01 Feb 2017
  11:44:55 GMT"
```

Figure 14.12 A response from an aggregated TimeGate, redirecting to http://archive.is/20131016225948/http://www.lanl.gov/.

separate aggregator at memgator.cs.odu.edu, and a query to this aggregator shows that Mementos for http://www.lanl.gov/ can be found in at least eight publicly accessible web archives with native Memento Protocol support (Figure 14.13). While the Internet Archive (archive.org) clearly has the most with 1,810 Mementos and Archive-It (a separate subscription service of the Internet Archive) is second with 311, there are still more than 250 Mementos in other web archives.

```
$ curl --silent http://memgator.
  cs.odu.edu/timemap/link/http://
  www.lanl.gov/ | grep datetime |
  awk '{print $1}' | awk -v FS=/
  '{print $3}' | sort | uniq -c
  36 archive.is
3 arquivo.pt
6 swap.stanford.edu
311 wayback.archive-it.org
1 wayback.vefsafn.is
1810 web.archive.org
228 webarchive.loc.gov
1 webarchive.parliament.uk
```

Figure 14.13 The processed TimeMap showing the hostnames of the eight public web archives with Mementos for http://www.lanl.gov/ and their respective Memento counts.

The MemGator software is available for download and local installation (Alam and Nelson, 2016), making it possible to set up a Memento aggregator with custom sets of web archives, suitable for local or private deployments.

RIGHT-CLICK TO THE PAST

To this point, we have discussed the Memento Protocol in terms of HTTP interactions, using the curl command line user-agent and examining raw HTTP responses. While this is necessary to understand the mechanics of the Memento Protocol, there are a range of more user-friendly clients and services using the Memento Protocol that provide a more seamless integration of the current and past web.

The easiest way to get started is to visit http://timetravel.mementoweb.org/ and input the desired URI-R and datetime. In Figure 14.14, we use http://www.lanl.gov/ and 2013-10-16, respectively; this is effectively the user-friendly version of the curl request shown in Figure 14.12. In Figure 14.15, we see the results, sorted by web archives with a Memento closest to the desired datetime,

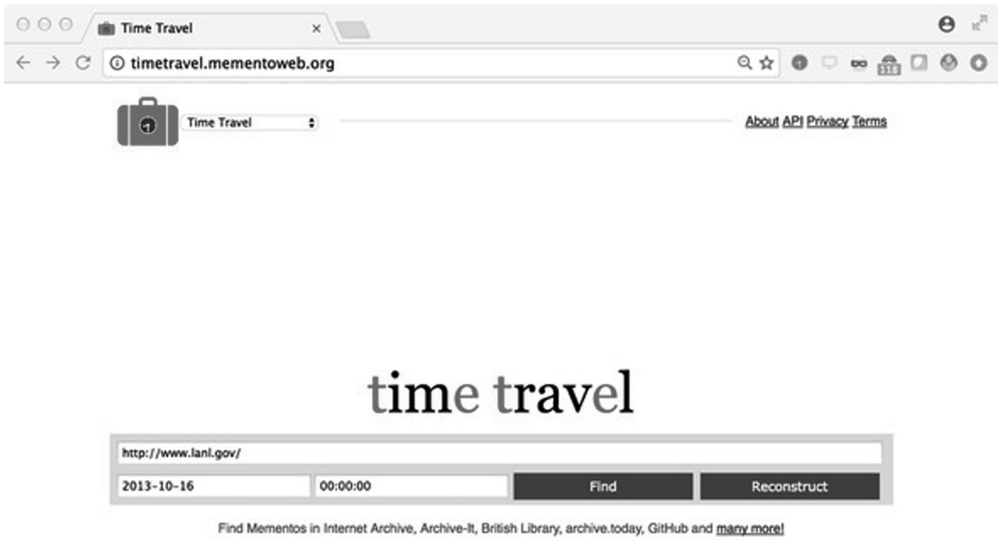

Figure 14.14 A request to the TimeTravel service with URI-R = http://www.lanl.gov/ and datetime=2013-10-16.

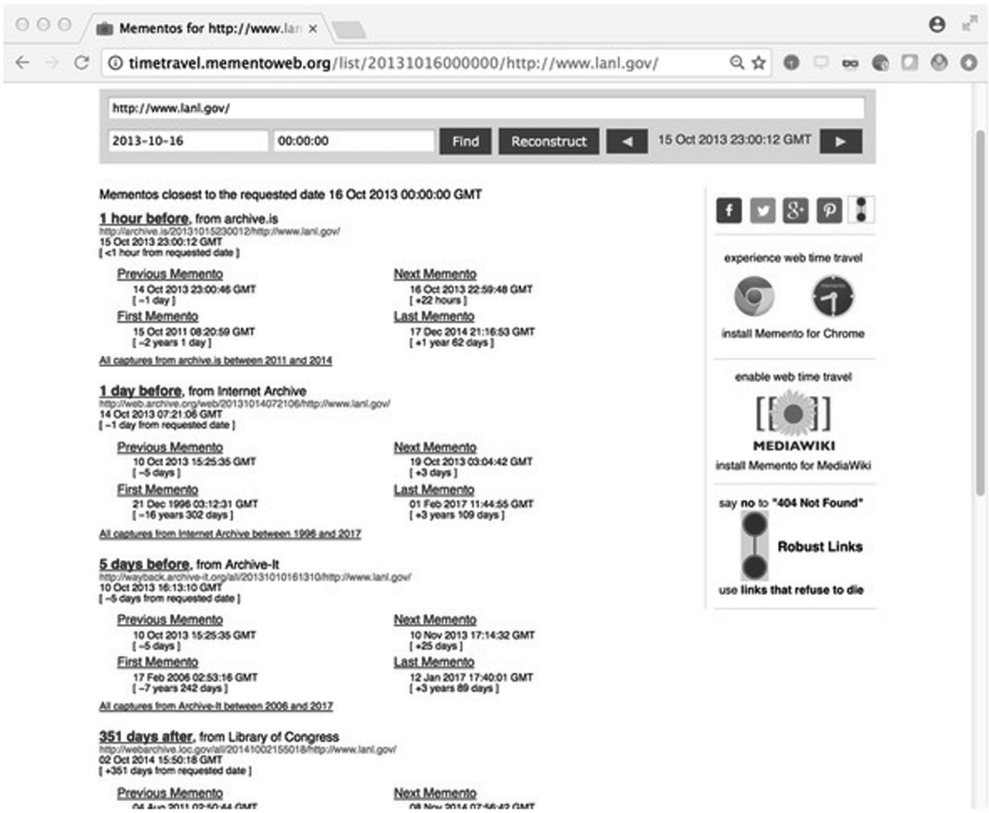

Figure 14.15 The response to the request shown in Figure 14.12, with seven archives holding Mementos for this URI-R (available at: http://timetravel.mementoweb.org/list/20131016000000/http://www.lanl.gov/).

with the archive.is Memento at the top of the list, followed by the Internet Archive, and a total of seven different archives (when the page scrolled to the bottom). The archives differ slightly from the ODU aggregator result shown in Figure 14.13, in part because of a prediction interface employed by the LANL Time Travel service, which is discussed below.

While the Time Travel service is an attractive and easy to use interface, it is still a destination separate from the live web, meaning that users need to know of its location and choose to navigate there to explore the past web. While this is suitable for extended sessions of interacting with the past web, such sessions do not reflect how most humans use web archives. In a study of accesses to the Internet Archive, AlNoamany found that 82% of human sessions begin with referrals from live web pages (the majority of which come from Wikipedia), and 86% of those links no longer exist on the live web (AlNoamany, 2014). In short, web archives are primarily used as a versioning system to supplement failures in the live web, so we should have user agents that support this modality.

There are several clients for providing Memento-based access, including the original MementoFox (now deprecated) (Sanderson et al., 2011), Mink (which also facilitates archiving of live web pages) (Kelly et al., 2014b), and Memento for iOS (Tweedy et al., 2013). We will focus on the 'Memento for Chrome' extension, which provides Chrome users the capability to 'right-click to the past' (Nelson, 2013). Regular clicks provide the expected navigation (i.e., staying on the live web), but users can choose to right-click on a link or just in the middle of the page for the current URI to seamlessly provide access to the same functionality shown in Figures 14.12, 14.14, and 14.15. Figure 14.16 shows the user setting the desired datetime (October 16, 2013) with a calendar widget (the extension will use this value in the Accept-Datetime request header). Figure 14.17 shows the user right-clicking in the middle of the page, indicating the desired URI-R is 'this' page, i.e., http://www.lanl.gov/. There are several datetime options, but here the user is selecting the datetime as set in Figure 14.16. After selecting the link in Figure 14.17, the client then communicates with the Time Travel aggregator (effectively issuing the request shown in Figure 14.12), and the result is the client is directed to the URI-M http://archive.is/20131016225948/http://www.lanl.gov/ (Figure 14.18). To navigate back to the live web, the user right-clicks in the middle of the page again and selects 'get at current date' (Figure 14.19). The described interactions are also possible for links in pages and embedded resources such as images.

ONGOING RESEARCH

The Memento Protocol has opened up entirely new areas of research regarding inter-archive access and collaboration. Though not an exhaustive list, we provide an overview of some of our recent research activities which the Memento Protocol has made possible and, in some cases, necessary.

Routing URI Requests to the Proper Web Archives

When there was only one archive (i.e., the Internet Archive), access was simple: the archive either had the desired Memento or it did not. The addition of multiple web archives, which began in earnest in the mid to late 2000s (Bailey et al., 2013; Gomes et al., 2011), increases the complexity significantly. Either the user has to navigate to different web archives and check for existence, which is limited by the user's time and knowledge of the archives, or a service has to aggregate access to these archives (cf. the Time Travel service in Figures 14.14 and 14.15). When the number of web archives an aggregator has to select is small, it is easy to broadcast the URI lookup to all

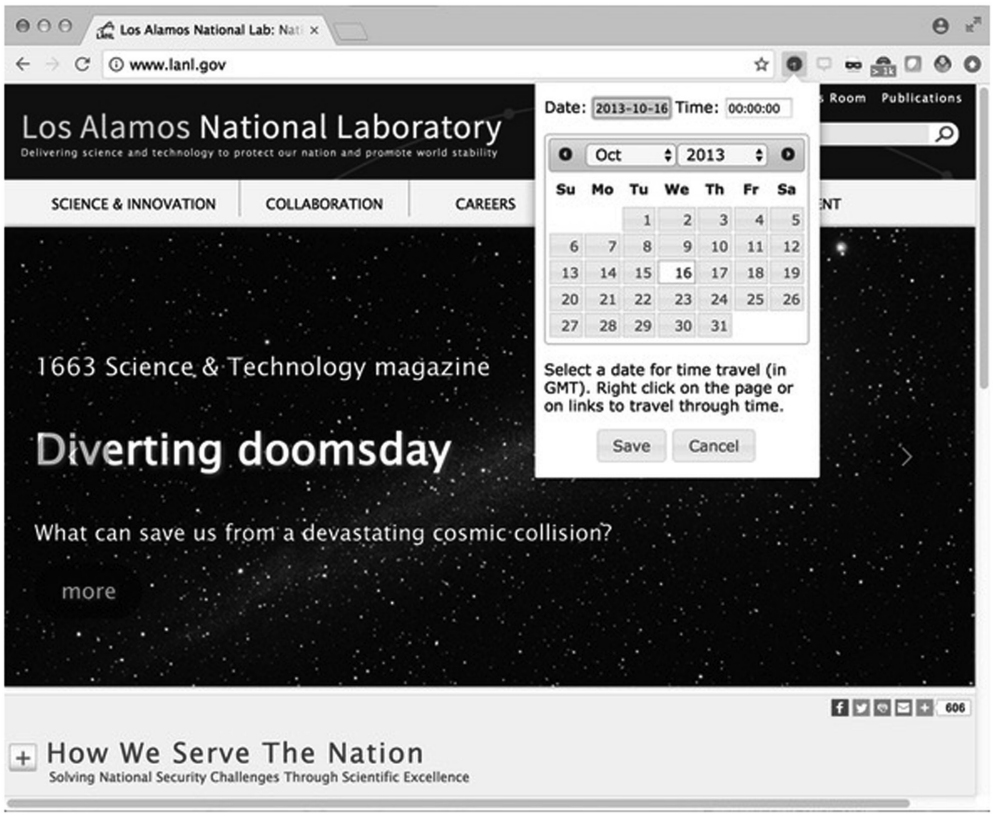

Figure 14.16 Setting the datetime to October 16, 2013 for http://www.lanl.gov/.

known web archives and then synthesize the results. As the number of web archives grows this become untenable, with responses waiting for the slowest archive to respond, while all along most of the web archives do not hold the desired Memento (i.e., a 404 HTTP response). This problem is well-known in the distributed search community as the query routing problem (Callan, 2002), even though URI lookups are not exactly like full-text queries in that the archives either have the Memento or they do not, and there is no ranked list of potentially relevant hits. For web archives, we want to predict where to send URI lookup requests such that we only query archives which are likely to have Mementos for the requested Original URI. This is not as obvious as it first seems – crawling scope is notoriously difficult and imprecise and archives often capture more than they intended.

In our earliest research in URI routing (AlSum et al., 2013, 2014), we proved that 52% of the time we can still produce a complete TimeMap querying only the top three (of a possible 12 available at that time) web archives even if we exclude the Internet Archive. Querying the top six web archives (excluding the Internet Archive) produces a complete TimeMap 77% of the time. In summary, despite a growing number of web archives, we need only query a few of them to get the complete list of all available Mementos, and even if the Internet Archive were to disappear, over one-half of the Original URIs would be unaffected.

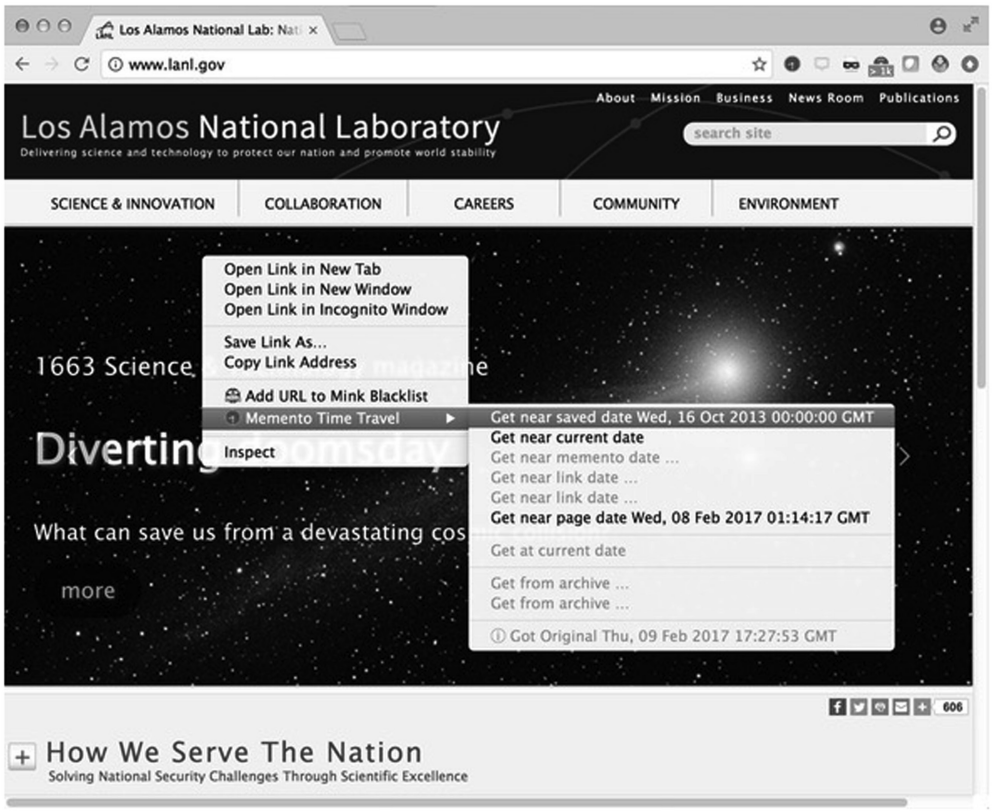

Figure 14.17 Right-clicking in the middle of the page to expose datetime negotiation options for http://www.lanl.gov/.

Of course, the question is: how do we know where to send the URI lookups? We have explored a variety of methods, including building web archive profiles using their CDX files (produced for playback in the Wayback Machine archival playback software) and HTTP access logs of web archives (Alam et al., 2015; Alam et al., 2016b), machine learning techniques based training data from responses of web archives for a fixed set of URI lookups (Bornand et al., 2016) (this method is in production in the Time Travel service shown in Figures 14.14 and 14.15), and – for web archives that have a full-text index – extracting terms from archived documents and using those as queries back into the archive (Alam et al., 2016a).

All of these methods have trade-offs. Using CDX files and HTTP access logs requires cooperation from the participating archives to make those files available for processing into a profile. Profiles built from learning from responses to lookups are sensitive to the URIs used in the training phase and thus will not reveal Original URIs that you do not know to ask for. Using keywords extracted from documents assumes a full-text search interface is available (which is often available only in smaller archives) and requires a large number of full-text queries to extract enough documents to profile its holdings. Our current research in this area involves developing approaches that are hybrids of the methods described above, as well as addressing

Figure 14.18 The user is now at the Memento http://archive.is/20131016225948/http://www.lanl.gov/.

optimum approaches for updates and reevaluating baseline profiles.

How to Use Multiple Archives

Figures 14.12, 14.14, and 14.15 show an aggregator consulting multiple web archives and then redirecting the requesting client to the web archive that contains the closest available Memento. But once the client follows that redirection, all the requests for embedded resources are relative to the archive holding the root HTML page (in the case of Figures 14.12, 14.14, and 14.15, all requests are relative to http://archive.is/). This means if there are links, images, or other embedded resources missing, the client is unable to take advantage of the other web archives that might hold those resources.

Some embedded resources, links, or even entire sites might not be present in the Internet Archive. One reason an entire site might not be present is because the site, for any number of reasons, is using a 'robots.txt' file on the live web to limit access by the Internet Archive, for both future crawls and any content they currently hold (Rossi, 2016). Figure 14.20 shows the Internet Archive's response for the URI-R https://www.quora.com/, and Figure 14.21 shows that there are several hundreds of Mementos in other web archives that do not implement the Internet Archive's robots.txt policy.

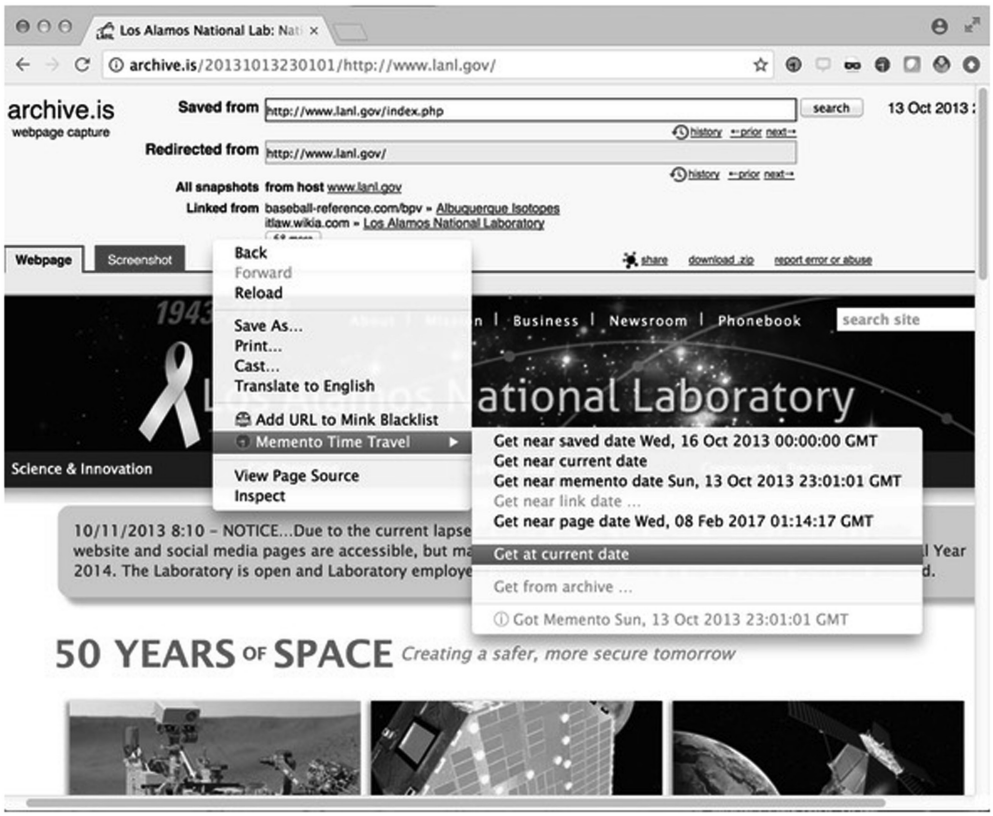

Figure 14.19 Right-clicking in the middle of the Memento to go back to the live web (i.e., from http://archive.is/20131016225948/http://www.lanl.gov/ back to http://www.lanl.gov/).

Memento Quality and Temporal Coherence

Although the gold standard for assessing web archiving quality is still human interaction with a Memento to ensure all embedded resources, links, and functionality are preserved, this level of assessment is clearly not scalable. We have been involved with a range of automated evaluations of the web archiving process, including the Archival Acid Test (Kelly et al., 2014a), which evaluates the capabilities of crawling and playback technology stacks (e.g., the Heritrix crawler (Mohr et al., 2004) and the Wayback Machine playback engine), and assessing Memento damage (Brunelle et al., 2014, 2015) which provides weights to missing embedded resources based on heuristics for determining if the missing resource was 'important'.

Of particular note is our work on 'Temporal Violations', which are combinations of root HTML pages and embedded resources (e.g., images, CSS, JavaScript) that are combined during archival replay to produce a combination that never existed on the live web (Ainsworth et al., 2014; 2015). This can happen when the root HTML page is crawled and the embedded images (for example) are modified in between the time the root HTML page was crawled and the images themselves were crawled.

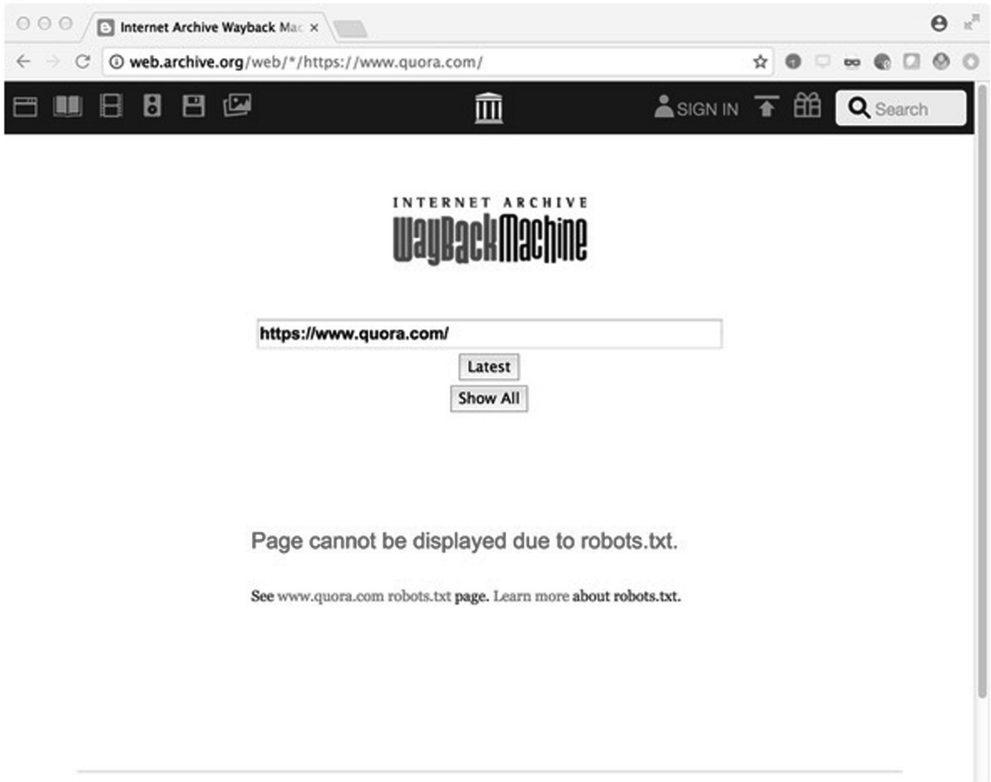

Figure 14.20 The Internet Archive may hold Mementos for https://www.quora.com/ but is blocking them due to the directives found in https://www.quora.com/robots.txt.

```
$ curl --silent -i http://memgator.
   cs.odu.edu/timemap/link/https://
   www.quora.com/ | grep datetime |
   awk '{print $1}' | awk -v FS=/
   '{print $3}' | sort | uniq -c
44 archive.is
3 arquivo.pt
219 wayback.archive-it.org
77 wayback.vefsafn.is
182 webarchive.loc.gov
4 webarchive.nationalarchives.gov.uk
6 webarchive.parliament.uk
```

Figure 14.21 https://www.quora.com/ is not in the Internet Archive but is archived 500+ times in seven other archives.

We defined four categories for discussing the temporal coherence of composite Mementos (root HTML plus embedded resources):

1 Prima Facie Coherent – the combination of embedded resources and a root HTML page can be shown to have existed as presented at the time of crawling (i.e., Memento-Datetime).
2 Prima Facie Violative – the embedded resources can be shown to have been modified since the Memento-Datetime of the root HTML page.
3 Possibly Coherent – the embedded resources have Memento-Datetimes earlier than their root HTML page, and although they cannot be shown to be coherent since embedded resources are typically static, they are possibly coherent.
4 Probably Violative – the embedded resources have Memento-Datetimes later than their root HTML page, and although they cannot be shown to be violative, as root HTML pages are often dynamically generated and modified at a faster rate than embedded resources, they are probably violative.

```
$ curl -I http://web.archive.org/web/20141113140512im_/http://www.lanl.gov/
  _assets/images/lanl-logo-footer.png
HTTP/1.1 200 OK
Server: Tengine/2.1.0
Date: Mon, 13 Feb 2017 05:46:48 GMT
Content-Type: image/png
Content-Length: 8719
Connection: keep-alive
Memento-Datetime: Thu, 13 Nov 2014 14:05:12 GMT
Link: <http://www.lanl.gov/_assets/images/lanl-logo-footer.png>; rel="original",
  <http://web.archive.org/web/timemap/link/http://www.lanl.gov/_assets/images/
  lanl-logo-footer.png>; rel="timemap"; type="application/link-format",
  <http://web.archive.org/web/http://www.lanl.gov/_assets/images/lanl-logo-
  footer.png>; rel="timegate", <http://web.archive.org/web/20120912040310/
  http://www.lanl.gov/_assets/images/lanl-logo-footer.png>; rel="first
  memento"; datetime="Wed, 12 Sep 2012 04:03:10 GMT", <http://web.archive.
  org/web/20141009211301/http://www.lanl.gov/_assets/images/lanl-logo-footer.
  png>; rel="prev memento"; datetime="Thu, 09 Oct 2014 21:13:01 GMT", <http://
  web.archive.org/web/20141113140512/http://www.lanl.gov/_assets/images/lanl-
  logo-footer.png>; rel="memento"; datetime="Thu, 13 Nov 2014 14:05:12 GMT",
  <http://web.archive.org/web/20141122060334/http://www.lanl.gov/_assets/
  images/lanl-logo-footer.png>; rel="next memento"; datetime="Sat, 22 Nov 2014
  06:03:34 GMT", <http://web.archive.org/web/20170204070650/http://www.lanl.
  gov/_assets/images/lanl-logo-footer.png>; rel="last memento"; datetime="Sat,
  04 Feb 2017 07:06:50 GMT"
X-Archive-Orig-last-modified: Tue, 28 Oct 2014 22:12:02 GMT
X-Archive-Orig-content-type: image/png
X-Archive-Orig-date: Thu, 13 Nov 2014 14:03:55 GMT
X-Archive-Orig-content-length: 8719
```

Figure 14.22 The Memento-Datetime and X-Archive-Orig-last-modified headers establish a range of temporal validity.

The combination of Last-Modified and Memento-Datetime headers are critical for determining a range of temporal validity for an embedded resource with respect to a root HTML page in which it appears. Embedded resources, such as images, CSS, and JavaScript, tend to be static files with Last-Modified response headers available (cf. Figure 14.1), and most HTML files are dynamically generated and thus rarely have the Last-Modified response header (cf. Figure 14.2). For example, in the Memento http://web.archive.org/web/20141031175656/https://www.lanl.gov/, with a Memento-Datetime of 'Fri, 31 Oct 2014 17:56:56 GMT', the logo at the bottom of the page (http://web.archive.org/web/20141113140512im_/http://www.lanl.gov/_assets/images/lanl-logo-footer.png) has a Memento-Datetime of 'Thu, 13 Nov 2014 14:05:12 GMT', which means it was crawled after the root HTML page in which it appears. However, as shown in Figure 14.22, the Last-Modified date of the logo as it was when the logo was crawled is echoed in the 'X-Archive-Orig-last-modified' response header (indeed, many of the original HTTP response headers are echoed by the web archive with the prefix 'X-Archive-Orig-').

In Figure 14.23, we see a NOAA page with a Memento-Datetime of January 29, 1999 with an embedded image as the primary content. However, when we dereference the URI-M for the image, we see that the Last-Modified and Memento-Datetime headers (Figure 14.24) have values of April 10, 2003. This combination of HTML page and embedded JPEG never existed on the live web.

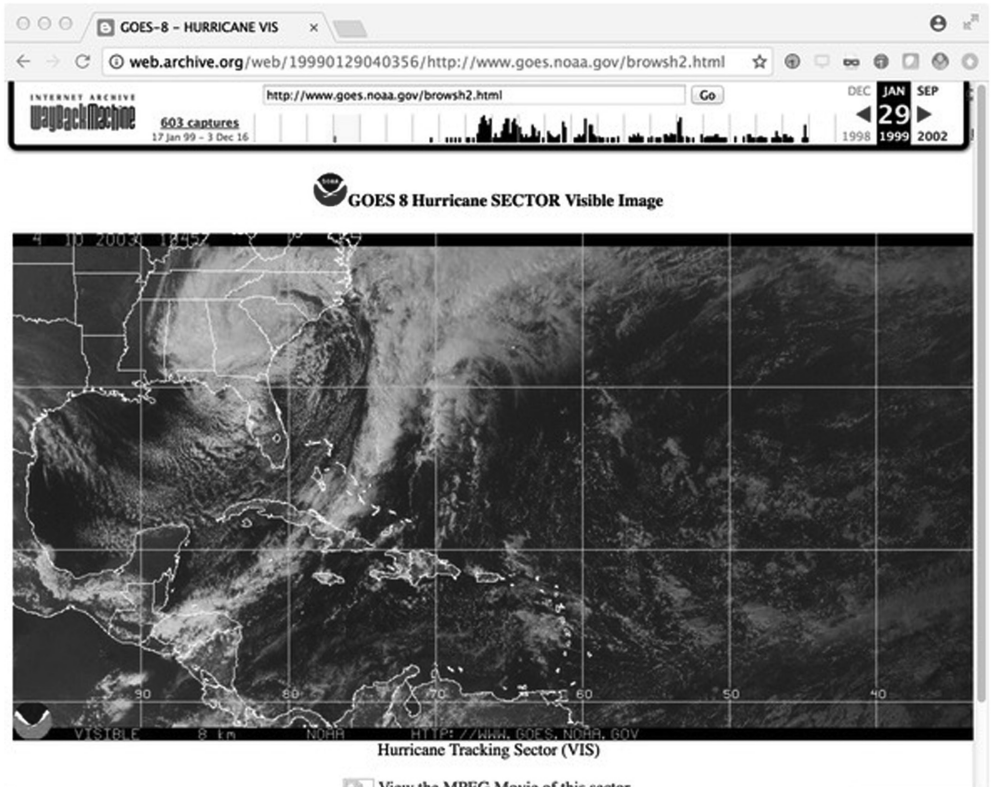

Figure 14.23 Memento http://web.archive.org/web/19990129040356/http://www.goes.noaa.gov/browsh2.html.

In our study of temporal violations (Ainsworth et al., 2015), we found that approximately 76% of composite Mementos were complete (i.e., missing no embedded Mementos), and utilizing additional Memento-enabled web archives could raise that number to 80% complete. More concerning is that 6% of composite Mementos are Prima Facie Violative and 2.5% are Probably Violative. When multiple archives are used, the Probably Violative composite Mementos actually goes up to 5%, in part because not all web archives provide the X-Archive-Orig-last-modified response header (Ainsworth, 2015). So, while multiple archives increase the completeness, they can potentially decrease the temporal coherence. When taken altogether, only 18% of composite Mementos are both complete and Prima Facie Coherent. In summary, web archives probably provide a faithful rendering of the past web approximately only one in five times.

Linking to Archives or the Live Web?

The existence of multiple web archives introduces the question, similar to the 'appropriate copy' problem of reference linking (Caplan and Arms, 1999), of which version to link to when writing HTML: the live web version, or an archived version, and, if an archived version, in which archive? Increased

```
$ curl -IL http://web.archive.org/web/19990129040356/http://www.goes.noaa.
  gov/GIFS/HUVS.JPG
HTTP/1.1 302 Found
Server: Tengine/2.1.0
Date: Tue, 14 Feb 2017 04:02:10 GMT
Content-Type: image/jpeg
Content-Length: 0
Location: /web/20030410203138/http://www.goes.noaa.gov/GIFS/HUVS.JPG
Link: <http://www.goes.noaa.gov:80/GIFS/HUVS.JPG>; rel="original"

HTTP/1.1 200 OK
Server: Tengine/2.1.0
Date: Tue, 14 Feb 2017 04:02:11 GMT
Content-Type: image/jpeg
Content-Length: 141380
Memento-Datetime: Thu, 10 Apr 2003 20:31:38 GMT
Link: <http://www.goes.noaa.gov/GIFS/HUVS.JPG>; rel="original", <http://
  web.archive.org/web/timemap/link/http://www.goes.noaa.gov/GIFS/HUVS.JPG>;
  rel="timemap"; type="application/link-format", <http://web.archive.org/
  web/http://www.goes.noaa.gov/GIFS/HUVS.JPG>; rel="timegate", <http://web.
  archive.org/web/20030410203138/http://www.goes.noaa.gov/GIFS/HUVS.JPG>;
  rel="first memento"; datetime="Thu, 10 Apr 2003 20:31:38 GMT", <http://web.
  archive.org/web/20030602014329/http://www.goes.noaa.gov/GIFS/HUVS.JPG>;
  rel="next memento"; datetime="Mon, 02 Jun 2003 01:43:29 GMT", <http://web.
  archive.org/web/20170201134641/http://www.goes.noaa.gov/GIFS/HUVS.JPG>;
  rel="last memento"; datetime="Wed, 01 Feb 2017 13:46:41 GMT"
X-Archive-Orig-last-modified: Thu, 10 Apr 2003 20:05:23 GMT
X-Archive-Orig-content-type: image/jpeg
X-Archive-Orig-date: Thu, 10 Apr 2003 20:31:33 GMT
X-Archive-Orig-content-length: 141380
X-Archive-Orig-server: Apache/2.0.45 (Unix)
```

Figure 14.24 Prima Facie Violative: the embedded JPEG from Figure 14.23 was actually modified and archived in 2003, not 1999.

interest in link rot ('soft 404s' (Bar-Yossef et al., 2004) or actual 404 HTTP responses) and content drift (the page persists, but the content is no longer relevant to the author's original intent when creating the link) in our scholarly, legal, and other corpora has raised interest in linking directly to Mementos instead of Original URIs (Jones et al., 2016b; Klein et al., 2014; Zittrain et al., 2014). The current practice in sites like Wikipedia is to replace broken links to the live web to point directly to a URI-M in the Internet Archive (AlNoamany, 2014). However, this fails to take advantage of the other public web archives that might also have copies. One could link to an aggregator, but there is a better way that allows providing both URI-R and URI-M values, along with a prefered datetime for use in archives the author of the HTML might not have knowledge of. This approach increases the chances to be able to revisit originally linked content by including information that can be used to look up Mementos in any archive (URI-R and archival datetime) in addition to the key that can only be used in a single archive (URI-M). This fallback mechanism is relevant because the brief history of web archives provides plenty of illustrations that their accessibility can be hindered by a wide range of challenges including technical failures, financial woes, take-down requests, and geo-politically induced censorship.

The Robust Links specification takes advantage of the HTML5 'data-' attribute for extensible metadata fields not otherwise

```
<a href="http://www.lanl.gov/"
data-versionurl="http://archive.
   is/3IEj0"
data-versiondate="2013-10-16">my
   robust link to the LANL home
   page</a>
```

Figure 14.25 Primary link is to URI-R, alternate link to URI-M, and a preferred datetime.

defined in HTML5 (Van de Sompel et al., 2015). For example, if we wanted to link to the URI-R with an alternate link to a URI-M, we could write the HTML shown in Figure 14.25. The URI-R and value in the 'data-versiondate' attribute can be combined for URI lookups in any web archive, for example, using the Memento Protocol.

Similarly, it is possible to reverse the primary and alternate link. Figure 14.26 shows the primary link to an aggregator service that will dynamically determine and redirect the client to the closest available Memento; the URI-R is the alternate link, and the datetime is also provided. If that aggregator is no longer online, the URI-R and datetime values can be combined for discovery in other web archives or aggregators.

Variations of the examples in Figures 14.25 and 14.26 are possible, but all require the desired datetime to be expressed either in the 'data-versiondate' link attribute or in the datePublished or dateModified attributes for the HTML meta tag if the datetime is applicable to the entire page. A demonstration of Robust Links in action can be found in the links of one of our recent papers (Van de Sompel and Nelson, 2015).

```
<a href="http://timetravel.mementoweb.
   org/memento/20131016000000/
   http://www.lanl.gov/"  data-
   originalurl="http://www.lanl.gov/"
   data-versiondate="2013-10-16">my
   robust link to the LANL home
   page</a>
```

Figure 14.26 Primary link is to an aggregator, alternate link to URI-R, and a preferred datetime.

CONCLUSIONS

The Memento Protocol is the de facto standard for integrating the now dozens of publicly accessible web archives, and it works equally well in intranets and private archives. For example, the browsers and tools we at LANL and ODU use to access the Internet Archive and other public web archives are the same tools we use to access our research group's private wiki. The Memento Protocol is an implementation of datetime negotiation in HTTP, fulfilling one of the original design goals of Tim Berners-Lee that was deferred in part because of the lack of temporal semantics in the Unix filesystem. The Memento Protocol provides a standardized, machine-readable mechanism for bi-directional linkage between the current and past webs, where before there had just been an ad hoc set of conventions and archive-specific heuristics for naming and accessing the past web. TimeMaps provide lists of Mementos for an Original Resource, and TimeGates provide the datetime negotiation to the closest available Memento.

In 2006, Julien Masanes expressed a vision of an interconnected, global grid of web archives:

> Such a grid should link Web archives so that they together form one global navigation space like the live Web itself. This is only possible if they are structured in a way close enough to the original Web and if they are openly accessible. (Masanes, 2006, p. 21)

The Memento Protocol achieves this goal by changing web archives from strictly destinations into infrastructure that supports the live web by providing a standardized and integrated versioning for the web indexed by global datetime. In the future when the web transitions from dozens to 100s or even 1,000s of web archives, public and private, the faithful and complete rendering of past web pages will depend on the ability to identify, query, and access the appropriate subset of web archives.

The history of the early twenty-first century cannot be told without significant evidence from web archives. The websites and their contents are fleeting, and are too culturally important to be left to the care of a single web archive. The web is distributed and largely uncoordinated, with interoperability emerging from simple protocols such as HTTP. So too must be our web archives: distributed and largely uncoordinated, with interoperability made possible via time semantics that are part of HTTP and finally transcend the limitations of the Unix operating system.

ACKNOWLEDGMENTS

The Memento Protocol was originally supported by grants from the Library of Congress. Additional research was supported in part by NSF IIS 1009392, NSF IIS 1526700, and the Andrew Mellon Foundation. Numerous people supported the Memento Protocol through software development and installation, many of whom appear below as our co-authors. Figures 14.16 through 14.19 are courtesy of the Los Alamos National Laboratory.

REFERENCES

Ainsworth, S.G. (2015) *Original header replay considered coherent*. [blog] Web Science and Digital Library Research Group. Available at: http://ws-dl.blogspot.com/2015/08/2015-08-28-original-header- replay.html [Accessed 1 Mar. 2018].

Ainsworth, S.G., Nelson, M.L., and Van de Sompel, H. (2015) 'Only one out of five archived web pages existed as presented', in: *HT '15, Proceedings of the 26th ACM Conference on Hypertext and Hypermedia*, [online] Guzelyurt: ACM, pp. 257–266. Available at: https://doi.org/10.1145/2700171.2791044 [Accessed 1 Mar. 2018].

Ainsworth, S.G., Nelson, M.L., and Van de Sompel, H. (2014) *A framework for evaluation of composite memento temporal coherence*. [online] Available at: http://arxiv.org/abs/1402.0928 [Accessed 1 Mar. 2018].

Alam, S., and Nelson, M.L. (2016) 'Memgator – a portable concurrent memento aggregator: Cross-platform CLI and server binaries in Go', in: *JCDL '16. Proceedings of the 16th ACM/IEEE-CS Joint Conference on Digital Libraries*, [online] Newark: ACM, pp. 243–244. Available at: https://doi.org/10.1145/2910896.2925452 [Accessed 1 Mar. 2018].

Alam, S., Nelson, M.L., Van de Sompel, H., Balakireva, L., Shankar, H., and Rosenthal, D.S.H. (2015) 'Web archive profiling through CDX summarization', in: *TPDL '15, Proceedings of Theory and Practice of Digital Libraries*, [online] Poznań: Springer, pp. 3–14. Available at: https://doi.org/10.1007/978-3-319-24592-8_1 [Accessed 1 Mar. 2018].

Alam, S., Nelson, M.L., Van de Sompel, H., and Rosenthal, D.S.H. (2016a) 'Web archive profiling through fulltext search', in: *TPDL '16, Proceedings of Theory and Practice of Digital Libraries*, [online] Hannover: Springer, pp. 121–132. Available at: https://doi.org/10.1007/978-3-319-43997-6_10 [Accessed 1 Mar. 2018].

Alam, S., Nelson, M.L., Van de Sompel, H., Balakireva, L.L., Shankar, H., and Rosenthal, D.S.H. (2016b) 'Web archive profiling through CDX summarization', *International Journal on Digital Libraries*, [online] 17(3):223–228. Available at: https://doi.org/10.1007/s0079 [Accessed 1 Mar. 2018].

AlNoamany, Y. (2014) '*Using Web Archives to Enrich the Live Web Experience Through Storytelling*'. PhD. Old Dominion University, Department of Computer Science.

AlSum, A., Weigle, M.C., Nelson, M.L., and Van de Sompel, H. (2013) 'Profiling web archive coverage for top-level domain and content language', in: *TPDL '13, Proceedings of Theory and Practice of Digital Libraries*, [online] Valetta: Springer, pp. 60–71. Available at: https://doi.org/10.1007/978-3-642-40501-3_7 [Accessed 1 Mar. 2018].

AlSum, A., Weigle, M.C., Nelson, M.L., and Van de Sompel, H. (2014) 'Profiling web archive coverage for top-level domain and content language', *International Journal on Digital Libraries*, [online] 14(3):149–166.

Available at: https://doi.org/10.1007/s00799-014-0118-y [Accessed 1 Mar. 2018].

Bailey, J., Grotke, A., Hanna, K., Hartman, C., McCain, E., Moffatt, C., and Taylor, N. (2013) *Web archiving in the United States: A 2013 survey*. [online] Available at: https://blogs.loc.gov/thesignal/2014/10/results-from-the-2013-ndsa-u-s-web-archiving-survey/ [Accessed 1 Mar. 2018].

Bar-Yossef, Z., Broder, A.Z., Kumar, R., and Tomkins, A. (2004) 'Sic transit gloria telae: Towards an understanding of the web's decay', in: *WWW '04, Proceedings of the 13th International Conference on World Wide Web*, [online] New York: ACM, pp. 328–337. Available at: https://doi.org/10.1145/988672.988716 [Accessed 1 Mar. 2018].

Berners-Lee, T. (1991) *The original http as defined in 1991*. [online] Available at: https://www.w3.org/Protocols/HTTP/AsImplemented.html [Accessed 1 Mar. 2018].

Berners-Lee, T. (1996) *Web architecture: Generic resources*. [online] Available at: http://www.w3.org/DesignIssues/Generic.html [Accessed 1 Mar. 2018].

Berners-Lee, T., Fielding, R., and Frystyk, H. (1996) *Hypertext Transfer Protocol – HTTP/1.0, Internet RFC 1945*. [online] Available at: https://tools.ietf.org/html/rfc1945 [Accessed 1 Mar. 2018].

Bornand, N.J., Balakireva, L., and Van de Sompel, H. (2016) 'Routing memento requests using binary classifiers', in: *JCDL '16, Proceedings of the 16th ACM/IEEE-CS Joint Conference on Digital Libraries*, [online] Newark: ACM, pp. 63–72. Available at: https://doi.org/10.1145/2910896.2910899 [Accessed 1 Mar. 2018].

Brunelle, J.F., Kelly, M., SalahEldeen, H., Weigle, M.C., and Nelson, M.L. (2014) 'Not all mementos are created equal: Measuring the impact of missing resources', in: *JCDL '14, Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries*, [online] London: ACM, pp. 321–330. Available at: https://doi.org/10.1109/JCDL.2014.6970187 [Accessed 1 Mar. 2018].

Brunelle, J.F., Kelly, M., SalahEldeen, H., Weigle, M.C., and Nelson, M.L. (2015) 'Not all mementos are created equal: Measuring the impact of missing resources', *International Journal on Digital Libraries*, [online] 16(3–4):283–301. Available at: https://doi.org/10.1007/s00799-015-0150-6 [Accessed 1 Mar. 2018].

Brunelle, J.F., Nelson, M.L., Balakireva, L., Sanderson, R., and Van de Sompel, H. (2013) 'Evaluating sitestory with the ApacheBench tool', in: *TPDL '13, International Conference on Theory and Practice of Digital Libraries*, [online] Valetta: Springer, pp. 204–215. Available at: https://doi.org/10.1007/978-3-642-40501-3_20 [Accessed 1 Mar. 2018].

Callan, J. (2002) 'Distributed Information Retrieval' in: Croft W.B. (ed) Advances in Information Retrieval. The Information Retrieval Series, vol 7. Springer, Boston, MA, [online]. Available at: https://doi.org/10.1007/0-306-47019-5_5 [Accessed 1 Mar. 2018].

Caplan, P., and Arms, W.Y. (1999) 'Reference linking for journal articles', *D-Lib Magazine*, [online] 5(7/8). Available at: https://doi.org/10.1045/july99-caplan [Accessed 1 Mar. 2018].

Crocker, D.H. (1982) *Standard for the Format of ARPA Internet Text Messages, Internet RFC-822*. [online] Available at: https://tools.ietf.org/html/rfc822 [Accessed 1 Mar. 2018].

Fielding, R., Gettys, J., Mogul, J., Frystyk, H., Masinter, L., Leach, P., and Berners-Lee, T. (1999) *Hypertext Transfer Protocol – HTTP/1.1, Internet RFC-2616.* [online] Available at: https://tools.ietf.org/html/rfc2616 [Accessed 1 Mar. 2018].

Fielding, R.T. (2000) '*Architectural Styles and the Design of Network-based Software Architectures*'. PhD. University of California, Irvine, Department of Computer Science.

Fuhrig, L.S. (2014) *Tracking down the elusive 'treasure house for learning'*. [blog] Smithsonian Institution Archives blog. Available at: https://siarchives.si.edu/blog/tracking-down-elusive-'treasure-house-learning' [Accessed 1 Mar. 2018].

Gomes, D., Miranda, J., and Costa, M. (2011) 'A survey on web archiving initiatives', in: *TPDL '11, Proceedings of Theory and Practice of Digital Libraries*, [online] Berlin: Springer, pp. 408–420. Available at: https://doi.org/10.1007/978-3-642-24469-8_41 [Accessed 1 Mar. 2018].

Jacobs, I., and Walsh, N. (2004) *Architecture of the world wide web, volume one. Technical Report W3C Recommendation 15 December*

2004, W3C. [online] Available at: https://www.w3.org/TR/webarch/ [Accessed 1 Mar. 2018].

Jones, S.M., Nelson, M.L., Shankar, H., and Van de Sompel, H. (2014) *Bringing web time travel to mediawiki: An assessment of the memento mediawiki extension*. [online] Available at: http://arxiv.org/abs/1405.2330 [Accessed 1 Mar. 2018].

Jones, S.M., Van de Sompel, H., and Nelson, M.L. (2016a) 'Avoiding spoilers: Wiki time travel with Sheldon Cooper', *International Journal on Digital Libraries*, [online] 19(1):77–93. Available at: https://doi.org/10.1007/s00799-016-0200-8 [Accessed 1 Mar. 2018].

Jones, S.M., Van de Sompel, H., Shankar, H., Klein, M., Tobin, R., and Grover, C. (2016b) 'Scholarly context adrift: Three out of four uri references lead to changed content', *PloS One*, [online] 11(12):e0167475. Available at: https://doi.org/10.1371/journal.pone.0171057 [Accessed 1 Mar. 2018].

Kelly, M., Nelson, M.L., and Weigle, M.C. (2014a) 'The archival acid test: Evaluating archive performance on advanced HTML and JavaScript', in: *JCDL '14, Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries*, [online] London: ACM, pp. 25–28. Available at: https://doi.org/10.1109/JCDL.2014.6970146 [Accessed 1 Mar. 2018].

Kelly, M., Nelson, M.L., and Weigle, M.C. (2014b) 'Mink: Integrating the live and archived web viewing experience using web browsers and Memento', in: *JCDL '14, Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries*, [online] London: ACM, pp. 267–276. Available at: https://doi.org/10.1109/JCDL.2014.6970229 [Accessed 1 Mar. 2018].

Klein, M., Van de Sompel, H., Sanderson, R., Shankar, H., Balakireva, L., Zhou, K., and Tobin, R. (2014) 'Scholarly context not found: One in five articles suffers from reference rot', *PloS One*, [online] 9(12):e115253. Available at: https://doi.org/10.1371/journal.pone.0115253 [Accessed 1 Mar. 2018].

Masanes, J. (2006) *Web Archiving*. Springer, Berlin, Heidelberg.

Mohr, G., Kimpton, M., Stack, M., and Ranitovic, I. (2004) 'Introduction to heritrix, an archival quality web crawler', in: *IWAW '04, 4th International Web Archiving Workshop*, [online] Bath. Available at: https://web.archive.org/web/20170809135759/http://iwaw.europarchive.org/04/Mohr.pdf [Accessed 1 Mar. 2018].

Negulescu, K.C. (2010) *Web archiving @ the internet Archive; presentation at the 2010 Digital Preservation Partners Meeting*. [online] Available at: http://www.digitalpreservation.gov/meetings/documents/ndiipp10/NDIIPP-072110FinalIA.ppt [Accessed 1 Mar. 2018].

Nelson, M.L. (2011) *Memento-Datetime is not Last-Modified*. [blog] Web Science and Digital Library Research Group. Available at: http://ws-dl.blogspot.com/2010/11/2010-11-05-memento-datetime-is-not-last.html [Accessed 1 Mar. 2018].

Nelson, M.L. (2013) *Right-click to the past – Memento for Chrome*. [blog] Web Science and Digital Library Research Group. Available at: http://ws-dl.blogspot.com/2013/10/2013-10-14-right-click-to-past-memento.html [Accessed 1 Mar. 2018].

Ritchie, D., and Thompson, K. (1974) 'The UNIX time-sharing system', *Communications of the ACM*, [online] 17(7):365–375. Available at: https://doi.org/10.1145/361011.361061 [Accessed 1 Mar. 2018].

Rossi, A. (2016) *Robots.txt Files and Archiving .gov and .mil Websites*. [blog] Internet Archive Blogs. Available at: https://blog.archive.org/2016/12/17/robots-txt-gov-mil-websites/ [Accessed 1 Mar. 2018].

Sanderson, R. (2012) 'Global web archive integration with Memento', in: *JCDL '06, Proceedings of the 12th ACM/IEEE Joint Conference on Digital Libraries*, [online] Washington: ACM, pp. 379–380. Available at: https://doi.org/10.1145/2232817.2232900 [Accessed 1 Mar. 2018].

Sanderson, R., Shankar, H., Ainsworth, S., McCown, F., and Adams, S. (2011) 'Implementing time travel for the web', *Code4Lib Journal*, [online] 13. Available at: http://journal.code4lib.org/articles/4979 [Accessed 1 Mar. 2018].

Tweedy, H., McCown, F., and Nelson, M.L. (2013) 'A Memento web browser for iOS', in: *JCDL '013, Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries*, [online] Indianapolis: ACM, pp. 371–372. Available at: https://doi.org/10.1145/2467696.2467764 [Accessed 1 Mar. 2018].

Van de Sompel, H., and Nelson, M.L. (2015) 'Reminiscing about 15 years of interoperability efforts', *D-Lib Magazine*, [online] 21(11/22). Available at: https://doi.org/10.1045/november2015-vandesompel [Accessed 1 Mar. 2018].

Van de Sompel, H., Nelson, M.L., and Sanderson, R. (2013) *HTTP framework for time-based access to resource states – Memento, Internet RFC 7089*. [online] Available at: http://tools.ietf.org/html/rfc7089 [Accessed 1 Mar. 2018].

Van de Sompel, H., Nelson, M.L., Sanderson, R., Balakireva, L.L., Ainsworth, S., and Shankar, H. (2009). *Memento: Time Travel for the Web*. [online] Available at: http://arxiv.org/abs/0911.1112 [Accessed 1 Mar. 2018].

Van de Sompel, H., Sanderson, R., Nelson, M.L., Balakireva, L.L., Shankar, H., and Ainsworth, S. (2010) 'An HTTP-based versioning mechanism for linked data', in: *LDOW '010, Proceedings of the Linked Data on the Web Workshop,* [online] Raleigh: CEUR 628. Available at: http://ceur-ws.org/Vol-628/ldow2010_paper13.pdf [Accessed 1 Mar. 2018].

Van de Sompel, H., Shankar, H., Wincewicz, R., and Nelson, M.L. (2015) *Robust links – link decoration*. [online] Available at: http://robustlinks.mementoweb.org/spec/ [Accessed 1 Mar. 2018].

Zittrain, J., Albert, K., and Lessig, L. (2014) 'Perma: Scoping and addressing the problem of link and reference rot in legal citations', *Legal Information Management*, [online] 14(02):88–99. Available at: https://doi.org/10.1017/S1472669614000255 [Accessed 1 Mar. 2018].