Only One Out of Five Archived Web Pages Existed as Presented

Scott G. Ainsworth
Old Dominion University
Norfolk, VA, USA
sainswor@cs.odu.edu

Michael L. Nelson
Old Dominion University
Norfolk, VA, USA
mln@cs.odu.edu

Herbert Van de Sompel
Los Alamos National Laboratory
Los Alamos, NM, USA
herbertv@lanl.gov

ABSTRACT

When a user retrieves a page from a web archive, the page is marked with the acquisition datetime of the root resource, which effectively asserts "this is how the page looked at a that datetime." However, embedded resources, such as images, are often archived at different datetimes than the main page. The presentation appears temporally coherent, but is composed from resources acquired over a wide range of datetimes. We examine the completeness and temporal coherence of composite archived resources (composite mementos) under two selection heuristics. The completeness and temporal coherence achieved using a single archive was compared to the results achieved using multiple archives. We found that at most 38.7% of composite mementos are both temporally coherent and that at most only 17.9% (roughly 1 in 5) are temporally coherent and 100% complete. Using multiple archives increases mean completeness by 3.1–4.1% but also reduces temporal coherence.

Categories and Subject Descriptors

H.3.7 [**Information Storage and Retrieval**]: Digital Libraries

General Terms

Design, Experimentation, Standardization

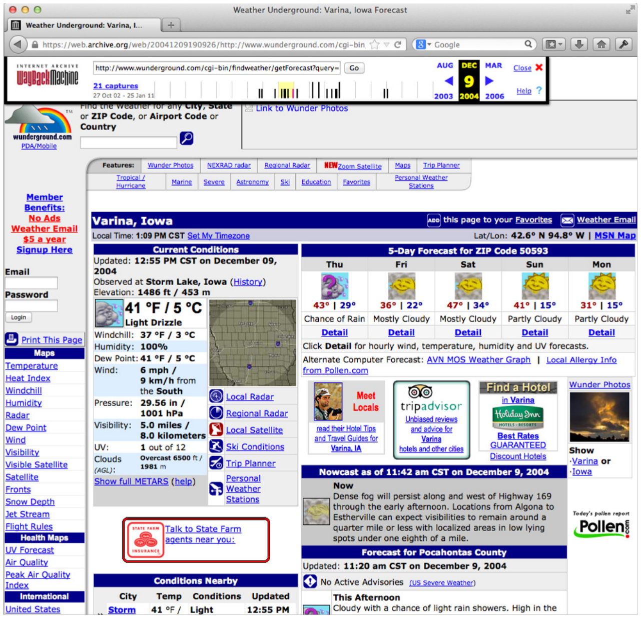
Keywords

Digital Preservation, Temporal Coherence, Web Architecture, Web Archiving, Resource Versioning, HTTP, Memento, RFC 7089

1. INTRODUCTION

When a user retrieves an archived page from a web archive such as the Internet Archive [28], the page is marked with a Memento-Datetime (the datetime the archived resource was acquired) of the root resource, which effectively asserts "this page appeared in this state at the datetime indicated." The page is recomposed from archived resources in an attempt to provide an appearance as similar as possible to that which the user would have experienced had the original page been visited on the indicated Memento-Datetime. The presentation appears as real as those from the live web. However, appearances can be deceiving.

Figure 1 shows a presentation for *wunderground.com* on Thursday 2004-12-09 19:09:26 (all times GMT). The large radar image near the page center shows the weather in Varina, Iowa, USA was clear and sunny. A closer look tells a different story. The daily image for Thursday shows cloudy with a chance of rain and the rest of the daily images are partly cloudy. These discrepancies indicate temporal incoherence between the archived root and the embedded archived resources.


Weather Underground: Varina, Iowa Forecast

https://web.archive.org/web/20041209190926/http://www.wunderground.com/cgi-bin

http://www.wunderground.com/cgi-bin/findweather/getForecast?query=

21 captures
27 Oct 02 - 25 Jan 11

AUG DEC MAR
9
2003 2004 2006

Find the weather for any City, State or ZIP Code, or Airport Code or Country

Varina, Iowa

Local Time: 1:09 PM CST Set My Timezone

Lat/Lon: 42.6° N 94.8° W | MSN Map

Current Conditions

Updated: 12:55 PM CST on December 09, 2004
Observed at Storm Lake, Iowa (History)
Elevation: 1486 ft / 453 m

41 °F / 5 °C
Light Drizzle

Windchill: 37 °F / 3 °C
Humidity: 100%
Dew Point: 41 °F / 5 °C
Wind: 6 mph / 9 km/h from the South
Pressure: 29.56 in / 1001 hPa
Visibility: 5.0 miles / 8.0 kilometers
UV: 1 out of 12
Clouds (AGL): Overcast 6500 ft / 1981 m

5-Day Forecast for ZIP Code 50593

Thu	Fri	Sat	Sun	Mon
43° | 29°	36° | 22°	47° | 34°	41° | 15°	31° | 15°
Chance of Rain	Mostly Cloudy	Mostly Cloudy	Partly Cloudy	Partly Cloudy
Detail	Detail	Detail	Detail	Detail

Click **Detail** for hourly wind, temperature, humidity and UV forecasts.

Alternate Computer Forecast: AVN MOS Weather Graph | Local Allergy Info from Pollen.com

Nowcast as of 11:42 am CST on December 9, 2004

Now
Dense fog will persist along and west of Highway 169 through the early afternoon. Locations from Algona to Estherville can expect visibilities to remain around a quarter mile or less with localized areas in low lying spots under one eighth of a mile.

Forecast for Pocahontas County

Updated: 11:20 am CST on December 9, 2004

No Active Advisories (US Severe Weather)

This Afternoon
Cloudy with a chance of light rain showers. High in the

Conditions Nearby

City	Temp	Conditions	Updated
Storm	41 °F /	Light	12:55 PM

Figure 1: Wunderground Composite Memento, Memento-Datetime 2004-12-09 19:09:26

An archived web page is a composite resource, generally a composition of a root HTML page, images, Javascripts, stylesheets, and other resources. As shown in Figure 2, although the *wunderground.com* root page itself was acquired on 2004-12-09, the embedded resources were acquired from −20 days before to +9 months after the root. Many archived resources are missing embedded resources. Some of these are apparent (as indicated at the top of Figure 2), but many are not. Web archive user interfaces attempt to create a smooth simulation of browsing the past web. Thus, the temporal incoherence shown in Figure 2 and hidden incompleteness are mostly ignored by these interfaces.

HT'15, September 1–4, 2015, Güzelyurt, TRNC, Cyprus
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-3395-5/15/09 . . . $15.00
DOI: http://dx.doi.org/10.1145/2700171.2791044.

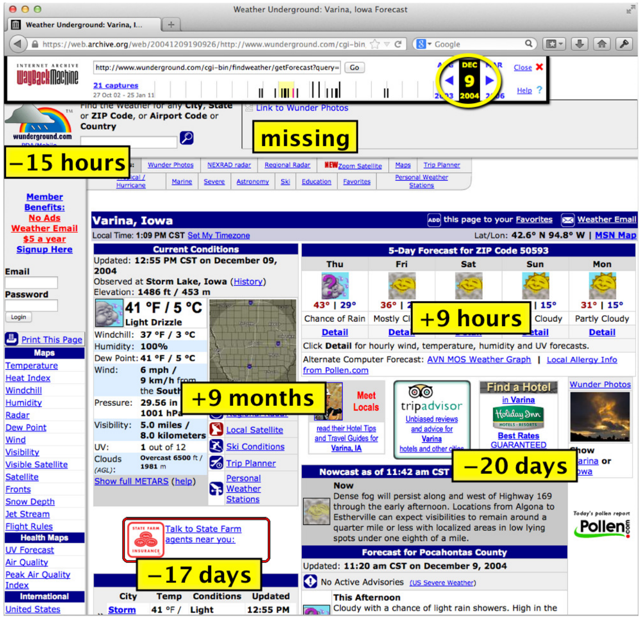


Figure 2: Wunderground Composite Memento, Memento-Datetime 2004-12-09 19:09:26

Public web archives now hold hundreds of billions of archived web pages. Indeed, as of May 25, 2015, the Internet Archive claims 479 billion pages [21]. These web archives have become trusted sources. For example, after the shoot down of Malaysian Flight 17, Arthur Bright, writing for the Christian Science Monitor, cited the Internet Archive when arguing that the flight had been shot down by pro-Russian rebels [10]. Knowing what a web site previously contained is also important in litigation and other legal proceedings. Howell [19] addresses these uses with an emphasis on the meeting for rules of evidence and admissibility requirements.

In the data we studied, it is common for the temporal spread between the oldest and newest captures to be weeks, months, and sometimes a year or more. Unexpected though were spreads exceeding five years; a few even exceed ten years. The difference between expectations and the data, along with our work on Memento [34], motivates us to explore archive completeness and temporal coherence of archived web pages. Our previous studies of existing web archive content examined the coverage on a broad scale [1] and the temporal drift that occurs as web archives are browsed [3]. This paper examines the completeness and temporal coherence of archived composite resources with a focus on these questions:

- How prevalent is temporal incoherence?
- Current web archive user interfaces enable access to a single archive's resources. If multiple archives are used, is temporal coherence improved?
- The *best* memento for an embedded resource can be selected using many different heuristics. Currently, most web archives use the Minimum Distance heuristic (see 4.2.1). How do other heuristics compare?

2. RELATED WORK

Large-scale web archiving requires resolution of issues and approaches on several axes. Although somewhat dated, Masanès [25] is an excellent introduction and covers a broad range of web archiving topics. Of significance to this research are the technical aspects of acquisition, quality, and completeness. An area not addressed by Masanès is standardized access to archived resources, which Van de Sompel et al. [35] addressed with IETF RFC 7089 [34].

2.1 Acquisition

Acquisition is the technical means of bringing content into a web archive. *Client-side* archiving works by emulating web users following links; Heritrix [27] is a widely-used tool. Most client-side archiving suffers from two major limitations: first, only linked resources are captured; second, root and embedded resources are commonly acquired at different times. (Some archives, such as WebCite [17], attempt to prevent the second.) *Transactional* archiving [16, 18], such as *SiteStory* [12], is specifically designed to overcome client-side limitations by inserting the acquisition process between users and the data source. Unique request-response pairs are archived, including requests for resources that are not linked and might not be discovered by client-side archiving. *Server-side* archiving makes a direct copy of the content from the server, bypassing HTTP altogether. This is a common approach for content management systems and wikis. Although conceptually simple, access to the resulting server-side archive can be difficult, requiring different URIs and navigational structures than the original. This can be mitigated by implementing the Memento Protocol, as Jones, et al. [20] did for MediaWiki.

2.2 Access

Until recently, there were neither standard methods nor a standard protocol for access to archived resources. Each archive provided (and still provides) a unique user interface (UI), such as the Wayback Machine [33] shown in Figures 1 and 2, for access to the archive's resources. In general, UI access to archives starts with a user-selected URI and datetime, after which the archive allows the user to simply click links to browse the collection.

Van de Sompel et al. addressed the lack of standards with Memento [35, 34]. Memento is an HTTP-based protocol that bridges web archives with current resources. Each original resource has zero or more archived resources (mementos) which encapsulate its state at various acquisition times. Memento provides a standard protocol for identifying and dereferencing mementos through datetime negotiation. Similar to the way clients use HTTP content negotiation, clients use datetime negotiation to request mementos for original resources by datetime.

2.3 Quality and Completeness

In general, quality is defined as fitting a particular use; objectively, it is defined as meeting measurable characteristics. Our examination of web archive content is concerned with the latter. For web archives, quality issues stem from difficulties inherent in acquiring content over HTTP [25]: content can be temporarily unavailable, leaving coverage gaps; web content changes more frequently than archival crawls can occur, creating temporal gaps; embedded resources may change after root resource acquisition but before the embedded resources themselves have been acquired, leading to temporal incoherence.

2.3.1 *Completeness (Coverage)*

When crawling to acquire content, the tradeoffs required and conditions encountered routinely lead to incomplete coverage. For example, the archive may not have the resources required to acquire and store all desired content; thus, only high priority content is crawled and stored. Desired content may not be available at crawl time due to server downtime or network disruption. The combination of compromises and resource unavailability create undesired, undocumented gaps in the archive.

Although much has been written on the technical, social, legal, and political issues of web archiving, little detailed research has been conducted on the coverage, completeness, and coherence of existing holdings. Day [14] surveyed web archives while investigating the methods and issues associated with web archiving, but did not address coverage. Thelwall touched on coverage when he addressed international bias in the Internet Archive [32], but did not directly address how much of the Web is covered. McCown and Nelson addressed coverage [26], but their research was limited to search engine caches. Ben Saad et al. [8, 7] addressed qualitative completeness through change detection to identify and archive important changes. Leveraging the Memento Protocol and pilot infrastructure, Ainsworth et al. [1] showed that 35–90% of publicly-accessible URIs have at least one publicly-accessible archived copy, 17–49% have two to five copies, 1–8% have six to ten copies, and 8–63% at least ten copies. The number of URI copies varies as a function of time, but only 14.6–31.3% of URIs are archived more than once per month.

The ad-hoc nature of web archiving causes many composite mementos to be incomplete. Brunelle et al. [11] studied the value of missing embedded resources on user perception of the damage caused. In this study, 19.7%–23.9% of composite mementos were found to be incomplete.

2.3.2 *Temporal Coherence*

Spaniol et al. define temporal coherence as meaning that a web archive's contents appear to be "as of" timepoint x or within interval $[x; y]$ [31]. The same constraints and conditions that negatively impact completeness also affect temporal coherence. Spaniol et al. [30] note that crawls may span hours or days, increasing the risk of temporal incoherence especially for large sites. Thus, the simple "as of" timepoint x or within interval $[x; y]$ requirement is *impossible to achieve* in a crawler-based paradigm. (It is not impossible in a transactional archiving or on-demand approach.) Also introduced is a model for identifying coherent sections of archives, which provides a measure of quality, and a crawling strategy which helps minimize temporal incoherence in web site captures. Spaniol et al. [31] also develop crawl and site coherence visualizations. Spaniol et al.'s work, while presenting a measure of quality, addresses the quality of new acquisition crawls; the quality of existing holdings is not addressed.

Denev et al. also address the quality of new acquisition crawls with the SHARC framework [15], which introduced a stochastic notion of *sharpness* (a quality measure). Site changes are modeled as Poisson processes with page-specific change rates, which allows reasoning on the expected sharpness of an acquisition crawl. The authors propose crawl-time quality assessments and page revisits to improve the quality of future crawls.

Ben Saad et al. [9] show that different methods and measures are required *a priori* and *a posterior*, that is during acquisition and post-acquisition respectively. Like Denev et al. [15], the *a priori* solution is designed to optimize the crawling process for archival quality. The *a posteriori* solution uses metadata collected by the *a priori* process to direct the user to the most coherent archived versions. Combining *a priori* and *a posterior* measures and methods into a single solution is shown to produce the most coherent result.

The above research shares a common thread: evaluation and control of completeness and temporal coherence during acquisition with the goal of improving future archive holdings. Our research focuses on the temporal quality of *existing* holdings.

One temporal quality issue we have documented is temporal drift as users navigate from page to page using web archives. The drift is subtle, but can be observed by examining Memento-Datetime. We examined temporal drift under two target datetime policies [3]. The sliding policy emulated web archive user interfaces by allowing the target datetime to shift; the sticky policy held the target datetime to the user's originally selected value. Although some drift is inevitable due to the sparse nature of web archive collections, we found that the sticky policy reduced median drift by 30 days compared to the sliding policy.

Continuing our temporal coherence focus, the present research addresses temporal coherence of composite resources under two heuristics and when embedded mementos are selected from the same archive as a web page or from multiple archives.

3. COMPOSITE MEMENTOS

3.1 Original Resources and Mementos

As mentioned in Subsection 2.2, Memento [34] extends HTTP to provide standard protocol for recognizing and accessing archived resources using datetime negotiation. Each original resource (URI-R) has zero or more mementos (URI-M$_i$), that encapsulate the URI-R's state at times t_i. A timemap (URI-T) provides a list of mementos for a URI-R. Each archived URI-R will have a timemap for each archive holding mementos for the URI-R.

Memento has become the de facto standard for web archive interoperability. Some web archives support Memento natively (Memento support was added to version 1.6 of the Wayback Machine in 2011), and for others the Los Alamos National Laboratory Research Library Prototype Team and Old Dominion University Computer Science Web Science and Digital Libraries Group have developed Memento proxies. The proxies allow the Memento aggregator to simultaneously access native and proxied Memento archives [6]. This study did not use the aggregator. All timemaps and mementos were collected directly from archives that support the Memento Protocol natively or by proxy.

3.2 Composite Mementos

Given a *root URI-R* (rURI-R), which contains *embedded URI-Rs* (eURI-Rs), which may contain eURI-Rs, etc., a Composite Memento comprises mementos for the rURI-R (rURI-M) and for all eURI-Rs (eURI-Ms); that is, all the URI-Ms needed to recompose the the original representation. Figure 3 shows a tree representation of a composite memento[1]. For HTML, a composite memento

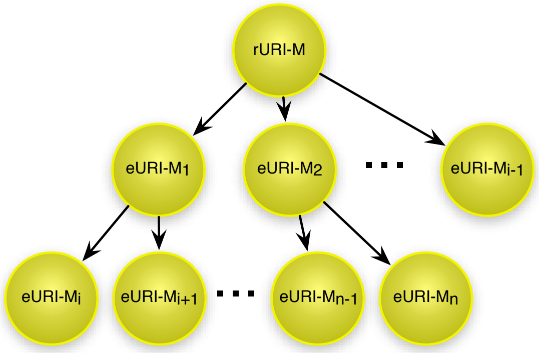


Figure 3: Composite Memento Tree

[1] Composite mementos are actually directed graphs, but in this context can be treated as trees without loss of generality.

Table 1: Sample Mementos from `http://www.rpgdreamers.com/`

URI-R	Last-Modified	Memento-Datetime	Delta	Root?	Bracket?	Coherence	Color
`http://www.rpgdreamers.com/`	N/A	2006-11-10 04:37:51	–	Y	–	C	Black
`.../aslmain.js`	2006-04-25 14:03:39	2006-11-10 23:44:02	–	N	Y	C	Green
`.../home_exp_leader_728x90_468x60.js`	2006-11-11 00:30:18	2006-11-11 01:02:19	19.9h	N	N	V	Red
`http://www.rpgdreamers.com/`	N/A	2006-12-06 08:04:42	–	Y	–	C	Black
`.../aslmain.js`	2006-04-25 14:03:39	2006-12-07 01:07:32	–	N	Y	C	Green
`.../home_exp_leader_728x90_468x60.js`	2006-12-05 17:54:31	2006-12-05 18:30:00	13.6h	N	N	PC	Blue
`.../show_ads.js`	N/A	2006-12-06 12:00:00	3.9h	N	N	PV	Yellow
`http://www.rpgdreamers.com/`	N/A	2007-01-01 21:31:16	–	Y	–	C	Black
`.../aslmain.js`	2006-04-25 14:03:39	2007-01-02 16:23:01	–	N	Y	C	Green
`.../home_exp_leader_728x90_468x60.js`	2006-12-09 14:41:27	2006-12-09 15:04:45	23.3d	N	N	PC	Blue
`.../show_ads.js`	N/A	2007-01-05 12:00:00	86.5d	N	N	PV	Yellow

generally includes an archived web page and archived embedded resources (e.g., images and stylesheets).

Recomposing a composite memento follows a process analogous to a web browser rendering a web page. The process starts with a rURI-R and target datetime, which are use to select a rURI-M. The rURI-M is dereferenced to retrieve its representation from the web archive. Recomposition continues recursively for each eURI-R until eURI-Ms have been retrieved for every rURI-R. It is likely that not every eURI-R will have been archived; these are considered missing.

3.3 Temporal Coherence

We define an eURI-M as temporally coherent with respect to an rURI-M when it can be shown that the eURI-M's representation existed at the rURI-M's Memento-Datetime [4]. However, temporal coherence is more nuanced than simply coherent or not. Determining the coherence state of an eURI-M requires examining the relationship between rURI-M using Memento-Datetime, Last-Modified datetime, and possibly content equality and similarity. Relationships between eURI-Ms and rURI-Ms form patterns. An example pattern, which we call [4] the *Two-Memento Bracket Pattern (2B)* is shown in Figure 4. Here, $\mathcal{L}$ represents a Last-Modified datetime and $\mathcal{T}$ is a Memento-Datetime. The horizontal axis represents time and the URI-Ms are represented as diamonds plotted at the URI-M's Memento-Datetime. The black diamond is the rURI-M; the left and right diamonds are two consecutive eURI-Ms (subscripts L and R indicate left and and right respectively). Note that the eURI-M$_R$ Last-Modified, $\mathcal{L}_{i,j}$, and its Memento-Datetime, $\mathcal{T}_{i,j}$, time *Bracket* the rURI-M's Memento-Datetime, $\mathcal{T}_0$, as shown by the thin green line. This time bracket shows that the eURI-M$_R$ representation existed when rURI-M was acquired. Thus, eURI-M$_R$ is temporally Coherent with respect to rURI-M.

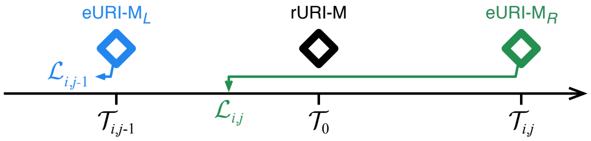


Figure 4: Two-Memento Bracket Pattern (2B)

The mementos in Table 1 are a subset of the mementos needed to recompose three `http://www.rpgdreamers.com/` composite mementos. Figure 5 is a scatter plot of these mementos. Each Composite Memento is a single row with diamonds representing individual rURI-Ms and eURI-Ms. The vertical axis is time. The horizontal axis is the delta between the rURI-M Memento-Datetime and the eURI-M Memento-Datetime, with the scale in days (d). For datetimes outside the two-day range, the delta is shown in the plot area. Composite Mementos are positioned vertically based on their rURI-M Memento-Datetime. Black diamonds are rURI-Ms. eURI-Ms are represented by colored diamonds with the color indicating there coherence state as defined by Ainsworth et al. [4]:

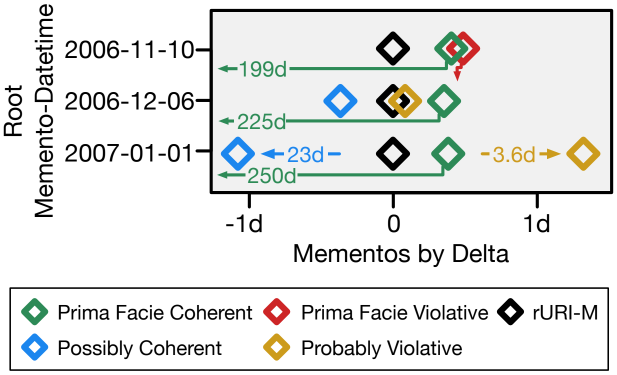


Figure 5: Scatter Plot Sample

Prima Facie Coherent (green). The eURI-M was acquired after the rURI-M Memento-Datetime and its Last-Modified datetime is on or before the rURI-M Memento-Datetime, bracketing rURI-M Memento-Datetime. Therefore, the eURI-M's representation existed at the time the rURI-M was acquired.

Prima Facie Violative (red). The eURI-M was acquired after the rURI-M and its Last-Modified datetime is after the rURI-M Memento-Datetime. Therefore, the eURI-M's representation did not exist at the time the rURI-M was acquired.

Possibly Coherent (blue). The eURI-M was acquired before the rURI-M. Therefore, its representation could have existed at the time the rURI-M was acquired.

Probably Violative (yellow). The eURI-M was acquired after the rURI-M Memento-Datetime and lacks a Last-Modified datetime. Without Last-Modified, the eURI-M cannot be considered Prima Facie Violative or Prima Facie Coherent. Finally, unlike Possibly Coherent, there is no evidence that the eURI-M's representation existed at rURI-M Memento-Datetime.

Table 2 contains temporal coherence statistics for the three sample composite mementos discussed above. The composite mementos have 37 or 39 eURI-Rs each, 72.5%–78.9% of which are Prima

Table 2: Sample Composite Memento Statistics

Description	’06-11-10	’06-12-06	’07-01-01	Mean
eURI-Ms	37	39	39	38.3
Coherent	78.9%	72.5%	72.5%	74.6%
Possibly Coherent	2.6%	12.5%	12.5%	9.3%
Probably Violative	0.0%	2.5%	2.5%	1.7%
Violative	7.9%	0.0%	0.0%	2.5%
Missing	10.5%	12.5%	12.5%	11.9%

Facie Coherent. Another 2.6–12.5% are Possibly Coherent. Prima Facie Violative (0–7.9%), Probably Violative (0–2.5%), and missing mementos (10.5–12.5%) account for the remainder. These are individual statistics for three mementos. (Subsection 3.3 and Table 9 present statistics for the entire sample set.)

The full scatter plot shown in Figure 6 reveals interesting patterns. Many embedded resources are archived infrequently; these create right-to-left diagonals. Also, some resources are archived much later than the root resources, which results in unexpectedly large deltas. For example, the 2005-03-10 03:21:27 composite memento (indicated by the dashed green line) includes several embedded resources that were archived well after the root. Of particular note is `media.gif`[2], which had only a single URI-M with Memento-Datetime 2011-06-06 20:10:22 and was acquired nearly seven years after the rURI-M. Full details for all sample URIs cannot be made available within this paper. We will provide the scatter plots and statistics for the entire sample set online at `http://coherence.cs.odu.edu`.

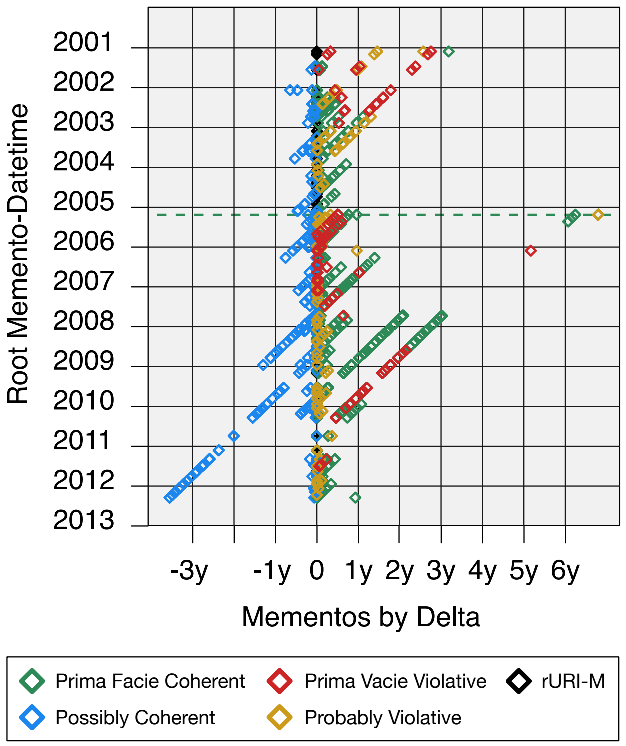


Figure 6: Full Scatter Plot

[2] `http://www.rpgdreamers.com/rpgworld/logo/media.gif`

Table 3: Archival Rates

Sample	This study	Jan’13 [1]	Jan’11 [3]
DMOZ	97.0%	95.2%	79%
Search Engine	55.2%	26.4%	19%
Bitly	28.3%	23.5%	16%
Delicious	95.0%	91.9%	68%
Aggregate	68.9%	59.4%	46%

4. EXPERIMENT

4.1 Sample URIs

Building on previous work, we used the same four URI sample sets (from DMOZ, search engine, Bitly, and Delicious) as in [1]. Each sample contains 1,000 randomly-selected URI-Rs for 4,000 URI-Rs total. Random selection details can be found in [2].

The percentage of sample URI-Rs found to be archived at the time of the experiment is shown in Table 3. There are several significant differences from our 2011 results [1, 36] and January 2013 results [3]. The 2013 results noted increased holdings availability provided by the Internet Archive [22], which helped increase archival rates. This experiment again found that archival rates increased. Most significant difference from [1] and [3] is the use of additional archives, including seven with primary languages other than English.

4.2 Parameters

The experiment recomposed four variants of each sample composite memento. Each variant used one of two memento selection heuristics and either a single archive or multiple archives.

4.2.1 Heuristics

Public web archives currently use a single heuristic to select mementos. We call this heuristic *Minimum Distance*. As explained in subsection 3.3, less than 50% of the embedded mementos selected by this heuristic are Prima Facie Coherent. Anticipating that a more sophisticated heuristic would improve coherence, this experiment also included a second heuristic we call *Last-Modified Datetime/Memento-Datetime Bracket*, or *Bracket* for short. Both heuristics are defined below:

Minimum Distance (Mindist) Select the URI-M with Memento-Datetime closest to the root’s (it can be before or after). The heuristic cost is absolute value of the difference between the two Memento-Datetimes.

Last-Modified/Memento-Datetime Bracket (Bracket) Establish URI-M lifetime using the range starting from Last-Modified datetime and ending with Memento-Datetime. This range is the time frame when the URI-M’s representation is known to have existed in its archived state. When an eURI-M’s lifetime overlaps an rURI-M Memento-Datetime, it is Prima Facie Coherent with cost 0. When the lifetime does not overlap, this heuristic falls back to a modified form of Mindist, which computes cost for both Memento-Datetime and Last-Modified, selecting the least as heuristic cost.

4.2.2 Source Constraint

Composite mementos are recomposed using a single or multiple archives. Using a single archive mimics the behavior of existing web archive user interfaces; all eURI-Ms are chosen from the same archive as the rURI-M. Using multiple archives enables eURI-Ms to be selected from multiple archives, possibly filling completeness and datetime gaps. The archives used are listed in Appendix A.

Table 4: rURI-R Timemap Statistics

Desc.	DMOZ	S.Eng.	Bitly	Deli	Total
URI-Rs	1,000	1,000	1,000	1,000	4,000
URI-Rs Archived	971	552	282	951	2,756
in Multiple Archives	14.5%	11.1%	23.0%	57.5%	29.5%
Timemaps	1,154	682	404	2,167	4,252
per URI-R	1.19	1.14	1.43	2.28	1.58
Mementos	120,765	11,508	36,816	174,214	343,303
per URI-R	124.37	20.85	130.55	183.19	124.57

4.3 Procedure

The examination of temporal spread was accomplished using the procedure described below. Timemaps and mementos were collected in May 2013. The proxies described in Subsection 3.1 were used to obtain timemaps for all archives except the Internet Archive, which was the only native Memento Protocol archive for the duration of preparation and data collection. Data collection was accomplished in two phases:

- Selection of rURI-Ms, and
- Recomposition of composite mementos.

Phase I. Selection of rURI-Ms

Phase I selected rURI-Ms by retrieving timemaps for all sample URI-Rs and choosing a subset of URI-Ms from the timemaps.

Process. We attempted to retrieve timemaps for all 4,000 sample rURI-Rs from 15 archives. Table 4 shows that 2,756 sample rURI-Rs had at least one timemap available; these are considered archived. Of the archived rURI-Rs, 29.5% are held in multiple archives and had multiple timemaps available, with an average of 1.58 timemaps each. The timemaps listed 343,303 rURI-Ms, averaging 124.57 per archived rURI-R. Of these 343,303 rURI-Ms, 88,054 (25.6%) were sampled for retrieval and evaluation. Specifically, we selected the rURI-M that was acquired closest to midnight on the 12th of each month during which the the rURI-R was archived.

Phase II. Recomposition of composite mementos

This phase recomposed composite mementos for each rURI-M selected in Phase I. It was executed four times, once for each combination of heuristic and source constraint. For each selected root URI-M, the following steps were performed.

Step 1. Download rURI-M. The rURI-M was retrieved using *curl*. A failure at this step stopped recomposition.

Of the 88,054 rURI-Ms, 93.6% were available and returned a 200 HTTP status. As shown in Table 5, the most common failure status was HTTP 503, which generally means the URI-M is not currently available (perhaps its data store is offline). Status 503 is generally considered temporary, so each 503 was retried twice over the course of a week before considering the 503 final for this study. The next most common failures were 403 and 404, access restricted and the memento does not exist, respectively.

Step 2. Extract eURI-Rs. The BeautifulSoup Python package (version 4.1.3) was used to extract eURI-Rs from HTML documents. Additionally, eURI-Rs were extracted from CSS files using regular expressions. If the memento did not contain eURI-Rs (or could not; e.g., images), processing finished.

Note: Steps 3–6 were repeated for each eURI-R found in step 2.

Step 3. Retrieve the eURI-R timemap. The eURI-R timemaps were retrieved using *curl*. Status 503 Failures were retried twice. If timemaps could not be retrieved, processing finished.

Table 5: rURI-M Statuses

Status	DMOZ	S.Eng.	Bitly	Deli.	Total	Percent
200	46,242	4,461	2,876	28,846	82,425	93.6%
503	2,989	254	155	1,046	4,444	5.0%
404	324	42	26	191	583	0.7%
403	161	6	28	193	388	0.4%
400	41	10	10	42	103	0.1%
Other	34	17	7	53	111	0.1%
Total	49,791	4,790	3,102	30,371	88,054	100.0%

Step 4. Select eURI-M for eURI-R . A heuristic was used to select the best eURI-M, which is the eURI-M with lowest score according to the heuristic in use. For the Mindist heuristic, the closest eURI-M was selected directly from the eURI-R's timemaps. For the Bracket heuristic, two candidate eURI-Ms were selected: one acquired before and one acquired after the root's Memento-Datetime (final selection is part of step 5). If no closest memento was available, processing finished.

Step 5. Download embedded memento. The representation for the selected eURI-M was retrieved using *curl*. Status 503 Failures were retried twice. For the Bracket heuristic, the "on or after" eURI-M was retrieved first. If it bracketed the root, it was considered best. If it did not bracket the root, the Bracket heuristic fell back to Minimum Distance (which may have caused the "before" eURI-M to be selected and retrieved).

Step 6. Recursion. Steps 2–5 were repeated for HTML frames and other resources containing eURI-Rs.

As Table 6 shows, 1,619,805–1,623,354 eURI-Rs were found varying by source constraint and heuristic. Timemaps and mementos were found for 1,250,641–1,332,993 (73.1%–78.1%) of URI-Rs. rURI-Ms average 21.2–21.2 eURI-Rs each, of which only 14.2–15.1 had available eURI-Ms. The primary reason for not finding mementos was that 19.3%–24.4% of eURI-Rs were not archived, as shown by the *No timemaps* reason. The next most common cause was 404 status (2.5%-2.9% of URI-Ms), which indicates the archive was not able to acquire a copy of the eURI-R. Other reasons account for about 1% of eURI-Rs.

5. RESULTS

5.1 Completeness

Completeness is the ratio of available mementos to required mementos. A rURI-M with no eURI-Ms (for example a plain HTML page with no CSS or JavaScript) is inherently 100% complete. That same web page with a single embedded image is 50% complete if no mementos for the image can be found (i.e., were never archived or not retrievable). Completeness is affected by both source constraint and heuristic.

Source Constraint and Completeness

As shown in the first line of Table 7, using multiple archives improves completeness by 4.1% for both heuristics. This is less than we expected. We hypothesize that the primary cause is minimal URI-R and Memento-Datetime overlap across the archives studied.

Heuristic and Completeness

Changing heuristic had a negligible impact on completeness, differing by just 0.01% for both single- and multi-archive recompositions. This is as we anticipated because the Bracket heuristic is an enhancement of the Minimum Distance heuristic. What is a little

Table 6: eURI-M Retrieval Statistics

Description	Mindist Single	Mindist Multi	Bracket Single	Bracket Multi
Memento Counts				
eURI-Rs	1,620,597	1,623,354	1,619,805	1,623,127
per rURI-M	19.7	19.7	19.7	19.7
eURI-Ms available	1,250,641	1,330,858	1,252,125	1,332,993
per rURI-M	14.2	15.1	14.2	15.1
eURI-Ms not found	369,956	292,494	367,680	290,134
***Not-Found* Reasons**				
No timemaps	395,545	312,843	395,065	312,641
404	41,588	46,330	40,428	44,852
403	6,186	6,194	6,105	6,116
503	6,092	5,806	5,697	5,442
Other	2,970	3,746	2,810	3,508

Table 7: Completeness and Temporal Coherence

Description	Mindist Single	Mindist Multi	Bracket Single	Bracket Multi
Completeness				
Mean complete	76.1%	80.2%	76.2%	80.3%
Mean missing	23.9%	19.8%	23.8%	19.7%
Temporal Coherence				
Mean Prima Facie Coherent	41.0%	40.9%	54.7%	54.6%
Mean Possibly Coherent	27.3%	28.7%	12.8%	14.2%
Mean Probably Violative	2.5%	5.3%	2.5%	5.3%
Mean Prima Facie Violative	5.3%	5.3%	6.2%	6.2%

surprising is that there is any difference at all—both heuristics will either find or not find an eURI-M for the same URI-R and target datetime. The difference has two causes. First, the heuristics can select different eURI-Ms for the same URI-R and target datetime; one may be available and the other not. Second, even when both are available, they may have different eURI-Rs.

5.2 Temporal Coherence

Mean temporal coherence values for eURI-Ms are in the lower section of Table 7. These means represent the percentage of all required URI-Ms in the specified coherence state. Most of the available URI-Ms were Prima Facie Coherent, ranging from a mean of 41.0% for Mindist/Single to 54.7% for Bracket/Single. Adding Possibly Coherent increases the means to 63.8–69.6%. On the other hand, 7.8%–11.5% of required mementos were violative, with 5.3%–6.2% Prima Facie Violative.

Source Constraint and Temporal Coherence

Unlike completeness, multiple archives have little effect on temporal coherence, just 0.1%. However, using multiple archives appears to reduce Prima Facie Coherence. This is a side effect of the fact that during this study only the Internet Archive returned original Last-Modified headers[3]. Thus, only the Internet Archive's URI-Ms could be Prima Facie Coherent. When using multiple archives, Internet Archive rURI-Ms can be paired with eURI-Ms from other archives. Since Last-Modified is not available from other archives, those eURI-Ms cannot be Prima Facie Coherent. If paired with an Internet Archive rURI-M, the eURI-Ms may have been Prima Facie Coherent. Using multiple archives improves the Possibly Coherent mean. More available eURI-M increases the likelihood that the eURI-M selected will match a Possibly Coherent pattern. Multiple archives also increase Probably Violative counts. The primary reason is the increase in eURI-M without Last-Modified datetimes. Multiple archives also increase completeness by 4.1%, but 2.8% is an increase in Probably Violative URI-Ms. The violative eURI-Ms ratio stays the same with multiple archives.

Table 8: eURI-M Temporal Coherence Counts

Description	Mindist Single	Mindist Multi	Bracket Single	Bracket Multi
Prima Facie Coherent	622,565	621,447	864,736	859,625
Possibly coherent	497,405	466,046	244,104	215,585
Probably violative	104,376	53,734	104,339	53,694
Prima Facie Violative	100,760	103,662	114,062	117,469
Total	1,325,106	1,244,889	1,327,241	1,246,373

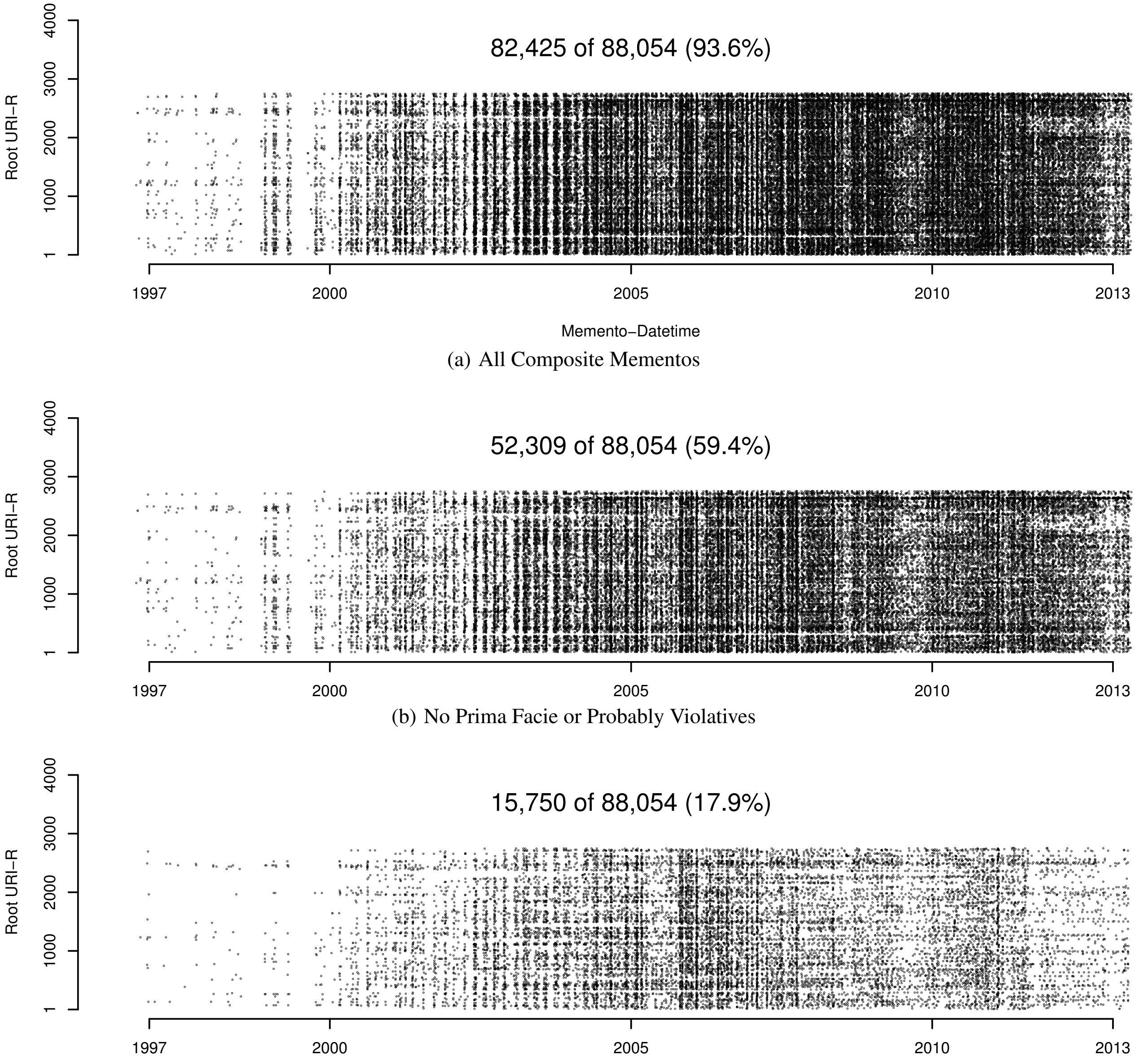
Heuristic and Temporal Coherence

Compared to Mindist, the Bracket heuristic increased Prima Facie Coherent eURI-Ms by 13.7%. However, Possibly Coherent eURI-Ms decreased by 14.5% and Prima Facie Violatives eURI-Ms by 0.9%. This appears to be reduction in coherence—quite unexpected. Indeed, we expected to see the coherence values increase and the violatives decrease. The Bracket heuristic makes two changes when Last-Modified is available. First, when selecting the best eURI-M and the rURI-M Memento-Datetime is bracketed by an eURI-M lifetime (see 4.2.1), then the eURI-M is Prima Facie Coherent. Compared to Mindist's Memento-Datetime-only calculation, Bracket is much more likely to select a Prima Facie Coherent eURI-M over a Possibly Coherent eURI-M, increasing Prima Facie Coherent eURI-Ms and decreasing Possibly Coherent eURI-Ms as shown in Table 8. Second, when the root is not bracketed, Last-Modified is used in the fall-back Mindist calculation, which increases the selection of eURI-Ms captured after the root. All these are Prima Facie Violative by definition (see 3.3). However, because Bracket uses eURI-M lifetime instead of just a point in time whenever possible, even with the increase in violatives, Bracket still provides a more accurate representation of the composite memento.

Composite Memento Coherence

Figure 7 and Table 9 provide two other, more compelling, views of the same data. (The figure shows only the Bracket heuristic and Single Archive source constraint due to space limitations.) Each row on the plots represents one sample URI-R. The horizontal axis represents the rURI-M Memento-Datetime. Only 2,756 of the the rows contain data; the other 1,244 sample URI-Rs had no timemaps (see Table 4) or were not publicly available (i.e., 403 status). For the available URI-Rs, 88,054 rURI-Ms were selected (see 4.3). Of these, 93.6% could be retrieved and were recomposed as shown under *Bracket Single* column in the *All Coherence States* row. Excluding composite mementos containing Prima Facie or Probably Violative eURI-Ms, leaves 54.8%–65.1% (*Possibly Coherent & Prima Facie Coherent* row). And, as the last row of Table 9 shows, only 12.4%–17.9% are both 100% Prima Facie Coherent and 100% complete. Figure 7 provides a visual dipiction: 7(a) includes all composite mementos, 7(b) excludes composite mementos with Prima Facie or Probably Violatives, and 7(c) includes only composite mementos that are both 100% complete and composed only from Prima Facie Coherent eURI-Ms. These results are striking and could have significant implications for non-casual uses such as litigation [19] and historical research. Still, public web

[3] Several other web archives now also return original Last-Modified headers.



(a) All Composite Mementos

(b) No Prima Facie or Probably Violatives

(c) Complete and Prima Facie Coherent

Figure 7: Composite Memento Coherence

Table 9: Composite Memento Coherence

	Mindist Single		Mindist Multi		Bracket Single		Bracket Multi	
Description	Quantity	Percent	Quantity	Percent	Quantity	Percent	Quantity	Percent
All Root Mementos in Timemaps	88,054	100.0%	88,054	100.0%	88,054	100.0%	88,054	100.0%
All Composite Mementos (complete and incomplete)								
All Coherence States	82,425	93.6%	82,425	93.6%	82,425	93.6%	82,425	93.6%
Probably Violative through Prima Facie Coherent	62,749	71.3%	62,603	71.1%	60,525	68.7%	60,525	68.7%
Possibly Coherent & Prima Facie Coherent	54,120	61.5%	48,289	54.8%	52,309	59.4%	52,309	59.4%
Prima Facie Coherent only	23,147	26.3%	20,262	23.0%	34,074	38.7%	34,074	38.7%
Complete Composite Mementos Only								
All Coherence States	26,969	30.6%	29,200	33.2%	27,077	30.8%	27,077	30.8%
Probably Violative through Prima Facie Coherent	23,284	26.4%	25,180	28.6%	22,678	25.8%	22,678	25.8%
Possibly Coherent & Prima Facie Coherent	21,576	24.5%	21,976	25.0%	21,010	23.9%	21,010	23.9%
Prima Facie Coherent only	10,944	12.4%	10,901	12.4%	15,750	17.9%	15,750	17.9%

archive interfaces give no indication that on average only 17.9% of holdings are both 100% complete and 100% Prima Facie Coherent.

6. FUTURE WORK

6.1 Redirection and Missing Mementos

The research conducted so far has accepted the lists of URI-Ms in timemaps as ground truth; however, timemaps only tell part of the story. For example, a URI-M listed in a timemap may redirect to another with a different Memento-Datetime [5]. Redirections are likely to impact temporal coherence, which could change the coherence state of some eURI-Ms. Redirection could also be an indication of duplication; however, timemaps and archives provide no indication of cause for redirections, so we cannot be sure. Timemaps also list URI-Ms that do not exist or are not accessible, which this study simply considers missing. A heuristic that searches the timemap to locate reasonable substitutes for missing URI-Ms could be developed.

6.2 Duplicates and Similarity

This study calculated deltas as the simple difference between rURI-M and eURI-M Memento-Datetimes (as is common practice for web archives). We have observed that web archives frequently return identical representations for multiple Memento-Datetimes. Consider an rURI-M with a Memento-Datetime of July 5 and two eURI-Ms with Memento-Datetimes of July 1 and July 10. It appears that the best delta is four days (July 5 − July 1). However, if the eURI-Ms have identical representations (e.g., a logo that has not changed), then the July 1 eURI-M should be considered Prima Facie Coherent instead of Possibly Coherent. For file types such as images, duplicates are easily identified using message digest algorithms (e.g., SHA-256). However, text (and especially HTML), is commonly modified for branding and user interface purposes by every web archive we have studied. This is evident in Figure 1, which shows the banner added by the Internet Archive's Wayback Machine. The changes are inconsequential to human users but prevent simple machine duplicate detection. Lexical Signatures [29, 23] and the *SimHash* [13, 24] algorithm may be able to help determine if modified files similar enough to be considered semantically equivalent and thus Prima Facie or Possibly Coherent. Ainsworth et al. [4] define equality and similarity patterns that can be used once appropriate similarity methods and measures are identified.

6.3 Communicating Status

One significant issue with existing web archive user interfaces is that temporal coherence is not communicated to the user. An icon or symbol that quickly communicates the temporal coherence of composite mementos would be very useful.

7. CONCLUSIONS

This study addressed temporal coherence of composite mementos recomposed from existing web archive holdings, comparing the use of Single and Multiple Archives and two eURI-M selection heuristics: Bracket and Minimum Distance. Using a sample of 4,000 original URIs with 343,303 corresponding URI-Ms, 88,054 (25.6%) URI-Ms were selected for recomposition. Of these, 82,425 (93.6%) were available and recomposed. We found that using multiple archives improves completeness by 4.1% but has no significant impact on temporal coherence. We also found that the Bracket heuristic improved temporal coherence, in particular increasing the number of Prima Facie Coherent mementos by 13.7%, compared with the Minimum Distance heuristic typically employed by web archives. To maximize completeness, multiple archives provide the best results; changing heuristic provided no significant benefit. To maximize coherence, the Bracket heuristic with a single archive provides the best results (privided that the archive has significant holdings for the URI-R). Truly significant is the finding that at most 17.9% of composite mementos held by web archives are both complete and Prima Facie Coherent.

As of this study, only the Internet Archive returns the original Last-Modified header, which significantly improves both eURI-M selection and temporal coherence evaluation. We recommend that all web archives capture and return all original headers.

Acknowledgments

Memento is a joint project between the Los Alamos National Laboratory Research Library and Old Dominion University. We are grateful to the Internet Archive for their continued support of Memento access to their archive.

8. REFERENCES

[1] S. G. Ainsworth, A. AlSum, H. SalahEldeen, M. C. Weigle, and M. L. Nelson. How much of the Web is archived? In *Proceedings of JCDL'11*, pages 133–136, June 2011.

[2] S. G. Ainsworth, A. AlSum, H. SalahEldeen, M. C. Weigle, and M. L. Nelson. How much of the Web is archived? Technical Report arXiv:1212.6177, Old Dominion University, December 2012.

[3] S. G. Ainsworth and M. L. Nelson. Evaluating sliding and sticky target policies by measuring temporal drift in acyclic walks through a web archive. In *Proceedings of JCDL'13*, July 2013.

[4] S. G. Ainsworth, M. L. Nelson, and H. Van de Sompel. A framework for evaluation of composite memento temporal coherence. Technical Report arXiv:1402.0928, Old Dominion University, February 2014.

[5] A. AlSum, M. L. Nelson, R. Sanderson, and H. Van de Sompel. Archival HTTP redirection retrieval policies. In *Proceedings of WWW'13 Companion*, pages 1051–1058, Republic and Canton of Geneva, Switzerland, 2013.

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

[7] M. Ben Saad and S. Gançarski. Archiving the Web using page changes patterns: a case study. In *Proceedings of JCDL'11*, pages 113–122, 2011.

[8] M. Ben Saad and S. Gançarski. Improving the quality of web archives through the importance of changes. In *Proceedings of DEXA'11*, pages 394–409, 2011.

[9] M. Ben Saad, Z. Pehlivan, and S. Gançarski. Coherence-oriented crawling and navigation using patterns for web archives. In *Proceedings of TPDL'11*, pages 421–433, 2011.

[10] A. Bright. Web evidence points to pro-Russia rebels in downing of MH17. *Christian Science Monitor*, 2014.

[11] J. F. Brunelle, M. Kelly, H. SalahEldeen, M. C. Weigle, and M. L. Nelson. Not all mementos are created equal: Measuring the impact of missing resources. In *Proceedings of JCDL'14*, pages 321–330, September 2014.

[12] J. F. Brunelle, M. L. Nelson, L. Balakireva, R. Sanderson, and H. Van de Sompel. Evaluating the SiteStory transactional web archive with the ApacheBench tool. In *Proceedings of TPDL'13*, pages 204–215, 2012.

[13] M. S. Charikar. Similarity estimation techniques from rounding algorithms. In *Proceedings of STOC'02*, pages 380–388, New York, NY, USA, 2002.

[14] M. Day. Preserving the fabric of our lives: A survey of web preservation initiatives. In *Proceedings of ECDL'05*, pages 461–472, 2003.

[15] D. Denev, A. Mazeika, M. Spaniol, and G. Weikum. SHARC: Framework for quality-conscious web archiving. *Proceedings of the VLDB Endowment*, 2(1):586–597, August 2009.

[16] C. E. Dyreson, H. ling Lin, and Y. Wang. Managing versions of web documents in a transaction-time web server. In *Proceedings of WWW'04*, 2004.

[17] G. Eysenbach and M. Trudel. Going, going, still there: Using the WebCite service to permanently archive cited web pages. *Journal of Medical Internet Research*, 7(5), 2005.

[18] K. Fitch. Web site archiving: an approach to recording every materially different response produced by a website. In *9th Australasian World Wide Web Conference, Sanctuary Cove, Queensland, Australia,*, pages 5–9, 2003.

[19] B. A. Howell. Proving web history: How to use the Internet Archive. *Journal of Internet Law*, 9(8):3–9, 2006.

[20] S. M. Jones, M. L. Nelson, H. Shankar, and H. V. de Sompel. Bringing web time travel to MediaWiki: An assessment of the Memento MediaWiki Extension. Technical Report arXiv:1406.3876, Old Dominion University and Los Alamos National Laboratory, June 2014.

[21] B. Kahle. Wayback Machine just grew today to 479,160,477,000 pages. Go @internetarchive ! https://archive.org/web [Twitter post]. Retrieved from `https://twitter.com/brewster_kahle/status/603611567276589056`.

[22] B. Kahle. Wayback machine: Now with 240,000,000,000 URLs. `http://blog.archive.org/2013/01/09/updated-wayback/`, January 2013.

[23] M. Klein and M. L. Nelson. Revisiting lexical signatures to (re-)discover web pages. In B. Christensen-Dalsgaard, D. Castelli, B. Ammitzbøll Jurik, and J. Lippincott, editors, *Research and Advanced Technology for Digital Libraries*, volume 5173 of *Lecture Notes in Computer Science*, pages 371–382. Springer Berlin Heidelberg, 2008.

[24] G. S. Manku, A. Jain, and A. Das Sarma. Detecting near-duplicates for web crawling. In *Proceedings of WWW'07*, pages 141–150, New York, NY, USA, 2007.

[25] J. Masanès. *Web Archiving*. Springer, Heidelberg, 2006.

[26] F. McCown and M. L. Nelson. Characterization of search engine caches. In *Proceedings of IS&T Archiving 2007*, pages 48–52, May 2007. (Also available as arXiv:cs/0703083v2).

[27] G. Mohr, M. Stack, I. Rnitovic, D. Avery, and M. Kimpton. Introduction to Heritrix, an archival quality web crawler. In *Proceedings of IWAW'04*, September 2004.

[28] K. C. Negulescu. Web archiving @ the Internet Archive. `http://www.digitalpreservation.gov/news/events/ndiipp_meetings/ndiipp10/docs/July21/session09/NDIIPP072110FinalIA.ppt`, 2010.

[29] S.-T. Park, D. M. Pennock, C. L. Giles, and R. Krovetz. Analysis of lexical signatures for finding lost or related documents. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 11–18, New York, NY, USA, 2002.

[30] M. Spaniol, D. Denev, A. Mazeika, G. Weikum, and P. Senellart. Data quality in web archiving. In *Proceedings of WICOW'09*, pages 19–26, 2009.

[31] M. Spaniol, A. Mazeika, D. Denev, and G. Weikum. "Catch me if you can": Visual analysis of coherence defects in web archiving. In *Proceedings of IWAW'09*, pages 27–37, 2009.

[32] M. Thelwall and L. Vaughan. A fair history of the Web? examining country balance in the Internet Archive. *Library & Information Science Research*, 26(2):162–176, 2004.

[33] B. Tofel. 'Wayback' for accessing web archives. In *Proceedings of IWAW'07)*, 2007.

[34] H. Van de Sompel, M. Nelson, and R. Sanderson. HTTP framework for time-based access to resource states—Memento (IETF RFC 7089), December 2013. `http://tools.ietf.org/html/rfc7089`.

[35] H. Van de Sompel, M. L. Nelson, R. Sanderson, L. L. Balakireva, S. Ainsworth, and H. Shankar. Memento: Time travel for the Web. Technical Report arXiv:0911.1112, 2009.

[36] M. C. Weigle. How much of the Web is archived? `http://ws-dl.blogspot.com/2011/06/2011-06-23-how-much-of-web-is-archived.html`, June 2011.

APPENDIX

A. WEB ARCHIVES

Table 10 lists the web archives used for this study.

Table 10: Web Archives Used in This Study

Archive	Home Page
Archiv Českého Webu	wayback.webarchiv.cz
Archive-It!	wayback.archive-it.org
Archief Web EU	www.archiefweb.eu
Archuivo de la Portguguesa	arquivo.pt
L'Arxiu Web de Catalunya	www.padi.cat
Library and Archives Canada	www.collectionscanada.gc.ca
Hrvatski arhiv weba	haw.nsk.hr
Internet Archive	www.archive.org
Icelandic Web Archive	wayback.vefsafn.is
Library of Congress Web Archives	webarchive.loc.gov
The National Archives (UK)	webarchive.nationalarchives.gov.uk
NTU Web Archive	webarchive.lib.ntu.edu.tw
UK Web Archive (British Library)	www.webarchive.org.uk
Wikia	wikia.com
Wikipedia	www.wikipedia.org