Noname manuscript No.
(will be inserted by the editor)

Not All Mementos Are Created Equal: Measuring The Impact Of Missing Resources

Justin F. Brunelle, Mat Kelly, Hany SalahEldeen, Michele C. Weigle, and Michael L. Nelson

Received: date / Accepted: date

Abstract Web archives do not capture every resource on every page that they attempt to archive. This results in archived pages missing a portion of their embedded resources. These embedded resources have varying historic, utility, and importance values. The proportion of missing embedded resources does not provide an accurate measure of their impact on the Web page; some embedded resources are more important to the utility of a page than others. We propose a method to measure the relative value of embedded resources and assign a damage rating to archived pages as a way to evaluate archival success. In this paper, we show that Web users' perceptions of damage are not accurately estimated by the proportion of missing embedded resources. The proportion of missing embedded resources is a less accurate estimate of resource damage than a random selection. We propose a damage rating algorithm that provides closer alignment to Web user perception, providing an overall improved agreement with users on memento damage by 17% and an improvement by 51% if the mementos are not similarly damaged. We use our algorithm to measure damage in the Internet Archive, showing that it is getting better at mitigating damage over time (going from 0.16 in 1998 to 0.13 in 2013). However, we show that a greater number of important embedded resources (2.05 per memento on average) are missing over time.

Keywords Web Architecture, Web Archiving, Digital Preservation

Old Dominion University, Department of Computer Science
Norfolk VA, 23529, USA
{jbrunelle,mkelly,hany,mweigle,mln}@cs.odu.edu

1 Introduction

Web archives are valuable cultural repositories that capture and store Web content. Users make use of archives like the Internet Archive [18, 28] to retrieve archived material [12, 16] for a variety of purposes and in a variety of ways [2]. However, the resources being requested by Web users may not be complete; embedded resources are sometimes missing from an archived Web page [3]. Missing embedded resources return a non-200 HTTP status (e.g., 404, 503) when their URI is dereferenced.

Large images are often more important to an archived page's utility than small images. Similarly, stylesheets that format visible content are more important to the representation of the page than stylesheets without significant formatting responsibilities. We provide a mechanism to assess the impact of missing embedded resources in the archives.

Throughout this paper we use Memento Framework terminology. Memento [29] is a framework that allows web users to browse in the temporal dimension by aggregating the offerings of the archives at a single point of access. Original (or live web) resources are identified by URI-R, and archived versions of URI-Rs are called *mementos* and are identified by URI-M. Memento TimeMaps are machine-readable lists of mementos (at the level of single-archives or aggregation-of-archives) sorted by archival date.

This research is motivated by three factors. First, we want to understand how missing embedded resources impact Web user satisfaction (i.e., the utility of mementos). Using an algorithm to measure embedded resource importance, we determine whether an important embedded resource of the memento is missing (e.g., a main image or video essential to the user's understanding of the page), or the missing embedded resource is

a spacer image or a small button logo that contributes little to the memento's utility for the user. We propose a method of weighting embedded resources in a memento according to importance. We show that this is an improved damage rating over an unweighted count of missing embedded resources. We use Amazon's Mechanical Turk to compare our algorithm to Web users' notion of damage and to show an improvement over the unweighted count of missing embedded resources.

Second, we use our algorithm to assess the damage of mementos in the Internet Archive. We use the unweighted measure of damage as the proportion of missing embedded resources to all requested resources (M_m) and compare it to our algorithm's calculation of damage (D_m).

Third and finally, we measure damage in the Internet Archive over time using our weighted algorithm. We then describe how this algorithm can be used for future enhancements of the Heritrix crawler [17, 23] and archival processes.

2 Motivating Examples

We use the XKCD Web page as an example of a resource with embedded resources of differing importance. We captured the URI-R using the wget [1] command[1] and manually inflicted damage on a local memento of `http://www.xkcd.com/` by removing embedded images. We used PhantomJS [19] to dereference the URI-M, take a PNG snapshot of the representation, and record the resulting HTTP response headers of the embedded resources. We created three mementos of the URI-R: one duplicating its live Web counterpart (m_0), one with the central comic image removed (m_1), and one with two logo images removed (m_2). The snapshots taken by PhantomJS are provided in Figures 1(a), 1(b), and 1(c). As shown in the captions, the proportion of embedded missing resources (M_m) varies among the mementos.

The live XKCD site is missing two embedded stylesheets, as are m_0, m_1, and m_2 since they are copies of the live site. We verified that our memento m_0 has a M_m value identical to its live Web counterpart – the live resource and m_0 are both missing the same embedded resources (M_m=0.17). In Figure 1(a), m_0 has multiple embedded resources, but we focus on the three identified by the red arrows: the XKCD logo, the main comic image, and the banner of comics. The central image is most important to the utility of the page – without the main comic image, the user does not obtain the information from the page that the author intended (Figure 1(b)). The logo and banner are not essential to the user's understanding of the XKCD content (Figure 1(c)).

Cascading Stylesheets (CSS) also differ in importance. Some stylesheets are responsible for formatting small portions of a page, while others are responsible for placing images and other content or even organizing the entire page for the user. Figure 1(d) shows a memento of a URI-R that is missing a single stylesheet. This stylesheet is responsible for a large amount of information in the representation and without it, the meaning and utility of the memento changes. Figure 1(e) shows a memento that is properly styled but is missing two stylesheets that are not responsible for the majority of the content organization and the memento is still properly styled without them.

As we have discussed, the percentage of successfully dereferenced embedded resources is not the only factor in determining memento quality. In support of that principle, we refer to Figure 1(e) in which M_m=0.2 (6/30). However, it appears to be well-preserved. In our XKCD example, Figure 1(c) is missing two images (M_m=0.24) yet maintains more important embedded mementos than Figure 1(b) (M_m=0.29). These examples support the motivation of our research and demonstrate the need for evaluation criteria that assesses perceived memento damage.

3 Related Work

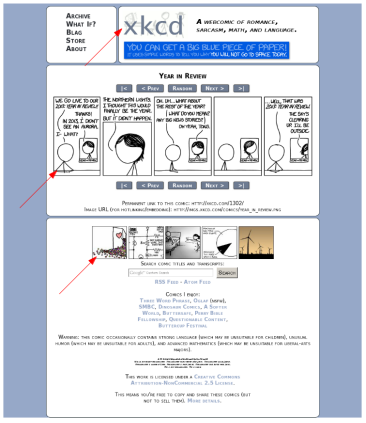
SalahEldeen et al. have studied the rate at which live resources disappear from the Web. In a study of the Egyptian Revolution, SalahEldeen found that 11% of the resources shared over Twitter were missing after one year [21, 22].

Kelly et al. studied the factors influencing archivability, including accessibility standards and their impact on memento completeness [14]. In this work, Kelly used a yearly sampling method to select mementos for testing. We use a similar method in this work to study memento damage.

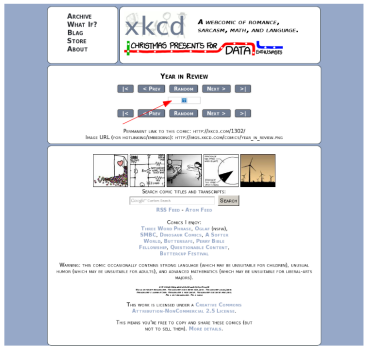
Spaniol has measured the quality of Web archives based on matching crawler strategies with resource change rates [7, 26, 27]. Ben Saad and Gançarski performed a similar study regarding the importance of changes on a page [4]. Gray and Martin created a framework for high quality mementos and assessing their quality by measuring the missing embedded resources [11]. While these studies focused on memento completeness and site coverage, we focus on assessing the importance of the artifacts that are missing.

Fersini et al. studied the importance of information blocks of a rendered Web page, finding that blocks with

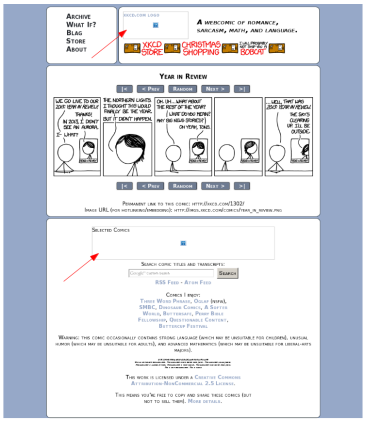
[1] We executed the wget command with parameters as follows: `wget -E -H -k -K -p http://www.xkcd.com/`



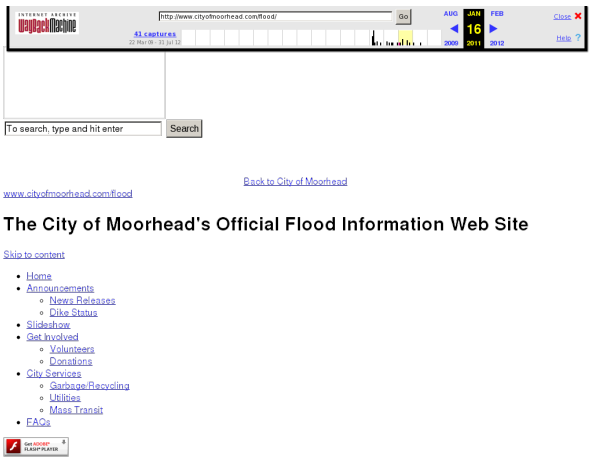
(a) All three of the embedded images are included in m_0 and identified by the red arrows (M_m=*0.17*).



(b) We removed the large, central image (that is the main content of the page) from m_1, identified by the red arrow (M_m=*0.24*).



(c) We removed the XKCD logo and banner of comics from m_2, identified by the red arrows (M_m=*0.29*).



(d) This memento (URI-M `http://web.archive.org/web/20110116022653/http://www.cityofmoorhead.com/flood/?`) is missing two stylesheets which changes the entire appearance and utility of the memento (M_m=*0.38*).

(e) Meanwhile, this memento (URI-M `http://web.archive.org/web/20060102083228/http://www.ascc.edu/`) is missing two stylesheets (along with two images) but does not appear damaged (M_m=*0.20*).

Fig. 1 Mementos have different meanings and usefulness depending on which embedded resources are missing from the memento (and the proportion of missing resources, M_m).

more images are more important [10]. Singh et al. found that multimedia within a page is essential for user understanding [24]. Ye et al. found that the information blocks close to the center of the viewport contain important information, while "noise" – or unimportant content – occurs on the fringes or edges of the page [30]. Kohlschütter et al. also found that important content was located in the center of pages [15]. Centrality is a way for authors to convey importance of information to their users. For example, images in the center of the viewport are more important or contribute to the users' understanding of a page than those positions on the fringes or outside the viewport of a page. Using these prior findings, we constructed an algorithm to assess the importance of embedded resources based on their MIME type, location in the viewport, and size in pixels.

Banos et al. created an algorithm to evaluate archival success based on adherence to standards for the purpose of assigning a resource archivability score [3]. Zhang et al. studied human perception and human ability to recognize differences in images effectively determining

human perception limitations for images at the pixel level [31]. Rademacher et al. used human perception to identify the visual factors that distinguished computer generated images from photographs [20]. We use human perception in a similar way to identify levels of memento damage.

The algorithm proposed in this paper determines the importance of embedded resources. Song et al. outlined an algorithm for determining the importance of sections of Web pages based on their content, size, and position [25]. We extend this algorithm (using many of the same principles) to measure the importance of missing embedded resources.

ΔM_m	Splits						
	5-0	4-1	3-2	2-3	1-4	0-5	Total
1.0							0.00
0.9							0.00
0.8	4						0.07
0.7							0.00
0.6							0.00
0.5	1	1					0.04
0.4							0.00
0.3	15	5					0.36
0.2	2						0.04
0.1	5	4	4	2		1	0.29
0.0	5	3	1	3			0.22
Total	0.58	0.23	0.09	0.09	0.00	0.02	1.0

Table 2 The turkers selected m_0 as the preferred memento 81% of the time, and more consistently for larger ΔM_m values.

Turker Assesment	M_m	
	Select m_0	Select m_1
m_0	44	0
m_1	11	0

Table 3 Confusion matrix of the turker assessments of the m_0 vs m_1 comparison test.

4 Users' Perception of Damage

As archivists, our perception of damage differs from that of more traditional Web users. To determine if M_m (percent missing) is a good estimate of human perception of damage, we used Amazon's Mechanical Turk to measure human agreement with M_m.

To ensure that Mechanical Turk workers (or more colloquially, "turkers") could evaluate damage, we presented turkers with pairs of mementos with varying levels of damage and asked them to select the memento they preferred to keep if given a choice between the two.

We captured 11 hand-selected URI-Rs (Table 1) on a local server and created five versions of the mementos for each URI-R. We manually inflicted damage to the mementos to create the five categories of damage. For the category *missing image*, we removed a prominent image (empirically identified as important) from the memento. For the category *missing css*, we removed a prominent CSS file to cause formatting issues in the memento; we empirically selected the CSS file to remove based on the greatest human-perceived detrimental impact to the page layout. We also created the categories *missing all images* (we removed every embedded image), *missing all resources* (we removed all embedded resources), and *original* (the URI-M was a direct copy of the live resource) and measured the M_m of each URI-M in each category. We refer to the four categories of damaged mementos in aggregate as m_1 and the *original* as m_0. These categories created a variety of damage ratings by a variety of missing embedded resources for identical URI-Rs at an identical time point to provide a wide spectrum of damaged mementos for turkers to evaluate.

With the goal of determining whether or not turkers can recognize damage in a memento, we presented the turkers with a m_1 and its m_0 counterpart (that is, a "damaged" and its *ground-truth* memento) and asked the turkers "We saved two pages for you. For which page did we do a better job?". For each URI-R, a pair of mementos consisting of m_0 and one of the four categories of m_1 were evaluated by five turkers for a total of 280 evaluations.

We show the judgement splits from the turker evaluations in Table 2. The judgement splits refer to the number of turkers that selected the correct-incorrect version. For example, a 0-5 split means all five turkers selected the m_1 (an incorrect selection), a 5-0 split means all five turkers selected the m_0 memento (the correct selection), and a 3-2 split means three turkers selected the m_0 memento and two selected the m_1 (a correct selection by the majority, but still a split decision among the turkers). For the purposes of this paper, we consider only 5-0 and 4-1 splits as agreement and all other splits as disagreement. ΔM_m refers to the delta between M_{m_0} and M_{m_1}.

The turkers selected m_0 as the preferred option (less damaged memento) 81% of the time (226/280). As ΔM_m shrinks, turker agreement is more consistent.

Regardless of ΔM_m, 81% of the evaluations agreed with M_m as a suitable damage metric (5-0 and 4-1 splits). Turkers were unsure about the damage (3-2 and 2-3 splits) 18% of the time and incorrectly identified damage only once. The average ΔM_m for the unsure selections was ¡0.01, and the only 0-5 split had a ΔM_m of 0.014, suggesting that confusion or disagreement occurs more often when the damage delta is smaller.

URI-R	m_0	M_m missing image	missing css	missing all images	missing all
http://www.cs.odu.edu/~mln/	0.14	0.43	0.29	0.43	0.43
http://activehistory.ca/2013/06/myspace-is-cool-again-too-bad-they-destroyed-history-along-the-way/comment-page-1/	0.0	0.32	0.32	0.57	0.85
http://www.albop.com/	0.0	0.13	0.0	0.50	0.50
http://www.cs.odu.edu/	0.10	0.13	0.11	0.82	0.81
http://ws-dl.blogspot.com/2013/08/2013-07-26-web-archiving-and-digital.html	0.07	0.08	0.08	0.13	0.14
http://www.cnn.com/2013/08/19/tech/social-media/zuckerberg-facebook-hack/	0.19	0.22	0.28	0.46	0.57
http://xkcd.com/	0.14	0.38	0.31	0.53	0.54
http://www.mozilla.org/	0.80	0.80	0.80	0.877	0.89
http://www.ehow.com/	0.05	0.05	0.06	0.11	0.33
http://google.com/	0.0	0.0	0.0	0.0	1.0
http://php.net/	0.32	0.33	0.33	0.37	0.37

Table 1 The 11 URI-Rs used to create the manually damaged dataset. M_m values are provided for each m_1.

Confusion matrices provide a consolidated view of an algorithm's performance. The top left quadrant shows the number of true positives, the top right shows the number of false negatives, the bottom left shows false positives, and the bottom right shows true negatives. The algorithm's accuracy ((True Positives + True Negatives) / (All Positives and Negatives)) and harmonic mean (or F_1 Score: 2 * True Positives / (2 * True Positives + False Positives + False Negatives)) are calculated using a confusion matrix. A harmonic mean provides an average (in this case, of the algorithm's success rate) and is sensitive to small values and outliers.

From the confusion matrix (Table 3), we can calculate the accuracy of m_0 vs m_1 as 0.80 with a harmonic mean of 0.88. Turker agreement does not match M_m 100% of the time with the m_0 vs m_1 test because of phenomena with aesthetics and human perception.

5 Evaluating Organic Damage

Because the turkers identified m_0 in the m_0 vs m_1 in 81% of the comparisons, we used turkers to evaluate our measured damage of mementos found in the Internet Archive.

This experiment uses the same set of 2,000 URI-Rs as in our previous work [6], which was sampled from Twitter and Archive-It. The first dataset, the *Twitter* set, consists of Bitly URIs shared over Twitter. The second dataset, the *Archive-It* set, was sampled from Archive-It collections. The Archive-It collections are created and curated by human users often corresponding to a certain event (e.g., National September 11 Memorial Museum) or a specific set of Web sites (e.g., City of San Francisco). We discarded non-HTML representations (e.g., JPEG and PDF) from both sets for a final dataset of 1,861 URI-Rs. Non-HTML representations do not contribute to this study since they do not have embedded resources. There is no overlap between the two sets.

Using this set of URI-Rs, we measured the damage of one memento per year from the Internet Archive TimeMap of each of the 1,861 URI-Rs, resulting in 45,341 URI-Ms. We randomly selected a subset of 100 URI-Ms from this set. Similar to the evaluation in Section 4, we gave turkers two mementos (we will generalize these to m_2 and m_3) from consecutive years from the same TimeMap and asked the turkers to select the less damaged memento ("We saved two pages for you. For which page did we do a better job?"). Because m_2 and m_3 are observed from the Internet Archive, neither is considered a *ground-truth*. We measured the damage M_m of mementos in the Internet Archive and compared it to the turker perception of the utility of the mementos.

Contrary to the test in Section 4, as ΔM_m grows, the turkers are not as effective at selecting the less damaged memento (the splits are shown in Table 4). The turkers only agree with M_m 12% of the time and completely disagree with M_m (1-4 and 0-5 splits) 44% of the time. This discrepancy demonstrates that turker assessment of damage does not match M_m. Additionally, we see that the turkers performed well when comparing m_0 vs m_1 (original vs damaged) but struggle to compare m_2 vs m_3 (damaged vs damaged).

From the confusion matrix (Table 5), we can calculate the accuracy of turker selections of m_2 vs m_3

ΔM_m	Splits						
	5-0	4-1	3-2	2-3	1-4	0-5	Total
1.0					1		0.01
0.9							0.00
0.8							0.00
0.7		1					0.01
0.6					1		0.01
0.5							0.00
0.4		1					0.01
0.3	1		3	4	1	2	0.11
0.2		5	6	5	12	9	0.37
0.1	4	5	10	11	15	3	0.48
0.0							0.00
Total	0.05	0.12	0.19	0.20	0.30	0.14	1.0

Table 4 The turker evaluations of the m_2 vs m_3 comparisons when using M_m as a damage measurement.

Turker Assesment	M_m	
	Select m_2	Select m_3
m_2	29	24
m_3	23	24

Table 5 Confusion matrix of the turker assessments of the m_2 vs m_3 comparison test against M_m.

Damage Calculation	AUC	F_1	Accuracy
M_m	0.472	0.55	0.46
M_{m_0}	0.789	0.88	0.80

Table 6 When compared to random, M_m performs worse than random selection and is worse than the optimal performance of m_0 vs m_1.

agreement with M_m is 0.46 with a harmonic mean (F_1) of 0.55. In a Receiver Operating Characteristic (ROC) curve [9], we calculated the Area Under the ROC Curve (AUC) for the results of the turker evaluations of m_2 vs m_3 against M_m and the results of the manually damaged m_0 vs m_1 test (as the optimal performance). The AUC of M_m is lower (AUC=0.472) than random (AUC=0.500) as shown in Table 6, meaning that M_m performed worse than random for matching turker perception of damage and far worse than the optimal performance (AUC=0.789), a further indicator that M_m is not a suitable metric for measuring memento damage.

6 Calculating Memento Damage

With M_m not matching Web users' perception of damage, we propose a new algorithm for assessing memento damage. Our proposed algorithm is based on the MIME type, size, and location of the embedded resource.

We define D_m as the damage rating, or cumulative damage, of a memento m in Equation 1. D_m is a normalized value ranging from $[0, 1]$. We calculate the potential damage of a memento and the actual damage of a memento and express the damage rating as the ratio of actual to potential damage. Notionally, potential damage is the cumulative importance of all embedded resources in the memento, while actual damage is only the importance of those embedded resources that are unsuccessfully dereferenced, or missing.

$$D_m = \frac{D_{m_{actual}}}{D_{m_{potential}}} \tag{1}$$

To determine potential and actual damage, we first define the set of all embedded resources R and the set of all missing resources R_r in Equation 2.

$$\begin{aligned} R &= \{\text{All embedded resources requested}\} \\ R_r &= \{\text{All missing embedded resources}\} \\ R_r &\subseteq R \end{aligned} \tag{2}$$

We calculate the importance of each embedded resource in the set R. The sum of each embedded resource is the potential damage $D_{m_{potential}}$ (Equation 3). Important resources are assigned additional weights to increase their relative value over unimportant resources (Equations 5 - 6).

$$\begin{aligned} D_{m_{potential}} = & \frac{\sum_{i=1}^{n_{[I,MM]}} D_{[I|MM]}(i)}{n_{[I|MM]}} + \frac{\sum_{i=1}^{n_C} D_C(i)}{n_C} \\ & \forall\{\text{I=Images, MM=Multimedia, C=CSS}\} \\ & n \in R \end{aligned} \tag{3}$$

Actual damage ($D_{m_{actual}}$, defined in Equation 4) is identical to $D_{m_{potential}}$ except it is computed using only the missing embedded resource set R_r.

$$\begin{aligned} D_{m_{actual}} = & \frac{\sum_{i=1}^{n_{[I,MM]}} D_{[I,MM]}(i)}{n_{[I|MM]}} + \frac{\sum_{i=1}^{n_C} D_C(i)}{n_C} \\ & \forall\{\text{I=Images, MM=Multimedia, C=CSS}\} \\ & n \in R_r \end{aligned} \tag{4}$$

In M_m, all embedded resources are treated as equal; all embedded resources are assigned a value of 1 with $weight$=1.0 applied. The potential damage is therefore the number of embedded resources, and the actual damage is the number of missing embedded resources. M_m is the unweighted ratio of missing embedded resources to total embedded resources.

We assign additional weights to important embedded resources at the expense of less important mementos. When a weight w is given to an embedded resource, all n embedded resources lose $\frac{w}{n}$ importance,

which redistributes the importance between embedded resources while keeping the sum of all importance constant. Images receive weights for image size and centrality (Equation 5). We use the pixel area (width x height) of the image and the page size along with a weight for horizontal and vertical central dividing line overlap by the image.

$$\begin{aligned} D_{[I|MM]} &= 1 + \frac{width * height}{\text{Page Size (pixels)}} \\ &+ (w_{\text{horizontal}} \iff \text{Overlaps horizontal center}) \\ &+ (w_{\text{vertical}} \iff \text{Overlaps vertical center}) \\ w_{\text{horizontal}} &= 0.25 \\ w_{\text{vertical}} &= 0.25 \end{aligned} \tag{5}$$

Embedded multimedia importance (D_{MM}) is calculated identically to image importance D_I, and we represent both in the same equation $D_{[I|MM]}$. Because size and centrality determine multimedia importance, we omit audio and other non-visual multimedia resources. We also classify Flash movies as multimedia.

Equation 6 outlines the damage from missing stylesheets, including a factor for a style threshold w_{style}.

$$\begin{aligned} D_C = 1 + w_{style} &\iff \\ &(¿75\% \text{ non-background in left two columns}) \\ &+ w_{tags} \iff \\ &(\text{tags in the DOM without matching CSS}) \\ w_{style} &= 0.50 \\ w_{tags} &= 0.50 \end{aligned} \tag{6}$$

Our intuition is that a missing important stylesheet will shift content to the left of the page rather than center content in the viewport, we divide a PNG snapshot of a memento into vertical thirds and measure the amount of content in each third. Traditional Web design (and particularly design enabled by stylesheets) evenly distributes content across each of the vertical thirds. If a stylesheet is missing *and* content appears to be shifted to primarily the left two-thirds, we assume the missing stylesheet was important to the distribution of content on the page.

When detecting content in the PNG snapshot, we use remaining CSS files and the HTML to determine the background color of the page. We measure the number of background and non-background colored pixels, with content being the number of non-background colored pixels. The proportion of non-background colored pixels in each vertical third gives us the amount of content in each partition.

The style threshold is determined as follows:

1. Determine background color
2. Render a PNG snapshot of the page
3. Divide PNG into vertical third partitions
4. Calculate number of pixels of the non-background color in each third for the viewport only (we used a 1024x768 viewport) and entire page
5. If ≤75% of the non-background colored pixels are in the left two thirds of the viewport, set $w_{style} = 0$ in Equation 6 (CSS file does not receive a weight)
6. If >75% of the non-background colored pixels are in the left two thirds of the viewport and left two thirds of the entire page and a stylesheet is missing, $w_{style} = 0.5$ in Equation 6 (CSS file does receive a weight)

For example, we created two mementos of the URI-R `http://www.pilotonline.com/` on a local server, one as it appears live (with all stylesheets – Figure 2(a)) and the other with its stylesheets removed (Figure 2(b)). The vertical partitions extend from the top of the PNG snapshot to the bottom. The percent of non-background color pixels in the viewports of our mementos are shown in their respective thirds in Figure 2. Notice that the non-background pixels (text, images, etc.) shift left when the CSS is missing. Intuitively, information is not meant to be displayed like the content in Figure 2(b).

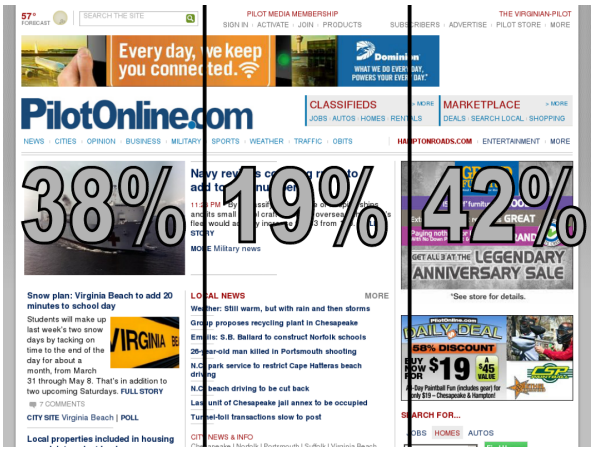
When we consider content outside of the viewport (Figures 3(a) and 3(b)), we see the same shift of content to the left when stylesheets are missing. However, the distribution of content in Figure 3(b) is more evenly distributed because the content has shifted down and fills out the middle and right vertical partitions more than in Figure 2(b). This is an indicator that the stylesheets that are missing in Figures 2(b) and 3(b) were important.

Along with the style threshold, the presence of tags on the page without a matching style suggests that the missing CSS contained the referenced formatting. If such tags exist without a matching style, $w_{tags} = 0.5$ in the Equation 6.

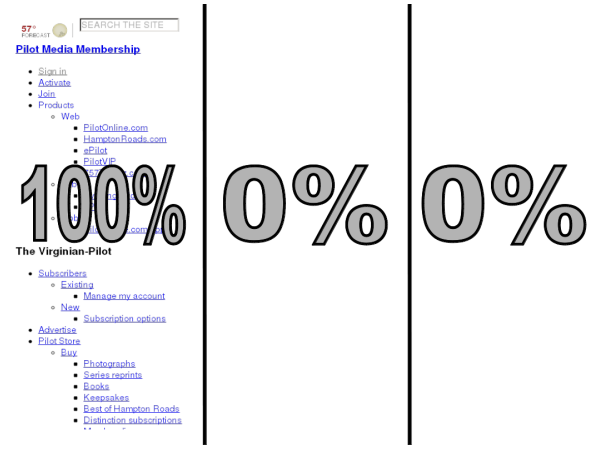
Embedded multimedia, images, and stylesheets do not account for the entirety of a page's importance and usefulness. We assume that text, as defined by the DOM and included on the page, is available regardless of archival success and therefore does not contribute to the damage calculation.

Equations 1 - 6 are used to compute D_m:

1. Load URI-M with PhantomJS
2. Find Potential Damage $D_{m_{potential}}$ (1)
   (a) Determine CSS importance D_C (6)
   (b) Determine Multimedia importance D_{MM} (5)
   (c) Determine Image importance D_I (5)



(a) We calculated that the non-background color is more evenly distributed between the three vertical partitions of the Pilot Online page with its stylesheet included than when it is missing.



(b) We calculated that the non-background color is most prevalent in the left-most vertical partition of the viewport of the Pilot Online page when it is missing its stylesheet.

Fig. 2 Missing stylesheets causes content to shift left. We show the percent of content in the vertical partitions of the viewport.

(a) When considering the entire page, the content of the page is distributed 33% in the left, 26% in the middle, and 41% in the right partitions when the stylesheet is present.

(b) When considering the entire page, the content of the page is distributed 84% in the left, 15% in the middle, and 1% in the right partitions when the stylesheet is missing.

Fig. 3 Missing stylesheets causes content to shift left.We show the percent of content in the vertical partitions of the page.

Figure	D_m	M_m
1(a)	*0.09*	*0.17*
1(b)	*0.41*	*0.24*
1(c)	*0.36*	*0.29*
1(d)	*0.59*	*0.38*
1(e)	*0.003*	*0.20*

Table 7 D_m vs M_m for the images in Figure 1. Note M_m¿D_m in 2 of 5 cases.

3. Determine proportion of unsuccessfully dereferenced embedded resources M_m
4. Find Actual Damage $D_{m_{actual}}$ (same as Step 3, but with only those URI-Ms unsuccessfully dereferenced)
5. Determine total damage D_m=[0,1] (1)

With D_m defined, we revisit the examples presented in Section 2. The values for D_m and M_m are listed in Table 7. Note that the damage ratings are closer to our empirical human assessment of memento quality than the proportion of the embedded resources that are missing.

Not all pages and page construction methods can be evaluated by this algorithm. An edge case not handled by this algorithm is any page constructed with iframes. Our algorithm uses JavaScript to determine the rendered location of embedded multimedia and images. When the embedded media is in a page embedded within another page, our algorithm does not provide the accurate rendered location. For this reason, we exclude iframes from our algorithm. We also exclude missing audio-only multimedia since the sound has no visual impact on the page, and sensory importance beyond sight is not considered in this algorithm.

While D_m includes multimedia calculations, multimedia resources are rarely embedded in our mementos (only observed twice in our entire set of 45,341 URI-Ms). We observed that multimedia is often loaded by JavaScript files embedded in the document object model (DOM); this prevents the multimedia files from being loaded into the archives since archival crawlers (at the time of this experiment) do not execute client-side

ΔD_m	Splits						
	5-0	4-1	3-2	2-3	1-4	0-5	Total
1.0							0.00
0.9		1					0.01
0.8							0.00
0.7							0.00
0.6			1				0.01
0.5							0.00
0.4	4	1					0.05
0.3	2	2	3				0.07
0.2		2	1	2	2	1	0.08
0.1	4	16	27	15	12	3	0.77
0.0							0.00
Total	0.10	0.22	0.32	0.17	0.14	0.04	1.0

Table 8 The turker evaluations of the m_2 vs m_3 comparisons when using D_m as a damage measurement.

Turker Assesment	D_m Select m_2	D_m Select m_3
m_2	45	32
m_3	8	14

Table 9 Confusion matrix of the turker assessments of the m_2 vs m_3 comparison test against D_m.

JavaScript and therefore do not discover the requested files.

7 Damage in the Archives

Having defined an algorithm for measuring D_m, we measured D_m values for each of the 45,341 URI-Ms from Section 5. We used these measurements to assess D_m's performance relative to turker assessment and perform damage measurements in the Internet Archive.

7.1 Turker Assessment of D_m

We compared D_m to turker assessment and M_m. As shown in Table 8, D_m agrees with turker assessment of damage 32% of the time, an increase of 18% over M_m. Additionally, 49% tie with a 3-2 or 2-3 split and only 16% of the turker evaluations disagreed with the D_m measure. Turkers agree more consistently when ΔD_m is larger. If we only consider ΔD_m ¿ 0.30, the turkers agree with D_m 71% of the time. However with ΔM_m ¿ 0.30, the turkers agree only 20% of the time.

From the confusion matrix in Table 9, we determine that the accuracy of D_m when comparing m_2 vs m_3 is 0.60, and the harmonic mean is 0.69. This is an improvement of 0.14 over the accuracy of M_m and an improvement over the harmonic mean of M_m by 0.14, showing that D_m measures damage closer to turker perception. We also calculated the AUC in a ROC curve for D_m and compared it to M_m and the optimal performance of the m_0 vs m_1 test. As shown in Table 10, D_m has an AUC of 0.584, an increase in 0.108 over M_m, showing that D_m outperforms M_m and is closer to the optimal performance of m_0 vs m_1 (AUC=0.789).

Damage Calculation	AUC	F_1	Accuracy
M_m	0.472	0.55	0.46
D_m	0.584	0.69	0.60
M_{m_0}	0.789	0.88	0.80

Table 10 D_m provides a closer estimate of turker perception of damage and our optimal performance of m_0 vs m_1 than M_m.

7.2 Measuring the Internet Archive

With D_m validated as aligning closer to turker evaluations than M_m, we used D_m to evaluate the Internet Archive's performance. Our measurement shows that only 46% of the 45,341 URI-Ms listed in the 1,861 TimeMaps are complete – that is, 54% of all URI-Ms listed in the Internet Archive TimeMaps we studied are missing at least one embedded resource[2]. In Figure 4, we show the average number of missing embedded resources M_m along with the average calculated damage D_m per URI-M per year.

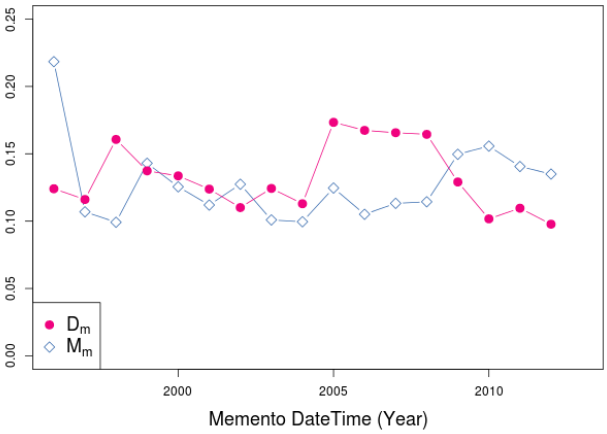


Fig. 4 The average embedded resources missed per memento per year as compared to damage per memento per year ($\overline{D_m}$=0.128, $\overline{M_m}$=0.132).

Because the number of missed mementos is important to M_m and D_m, we investigated the occurrence

[2] The Internet Archive performs URI canonicalization very well, and is assumed to not be a source of missing resources.

of missing and successfully dereferenced embedded resources. Most mementos are missing very few embedded resources with most missing 1-10 embedded resources (Figure 5), ($\mu = 1.7$, $\sigma = 4.6$). We calculate that 61% of mementos are missing 3 or fewer embedded resources, and 85% of mementos are missing 6 or fewer embedded resources. While the number of successfully dereferenced embedded resources in mementos is more evenly distributed (Figure 6), most mementos have very few embedded resources ($\mu = 17.6$, $\sigma = 86$).

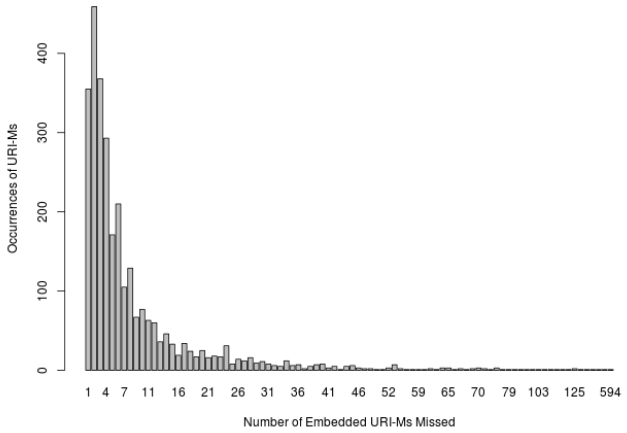


Fig. 5 The distribution of the number of missing embedded resources per URI-M.

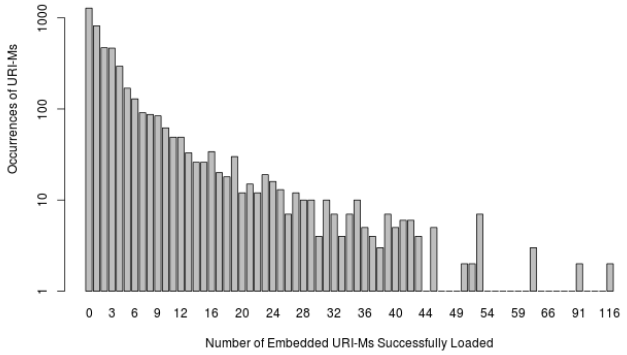


Fig. 6 The number of successfully dereferenced resources is more evenly distributed than those missing (Figure 5).

In aggregate, we observed that 45,009 of 292,192 embedded resources were missing, meaning 15% of the embedded resources in the dataset are missing. Of these, 25,848 (57% of the missing URI-Ms) were important, meaning they were assigned an additional weight by D_m (Equations 5 and 6). The average damage of all measured mementos was 0.132.

The yearly $\overline{D_m}$ goes from an average of 0.16 in 1998 to 0.13 in 2013. That means the Internet Archive is doing a better job (over time) reducing the total memento damage in its collection. However, the number of missing *important* resources (resources with an importance ¿1 due to added weights) is increasing, going from an average of 1.30 important resources per memento in 1997 to 2.38 important resources per memento in 2013 for an average of 2.05 missing per memento. Meanwhile, the number of unimportant missing embedded resources (damage rating ≤ 1) per memento is increasing at a lesser rate, going from 1.35 in 1997 to 1.64 in 2013. This suggests that while the Internet Archive is getting better overall at mitigating damage as much as possible, the archive is missing an increasing number of embedded resources deemed important.

The distribution of file types missing per memento (Figure 7) shows that most URI-Ms are missing ≥ 1 embedded resource and that stylesheets and JavaScript files are increasingly missing over time. Missing JavaScript may lead to additional missing files (such as multimedia). Images are missing at varying rates per memento.

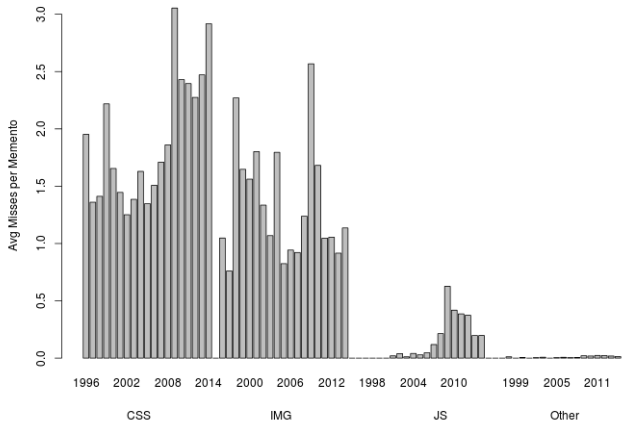


Fig. 7 The number of missed embedded resources per memento per year and MIME type.

7.3 Measuring WebCite

Contingency section

In an effort to measure a less prominent and different type of archive, we used the damage algorithm to determine M_m and D_m of WebCite[3][8]. WebCite is different from the Internet Archive's Heritrix crawler in that it is a page-at-a-time archiving tool. While the Internet Archive's goal is to archive the Web using the Heritrix crawler to identify crawl targets, add their URI-Rs to a crawl frontier, and crawl the frontier, WebCite and other page-at-a-time archivers allow users to submit URI-Rs for archiving, and WebCite immediately archives the resource[4]. This section identifies our damage measurement of this page-at-a-time archiver and outlines the differences between Heritrix and WebCite.

Our dataset has 992 mementos from 1,861 TimeMaps in the collection. The earliest available memento is from 2007, and the latest were from 2014. Only six mementos are available from 2014; we will focus on 2007-2013 as the target years of investigation. The average D_M – over all years – of the collection is $\mu = 0.397$ ($\sigma = 0.194$), and the average M_M is $\mu = 0.176$ ($\sigma = 0.0926$). All of the mementos in this collections are missing at least one embedded resource – 100% of the mementos are incomplete.

As shown in Figure 8, the average D_M in WebCite is increasing over time, going from 0.375 in 2007 to 0.475 in 2013 (the sample size in 2014 is too small to reliably draw conclusions). Meanwhile, the average M_M remains steady, going from 0.135 in 2007 to 0.139 in 2013. Only slight variation occurs, peaking at 0.287 in 2011.

Compared to the Internet Archive, WebCite has a higher damage value as well as missing more embedded resources. Additionally, the damage rating per memento is higher, indicating that more missing embedded resources are important (3,514 or 41.7%) than in the Internet Archive.

WebCite is missing, on average, 10.1 embedded resources per memento ($\sigma = 8.0$). Across the entire collection, 8,420 of 54,824, or 15.4% of the embedded resources were missing in our investigation. We calculate that 56% of mementos are missing 3 or fewer embedded resources, and 74% of mementos are missing 6 or fewer embedded resources (Figure 9 and 10).

The distribution of file types missing per memento (Figure 11) shows that most URI-Ms are missing ≥ 1 embedded image and CSS resources. WebCite has a lower occurence of missing stylesheets, but a higher occurrence of missing images. This will impact future work with D_M if we change the weighting of the importance of missing embedded resources – if we weight missing CSS as having a higher impact on overall resource damage, WebCite's collection might have a lower average damage rating. However, more investigation is needed before this conclusion may be reached. Our previous investigation [6] showed that WebCite has difficulties when encountering JavaScript and embedded iframes. However, its immediate archiving policies provide immediate results as opposed to crawlers that may incur delay between the time a URI-R is added to the frontier and a memento is created. WebCite's difficulties with JavaScript may contribute to the missing embedded resources if they were loaded through JavaScript.

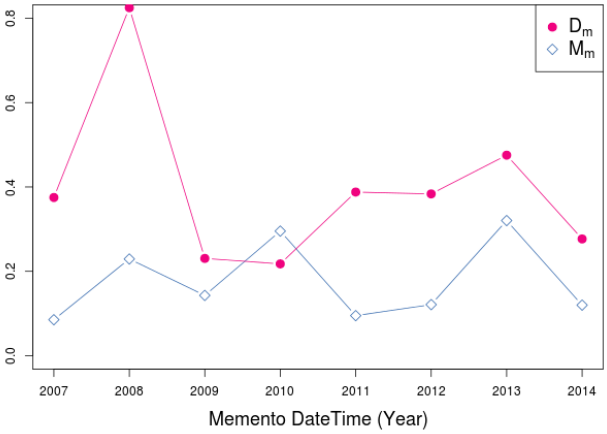


Fig. 8 The average embedded resources missed per memento per year as compared to damage per memento per year ($\overline{D_m}$=0.194, $\overline{M_m}$=0.176).

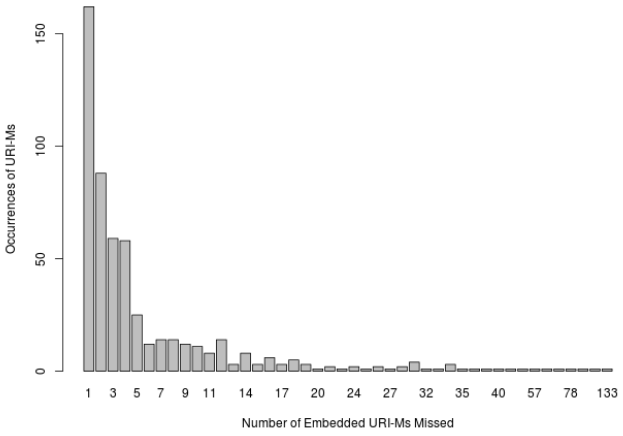


Fig. 9 The distribution of the number of missing embedded resources per URI-M in WebCite.

[3] `http://webcitation.org/`

[4] The Internet Archive has recently added an *on-demand* archiving utility at `http://archive.org/web/` under the heading "Save Page Now."

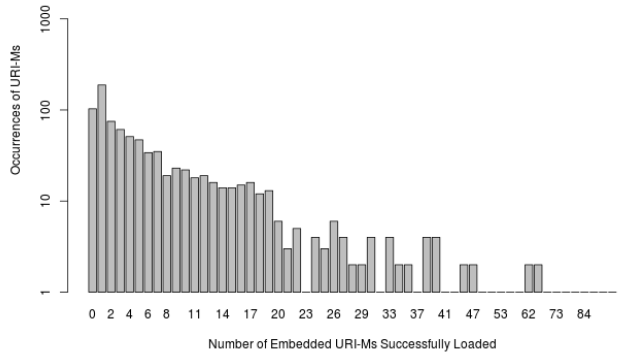


Fig. 10 In WebCite, the number of successfully dereferenced resources is more evenly? distributed than those missing (Figure 5).

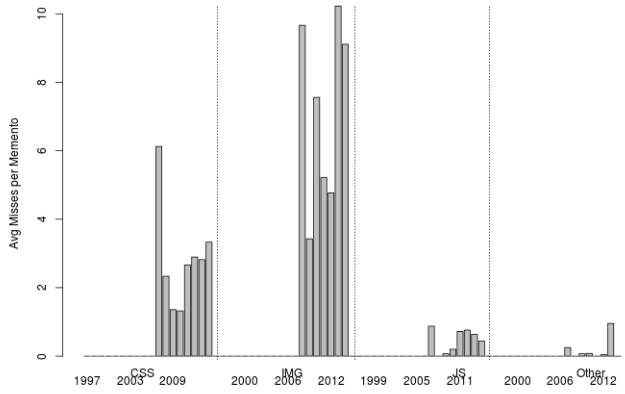


Fig. 11 The number of missed embedded resources per WebCite memento per year and MIME type.

7.4 Impact of JavaScript on Damage

As a preliminary investigation of the impact of JavaScript on archival tools, we set up an experiment to use Heritrix and PhantomJS [19] to crawl the same set of URI-Rs and measure the damage difference between the two set of mementos. Our goal is to understand how D_M is impacted by JavaScript by comparing mementos archive by a crawler that can execute JavaScript (PhantomJS) and a crawler that does not execute JavaScript (Heritrix).

7.4.1 PhantomJS vs Heritrix

Web crawlers operate by starting with a finite set of seed URI-Rs in a frontier – or list of crawl targets – and add to the frontier by extracting embedded resources and URI-Rs in the representations returned upon dereferencing the URI-R. This allows archival crawlers to discover embedded resources as well as new URI-Rs to crawl while creating mementos.

Representations of Web resources are increasingly reliant on JavaScript and other client-side technologies to load embedded resources and control the activity on the client. Web browsers use a JavaScript engine to execute the client side code; Web crawlers traditionally do not have such an engine or the ability to execute client-side code. The client-side code can be used to request additional data or resources (e.g., via Ajax) from servers after the initial page load. Crawlers are unable to discover the resources requested via Ajax and, therefore, are not adding these representations to their frontiers. The crawlers are missing embedded resources which ultimately causes the mementos of the crawled resources to be incomplete. We define *deferred representations* as those representations of resources that are difficult to archive because of their reliance on JavaScript and other client-side technologies to load embedded resources. We use the term *deferred* because the representation is not fully realized and constructed until *after* the JavaScript code is executed on the client.

To mitigate the impact web developers' practice of using JavaScript and Ajax to load embedded resources, crawlers like Heritrix have constructed approaches for extracting links from embedded JavaScript to be added to crawl frontiers (most recently, Google [5]). Even though it does not execute JavaScript, Heritrix does peek into the embedded JavaScript code to extract links where possible [13]. These processes rely on string matching and regular expressions to recognize URIs mentioned in the JavaScript; Heritrix 3.1.4 uses this approach. This is a sub-optimal approach because JavaScript may construct URIs during execution prior to requesting the request resulting in an incomplete URI extracted by the crawler.

Because archival crawlers' abilities differ from the abilities of browsers, the archives currently hold a representation of the Web from the point of view of crawlers and not Web users. That is, what we archive is increasingly different than what users experience. The intuitive solution to this challenge of archiving deferred representations is to provide crawlers with a JavaScript engine and allow headless browsing (i.e., allow a crawler to operate as would a browser) using a technology such as PhantomJS.

7.4.2 Crawling Deferred Representations

We sampled 50 URI-Rs by randomly generating bily.com URIs and identifying the URI-Rs to which the bitly URIs redirected. We then classified 50 URI-Rs as having deferred representations and crawled the set of URIs with Heritrix and PhantomJS.

During the Heritrix crawl, we used the 50 URI-Rs as a set of seed URIs and allowed Heritrix to create their mementos. The final frontier size of this crawl was 1,588 URIs of embedded resources used to create the mementos. Using our damage algorithm, we measured the damage of the mementos created by Heritrix and found that the average damage was 0.148. Recall that the measured avarage damage of the Internet Archive was 0.13.

To ensure the execution of JavaScript during the creation of mementos, we then crawled the 50 URI-Rs with PhantomJS. We recorded the embedded resources needed to create the representation, including those originating in the JavaScript. This created a frontier of 3,364 URIs which we used as a seed URI list in Heritrix. We then created the mementos using only the seed URI list, effectively creating mementos using the frontier list of PhantomJS. The damage of the mementos from this crawl was, on average, 0.1291.

PhantomJS provided a 13.5% improvement to the collection damage over Heritrix. This provides further evidence that JavaScript-dependent representations reduce the quality of mementos due to traditional crawlers' inability to execute JavaScript.

Not only does using PhantomJS provide a larger crawl frontier, but the damage rating of the resulting mementos is lower. In short, this initial investigation suggests that using PhantomJS mitigates the impact of JavaScript on resources with deferred representations and results in higher-quality mementos.

8 Conclusions

In this paper, we demonstrated that Web users (as represented by Mechanical Turk Workers) can correctly identify original mementos (m_0 vs m_1) 81% of the time when presented with an original and manually damaged pair of mementos. After randomly selecting 100 URI-Ms from the Internet Archive TimeMaps of 1,861 URI-Rs, we show that turkers' assessment of damage does not match that of M_m – in fact, their perception of damage more closely aligns to a random selection than with M_m.

To provide a damage metric closer to the perception of Web users, we proposed D_m, a damage calculation algorithm that estimates embedded resource importance to determine the perceived damage of mementos. Using turker evaluations, we showed that D_m aligns with turker perception 32% of the time when considering all ΔD_m values – an improvement of 17% over M_m. If we limit ΔD_m ¿ 0.30, we achieve an agreement of 71%, an improvement of 51% over M_m. We show that the performance of D_m is closer to that of the m_0 vs m_1 test than both M_m and a random selection.

We used D_m to measure the performance of the Internet Archive by measuring $\overline{D_m}$ of 1,861 URI-Rs. The average damage of the Internet Archive collection is 0.13 per memento and is missing 15% of its embedded resources. Mementos are missing 2.05 important resources on average. The Internet Archive has gotten better at mitigating damage over time, reducing D_m from 0.16 (1998) to 0.13 (2013).

With D_m, archival services can evaluate their performance and the quality of their mementos. The archives could measure a selection of mementos (either randomly sampled or by identifying those missing a proportion of embedded resources, such as ΔD_m ¿ 0.30) for damage to determine whether or not they have been satisfactorily archived. That is, with this algorithm, the archives can provide the greatest damage improvement through targeted repair efforts (e.g., which mementos require additional attention to ensure proper archiving?). Archives can also use historical damage ratings of a URI-R to identify memento improvements or changes.

This is a preliminary investigation of memento damage. We have shown that percentage of embedded resources missing is not an accurate representation of damage and have proposed a more accurate metric. Our future work will continue to improve upon the metric by using larger datasets, more turkers, and machine learning to further hone D_m. This will include a refinement of the relative weights of the embedded resources (e.g., the relative importance of CSS vs. images). We will also investigate the cumulative damage rating over time. For example, a logo that never changes over a 5 year period could have increased importance due to its use over multiple mementos. We plan to also measure the damage improvement of mementos if embedded resources are retroactively captured and included in past mementos. This cumulative damage improvement can help identify embedded resources that should be targeted by archives.

9 Acknowledgments

This work supported in part by the NSF (IIS 1009392) and the Library of Congress.

References

1. Introduction to GNU Wget. http://www.gnu.org/software/wget/ (2013)
2. Alnoamany, Y., Alsum, A., Weigle, M., Nelson, M.: Who and What Links to the Internet Archive. In: Proceedings of the Third International Conference on Theory and Practice of Digital Libraries, pp. 346–357. ACM (2013)
3. Banos, V., K. Yunhyong, S.R., Manolopoulos, Y.: CLEAR: a credible method to evaluate website archivability. In: Proceedings of the 9th International Conference on Preservation of Digital Objects (2013)
4. Ben Saad, M., Ganarski, S.: Archiving the web using page changes patterns: A case study. In: Proceedings of the 11th Annual International Joint Conference on Digital Libraries, pp. 113–122 (2011). URL http://www-poleia.lip6.fr/~bensaadm/JCDL2011.pdf
5. Brunelle, J.F.: Google and JavaScript. http://ws-dl.blogspot.com/2014/06/2014-06-18-google-and-javascript.html (2014)
6. Brunelle, J.F., Kelly, M., Weigle, M.C., Nelson, M.L.: The Impact of JavaScript on Archivability (2013). Submitted for publication
7. Denev, D., Mazeika, A., Spaniol, M., Weikum, G.: SHARC: framework for quality-conscious web archiving. Proceedings of the 35th International Conference on Very Large Data Bases **2**, 586–597 (2009)
8. Eysenbach, G., Trudel, M.: Going, going, still there: using the WebCite service to permanently archive cited web pages. Journal of Medical Internet Research **7**(5) (2005). DOI 10.2196/jmir.7.5.e60
9. Fawcett, T.: An introduction to ROC analysis. Pattern recognition letters **27**(8), 861–874 (2006)
10. Fersini, E., Messina, E., Archetti, F.: Enhancing web page classification through image-block importance analysis. Information Processing & Management **44**(4), 1431 – 1447 (2008)
11. Gray, G., Martin, S.: Choosing a sustainable web archiving method: A comparison of capture quality. D-Lib Magazine **19**(5) (2013)
12. Howell, B.A.: Proving Web History: How to Use the Internet Archive. Journal of Internet Law **9**(8), 3–9 (2006)
13. Jack, P.: Extractorhtml extract-javascript. https://webarchive.jira.com/wiki/display/Heritrix/ExtractorHTML+extract-javascript (2014)
14. Kelly, M., Brunelle, J.F., Weigle, M.C., Nelson, M.L.: On the Change in Archivability of Websites Over Time. In: Proceedings of the Third international conference on Theory and Practice of Digital Libraries (2013)
15. Kohlschütter, C., Fankhauser, P., Nejdl, W.: Boilerplate Detection Using Shallow Text Features. In: Proceedings of the Third ACM International Conference on Web Search and Data Mining, pp. 441–450 (2010)
16. Marshall, C.C., Shipman, F.M.: On the Institutional Archiving of Social Media. In: Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries, pp. 1–10 (2012)
17. Mohr, G., Kimpton, M., Stack, M., Ranitovic, I.: Introduction to Heritrix, an archival quality web crawler. In: Proceedings of the 4th International Web Archiving Workshop (2004)
18. Negulescu, K.C.: Web Archiving @ the Internet Archive. Presentation at the 2010 Digital Preservation Partners Meeting, 2010 http://www.digitalpreservation.gov/meetings/documents/ndiipp10/NDIIPP072110FinalIA.ppt
19. PhantomJS: PhantomJS. http://phantomjs.org/ (2013)
20. Rademacher, P., Lengyel, J., Cutrell, E., Whitted, T.: Measuring the Perception of Visual Realism in Images. In: Rendering Techniques 2001, Eurographics, p. 235247 (2001)
21. SalahEldeen, H.M., Nelson, M.L.: Losing my revolution: how many resources shared on social media have been lost? In: Proceedings of the Second international conference on Theory and Practice of Digital Libraries, pp. 125–137 (2012). URL http://dx.doi.org/10.1007/978-3-642-33290-6_14
22. SalahEldeen, H.M., Nelson, M.L.: Resurrecting My Revolution: Using Social Link Neighborhood in Bringing Context to the Disappearing Web. In: Proceedings of the Third international conference on Theory and Practice of Digital Libraries, pp. 333–345 (2013)
23. Sigursson, K.: Incremental crawling with Heritrix. In: Proceedings of the 5th International Web Archiving Workshop (2005)
24. Singh, R., Bhhatarai, B.D.: Information-theoretic identification of content pages for analyzing user information needs and actions on the multimedia web. In: Proceedings of the 2009 ACM Symposium on Applied Computing, pp. 1806–1810 (2009). URL http://doi.acm.org/10.1145/1529282.1529686
25. Song, R., Liu, H., Wen, J.R., Ma, W.Y.: Learning block importance models for web pages. In: Proceedings of the 13th international conference on

World Wide Web, pp. 203–211 (2004)
26. Spaniol, M., Denev, D., Mazeika, A., Weikum, G., Senellart, P.: Data quality in web archiving. In: Proceedings of the 3rd Workshop on Information Credibility on the Web, pp. 19–26. ACM (2009). URL http://scholar.google.com/scholar?cluster=3117236869878631784
27. Spaniol, M., Mazeika, A., Denev, D., Weikum, G.: Catch me if you can: Visual Analysis of Coherence Defects in Web Archiving. In: Proceedings of The 9th International Web Archiving Workshop, pp. 27–37 (2009). URL http://scholar.google.com/scholar?cluster=8411863341920369411
28. Tofel, B.: 'Wayback' for Accessing Web Archives. In: Proceedings of the 7th International Web Archiving Workshop (2007)
29. Van de Sompel, H., Nelson, M.L., Sanderson, R., Balakireva, L.L., Ainsworth, S., Shankar, H.: Memento: Time Travel for the Web. Tech. Rep. arXiv:0911.1112, Los Alamos National Laboratory (2009). URL http://scholar.google.com/scholar?cluster=12425719883705050728
30. Yi, L., Liu, B., Li, X.: Eliminating noisy information in web pages for data mining. In: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 296–305 (2003). URL http://doi.acm.org/10.1145/956750.956785
31. Zhang, X., Lin, W., Xue, P.: Just-noticeable difference estimation with pixels in images. Journal of Visual Communication and Image Representation **19**(1), 30–41 (2008)