2018

It is Hard to Compute Fixity on Archived Web Pages

Mohamed Aturban
Old Dominion University

Michael L. Nelson
Old Dominion University

Michele C. Weigle
Old Dominion University

Original Publication Citation

Aturban, M., Nelson, M. L., & Weigle, M. C. (2018) *It is hard to compute fixity on archived web pages* [Conference Paper]. Web Archiving and Digital Libraries Workshop (WADL 2018), Fort Worth, Texas. http://hdl.handle.net/10919/97988

It is Hard to Compute Fixity on Archived Web Pages

Mohamed Aturban
Old Dominion University
Norfolk, Virginia 23529, USA
maturban@cs.odu.edu

Michael L. Nelson
Old Dominion University
Norfolk, Virginia 23529, USA
mln@cs.odu.edu

Michele C. Weigle
Old Dominion University
Norfolk, Virginia 23529, USA
mweigle@cs.odu.edu

1 INTRODUCTION

Checking fixity in web archives is performed to ensure archived resources, or mementos (denoted by URI-M), have remained unaltered since when they were captured. The final report of the PREMIS Working Group [2] defines information used for fixity as "information used to verify whether an object has been altered in an undocumented or unauthorized way." The common technique for checking fixity is to generate a current hash value (i.e., a message digest or a checksum) for a file using a cryptographic hash function (e.g., SHA-256) and compare it to the hash value generated originally. If they have different hash values, then the file has been changed, either maliciously or not. We implicitly trust content delivered by web archives, but with the current trend of extended use of other public and private web archives, we should consider the question of validity of archived web pages. Most web archives do not allow users to retrieve fixity information. More importantly, even if fixity information is accessible, it is provided by the same archive delivering the content. A part of our research is dedicated to establishing and checking the fixity of archived resources with the following requirements:

- Any user can generate fixity information, not only the archive
- Fixity information can be generated on the mementos playback

2 EXAMPLES OF HOW MEMENTOS CHANGE

We have found that the HTTP entity stored in an archive change for several reasons. One example is the embedded image http://perma-archives.org/warc/20170101182814id_/http://umich.edu/includes/image/type/gallery/id/113/name/ResearchDIL-19Aug14_DM%28136%29.jpg/width/152/height/152/mode/minfit/ in the archived page perma-archives.org/warc/20170101182813/http://umich.edu/. Calculating hashes on the same image downloaded at two different times produced different results as Figure 1 depicts. We used Resemble.js[1] to compare the two images pixel by pixel. The mismatched pixels are shown in Figure 1c in pink.

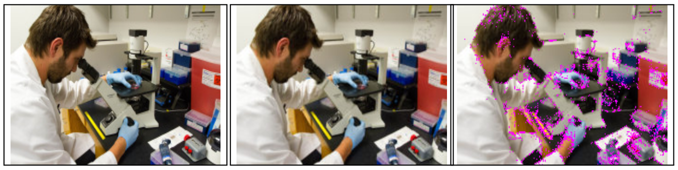
(a) On November 16, 2017, the hash ends in "...88c7".
(b) On December 25, 2017, the hash ends in "224b".
(c) Compare images (a) and (b). Mismatched pixels in pink.

Figure 1: The same image from perma-archives.org downloaded at two different times, produced two different hashes.

Figure 2 shows that we receive different entities for the same URI-M at different times. The memento is a stylesheet (CSS) file, and the URI-M is http://webarchive.proni.gov.uk/raw/20150303184134/http://fonts.googleapis.com/css?family=Droid+Serif.

Those two examples should never occur in a web archive. Add to those examples the known difficulties of client-side execution of JavaScript and network related transient error, and connection, fixity approaches for detecting tampering will produce many false positives.

```
@font-face {
  font-family: 'Droid Serif';
  font-style: normal;
  font-weight: normal;
  src: local('Droid Serif'),
       local('DroidSerif'),
       url('http://themes.googleusercontent.com/static/fonts/droidse
            rif/v2/0AKsP294HTD-nvJgucYTaJ0EAVxt0G0biEntp43Qt6E.ttf')
       format('truetype');
}
```

(a) Requesting the CSS file on November 11, 2017.

```
@font-face {
  font-family: 'Droid Serif';
  font-style: normal;
  font-weight: 400;
  src: local('Droid Serif Regular'),
       local('DroidSerif-Regular'),
       url(http://fonts.gstatic.com/s/droidserif/v7/0AKsP294HTD-nvJg
           ucYTaJ0EAVxt0G0biEntp43Qt6E.ttf)
       format('truetype');
}
```

(b) Requesting the CSS file on December 07, 2017.

Figure 2: Getting different content when requesting the same CSS file

3 QUANTIFYING CHANGES IN THE PLAYBACK OF MEMENTOS

We studied 18,472 mementos from 17 different web archives. We downloaded these mementos 10 times using Headless Chrome during 45 days between November 16, 2017 and December 31, 2017. The main aim of this study is to learn how the playback of these archived web pages changes during this period of time. Identifying and quantifying the types of changes present in today's archives will help us to differentiate between malicious and non-malicious changes in mementos in the future. Understanding these changes is important because conventional archival approaches regarding fixity are not applicable for web archives [1]. Table 1 shows the final number of selected mementos (URI-Ms) per archive. After downloading each memento 10 times over the 45 days, we quantified the following types of changes in the memento:

Table 1: The number of URI-Ms per archive. Total URI-Ms of 18,472

Archive	URI-Ms	Archive	URI-Ms
web.archive.org	1,600	archive.is	1,600
archive.bibalex.org	1,600	webarchive.loc.gov	1,600
arquivo.pt	1,600	webcitation.org	1,600
wayback.vefsafn.is	1,600	wayback.archive-it.org	1,407
swap.stanford.edu	1,233	nationalarchives.gov.uk	1,011
europarchive.org	990	webharvest.gov	733
veebiarhiiv.digar.ee	518	webarchive.proni.gov.uk	477
webarchive.org.uk	362	collectionscanada.gc.ca	359
perma-archives.org	182		

TimeMaps: Changes in TimeMaps can affect how a composite memento is constructed. The same memento might redirect differently

[1] https://github.com/Huddle/Resemble.js

each time it is requested (i.e., a change in the "Location" HTTP header).
HTTP entity body. Changes in the HTTP entity may occur because of dynamic content or random content generated by JavaScript.
Transient error: There are many types of transient errors. For example, web servers send back a "500" status when unable to handle the request, or an HTTP request gets a connection timeout error.
HTTP response headers: For instance, the MIME type (i.e., Content-Type Response header) of a resource might be converted (e.g., from GIF to PNG), or the server could return a "Memento-Datetime" header with a different datetime value each time.
HTTP status code: A web archive could respond with different HTTP status code when requesting the same URI-M. For example, the archive returns "404 Not Found" for a previously "200 Ok" resource because it was deleted from the server.
Other. This would include any other type of change than those mentioned above. For example, similar to HTTP entity, URI-Ms of an embedded resource of a memento may have random values generated by JavaScript code, such as values associated with the current datetime, geolocation, weather, etc.

We found that 19.48% of mementos (3, 599 out of 18, 472 URI-Ms) have changed at least one time within the 10 downloads as Table 2 shows. All archives except `archive.is` have at least one memento with a change type of "other". Similarly, all archives had some mementos experience an "entity" change, except `archive.is`, `europarchive.org`, and `stanford.edu`. The percentage of mementos with "Response headers" change does not exceed 8%. The "Transient error" change occurs in the fewest archives, but as mentioned earlier, 54% of `perma-archives`'s mementos experienced this type of change. As Figure 3 shows, all types of changes are noticed in mementos from `archive.org`. Only five of these mementos experience an "entity" change. About 54% (98 out of 182) mementos from `perma-archives.org` produced different hash values because of the "Transient error" (i.e., returning "5xx" HTTP status code). Approximately half of `webarchive.org.uk`'s mementos produced different hashes because of the "other" type of change. In general, transient errors and some HTTP status code changes are not unexpected, but these types of changes will make consistently computing fixity of archived resources challenging.

TABLE 2: Number of mementos with at least one change.

Archive	URI-Ms	URI-Ms with changes (%)
web.archive.org	1,600	673 (42.06)
archive.is	1,600	6 (0.38)
archive.bibalex.org	1,600	300 (18.75)
webarchive.loc.gov	1,600	88 (0.55)
arquivo.pt	1,600	807 (50.44)
webcitation.org	1,600	365 (22.81)
wayback.vefsafn.is	1,600	378 (23.62)
wayback.archive-it.org	1,407	220 (15.64)
swap.stanford.edu	1,233	96 (7.79)
nationalarchives.gov.uk	1,011	37 (3.66)
europarchive.org	990	24 (2.42)
webharvest.gov	733	150 (20.46)
veebiarhiiv.digar.ee	518	16 (3.09)
webarchive.proni.gov.uk	477	16 (3.35)
webarchive.org.uk	362	256 (70.72)
collectionscanada.gc.ca	359	45 (12.53)
perma-archives.org	182	122 (67.03)
(total)		**3,599 (19.48)**

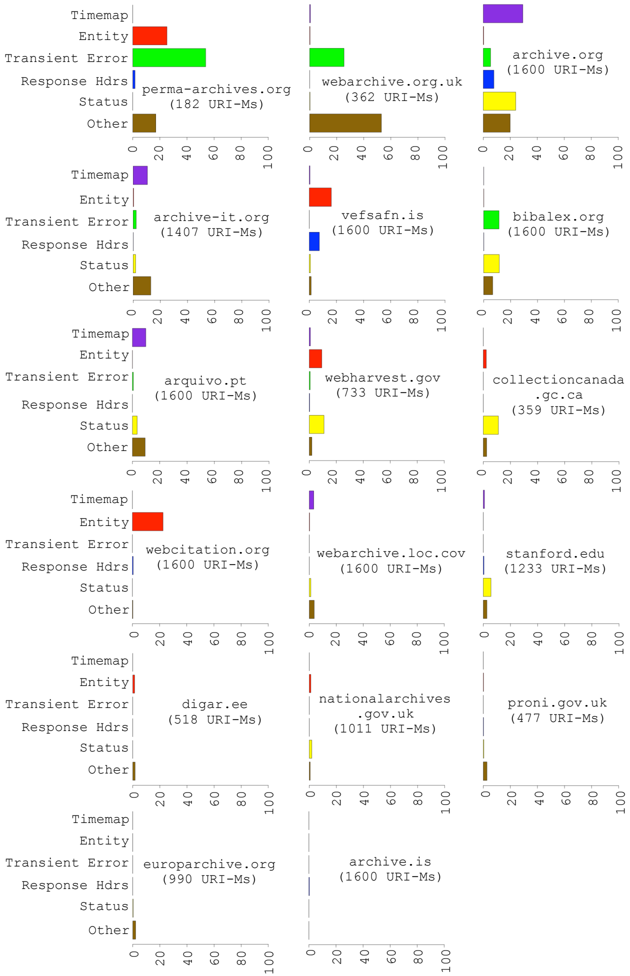

FIGURE 3: Different types of changes in mementos per archive.

4 CONCLUSIONS

A change in a memento may indicate malicious modification, but as we show, changes are caused by different playback-related issues. In general, we can categorize the cause of changes on the playback of mementos as: (1) expected changes, (2) unexpected non-malicious changes, and (3) unexpected malicious changes. In this article, we identify and quantify changes in the playback of mementos in general. We are currently working toward defining and quantifying each category. Being able to differentiate between malicious and non-malicious changes in mementos is important and will help us to introduce new approaches for verifying fixity of memento as conventional approaches regarding fixity are not applicable in web archives.

5 ACKNOWLEDGEMENTS

This work is supported in part by The Andrew W. Mellon Foundation (AMF) grant 11600663.

REFERENCES

[1] Jefferson Bailey. 2012. File Fixity and Digital Preservation Storage: More Results from the NDSA Storage Survey. `https://blogs.loc.gov/thesignal/2012/03/file-fixity-and-digital-preservation-storage-more-results-from-the-ndsa-storage-survey/`.

[2] PREMIS Working Group and others. 2005. Data dictionary for preservation metadata: final report of the PREMIS Working Group. *OCLC Online Computer Library Center & Research Libraries Group, Dublin, Ohio, USA, Final report* (2005).