Blockchain Can Not Be Used To Verify Replayed Archived Web Pages

Michael L. Nelson

Old Dominion University
Web Science & Digital Libraries Research Group
@WebSciDL, @phonedude_mln

With:
ODU: Michele C. Weigle, Mohamed Aturban
Los Alamos National Laboratory: Herbert Van de Sompel, Martin Klein

Supported in part by The Andrew Mellon Foundation and the National Science Foundation
This is not what you think it is…

How OpenTimestamps 'Carbon Dated' (almost) The Entire Internet With One Bitcoin Transaction

May 25, 2017

tl;dr: You can now use our searchable database to download Bitcoin timestamps for items in the Internet Archive.

While that title sounds like clickbait, the hard work of the Internet Archive made it much more accurate than it sounds. They’re a San Francisco non-profit digital library that provides free public access to collections of digitized materials, ranging from software applications/games, music, movies/videos, moving images, and millions of public-domain books. But they’re perhaps best known for the Wayback Machine, an archive of hundreds of billions of website snapshots, providing a priceless historical record of the evolution of the web.

In short, if it’s on the internet, there’s a pretty good chance the Internet Archive has a copy of it.

But is that copy the right copy?

OpenTimestamps helps answer that question by cryptographically proving data existed in the past, long before an attacker would have had an opportunity or reason to forge or modify that data.

The OpenTimestamps team has timestamped every item in the Internet Archive - about 750,000,000 files in total - and made those timestamps publicly available via a searchable database. This means that right now you can get timestamps for every book, movie, song, computer program, legal document, etc. in the thousands of collections in the archive. In the future we hope to be able to work with the Internet Archive to extend this to timestamping website snapshots, and our infrastructure will continue.


CNI Fall 2018 Membership Meeting, 2018-12-11,
@phonedude_mln, @WebSciDL
This is not what you think it is…

“…right now you can get timestamps for every book, movie, song, computer program, legal document, etc. in the thousands of collections in the archive. In the future we hope to be able to work with the Internet Archive to extend this to timestamping website snapshots…”

TL;DR

Web archiving is not file backup.

Backup = prevent, detect, repair changes

Web archiving = continuous changes to replicate the past

*Naïve fixity techniques are not applicable for web archiving.*
Monitoring Fixity To Detect Tampering ==
Endless False Positives

A simplified workflow of web archiving

1) live web site
https://climate.nasa.gov/vital-signs/carbon-dioxide/

2) Crawled by any of several archival crawlers

3) Result stored in a WARC File
   (like tar or zip, but for Web archives)

4) WARC files are indexed, served by replay software
   (there are several variations of Wayback Machine)

5) User chooses date of capture (Memento-Datetime)

6) Page replayed with banner, rewritten links, etc.
(apologies to Peter Arnett)

“In order to save the page, we had to completely change it”

Yes, some archives (including most versions of Wayback) provide “raw” access, but modifications can still happen (how/why is beyond the scope of this presentation).
I've got mad HTML skillz

- January 31-February 1, 2019, NYC, ACM Publications Board Meeting
- December 10-11, 2018, Washington DC, CNI Fall 2018 Membership Meeting
  https://www.cni.org/events/membership-meetings/upcoming-meeting/fall-2018
- November 5, 2018, Chapel Hill, NC, Symposium on Blockchain and Trusted Repositories
  https://theknowledgetrust.org/events/symposium-on-blockchain-and-trusted-repositories/
- November 2, 2018, Blacksburg, VA, Va Tech Computer Science Graduate Research Seminar
  https://cs.vt.edu/News/Seminars/MichaelNelson.html
- September 26-29, 2018, Los Alamos, NM, LANL Scholarly Orphans Meeting
- June 9-15, 2018, Power Tour
- June 3-6, 2018, Fort Worth, TX, JCDL 2018
  https://2018.jcdl.org/
- May 17-20, 2018, Ocean City MD
- May 4-5, 2018, Kentucky Derby
- April 18-20, 2018, Arlington, VA, NSF Panel
- March 22-24, 2018, NYC, National Forum on Ethics and Archiving the Web
  https://eaw.rhizome.org/
- February 9, 2018, Washington DC, NEH-ODH Project Directors Meeting
  https://www.neh.gov/divisions/odh/grant-news/odh-ten-our-tenth-anniversary-project-directors-meeting
- February 5-6, 2018, NYC, ACM Publications Board
- December 11-12, 2017, St. Louis, MO, DocNow Advisory Board Meeting
- December 7-8, 2017, NYC, 3rd ACM Workshop on Reproducibility in Publication
- October 14-18, 2017, Los Alamos National Laboratory
- September 21-22, 2017, New York, NY, ACM Publications Board meeting
- June 19-23, 2017, Toronto, CA, JCDL 2017
  http://2017.jcdl.org/

https://www.cs.odu.edu/~mln/travel.html

CNI Fall 2018 Membership Meeting, 2018-12-11,
@phonedude_mln, @WebSciDL
Same page, archived at IA

Archival Metadata
The banner tells the user the original URL, which archive the page resides in, when it was archived, how many copies, etc.

Links are rewritten to point back into the archive, not the live web.


CNI Fall 2018 Membership Meeting, 2018-12-11,
@phonedude_mln, @WebSciDL
Same page, archived at IA

$ curl -s https://www.cs.odu.edu/~mln/travel.html | head -5
<body bgcolor=white>
<pre>
- January 31-February 1, 2019, NYC, ACM Publications Board Meeting
$ curl -s https://www.cs.odu.edu/~mln/travel.html | wc
  585  2361  26471
</pre>

- January 31-February 1, 2019, NYC, ACM Publications Board Meeting
- May 17-20, 2018, Ocean City MD
- May 4-5, 2018, Kentucky Derby
- April 18-20, 2018, Arlington, VA, NSF Panel
- March 22-24, 2018, NYC, National Forum on Ethics and Archiving the Web
  https://eaw.rhizome.org/
- February 9, 2018, Washington DC, NEH-ODH Project Directors Meeting
  https://www.neh.gov/divisions/odh/grant-news/odh-ten-our-tenth-anniversary-project-directors-meeting
- February 5-6, 2018, NYC, ACM Publications Board
- December 11-12, 2017, St. Louis, MO, DocNow Advisory Board Meeting
- December 7-8, 2017, NYC, 3rd ACM Workshop on Reproducibility in Publication
- October 14-18, 2017, Los Alamos National Laboratory
- September 21-22, 2017, New York, NY, ACM Publications Board meeting
January 31-February 1, 2019, NYC, ACM Publications Board Meeting

$ curl -s https://www.cs.odu.edu/~mln/travel.html | head -5
<body bgcolor=white>
<pre>
-January 31-February 1, 2019, NYC, ACM Publications Board Meeting
</pre>

$ curl -s https://www.cs.odu.edu/~mln/travel.html | wc
585  2361  26471

<body bgcolor=white>
<pre>
-January 31-February 1, 2019, NYC, ACM Publications Board Meeting
</pre>

618  2472  33787

<body bgcolor=white>
<pre>
-January 31-February 1, 2019, NYC, ACM Publications Board Meeting
</pre>

618  2472  33787
January 31-February 1, 2019, NYC, ACM Publications Board Meeting

December 10-11, 2018, Washington DC, CNI Fall 2018 Membership Meeting
https://www.cni.org/events/membership-meetings/upcoming-meeting/fall-2018

November 5, 2018, Chapel Hill, NC, Symposium on Blockchain and Trusted Repositories
https://theknowledgetrust.org/events/symposium-on-blockchain-and-trusted-repositories/

November 2, 2018, Blacksburg, VA, Va Tech Computer Science Graduate Research Seminar
https://cs.vt.edu/News/Seminars/MichaelNelson.html

September 26-29, 2018, Los Alamos, NM, LANL Scholarly Orphans Meeting

June 9-15, 2018, Power Tour

June 3-6, 2018, Fort Worth, TX, JCDL 2018
https://2018.jcdl.org/

May 17-20, 2018, Ocean City MD

May 4-5, 2018, Kentucky Derby

April 18-20, 2018, Arlington, VA, NSF Panel

March 22-24, 2018, NYC, National Forum on Ethics and Archiving the Web
https://eaw.rhizome.org/

February 9, 2018, Washington DC, NEH-ODH Project Directors Meeting
https://www.neh.gov/divisions/odh/grant-news/odh-ten-our-tenth-anniversary-project-directors-meeting

February 5-6, 2018, NYC, ACM Publications Board

December 11-12, 2017, St. Louis, MO, DocNow Advisory Board Meeting

December 7-8, 2017, NYC, 3rd ACM Workshop on Reproducibility in Publication

October 14-18, 2017, Los Alamos National Laboratory

September 21-22, 2017, New York, NY, ACM Publications Board meeting

June 19-23, 2017, Toronto, CA, JCDL 2017
http://2017.jcdl.org/
If we just had isolated, static pages (e.g., individual jpegs, pdfs, mp3s) then there’d be no problem.

But HTML has:
1) links,
2) embedded resources (including iframes), and
3) Javascript, which can modify the HTML.

And HTTP has no “bulk download”, so you can’t grab an entire site instantaneously.
We could hash the WARC file

```
$ md5sum climate.nasa.gov.warc.gz
652853fe1bc8cb273cdf73aad8a489ca climate.nasa.gov.warc.gz
```

But this nasa.gov page contains:
• 201 images
• 19 Javascript files
• 3 CSS files

At a large archive like IA they could be in multiple WARC files; worst case is 224 WARC files.

In general, the WARC file(s) corresponding to the replayed page will be unavailable to the user replaying the page.
We can detect changes in the root HTML

https://ws-dl.blogspot.com/2017/12/2017-12-11-difficulties-in-timestamping.html

CNI Fall 2018 Membership Meeting, 2018-12-11,
@phonedude_mln, @WebSciDL
But what if the change is in an embedded resource?
Clearly we need to render the entire page, then compute the hash.

Unfortunately, that’s not easy.
Load the archived page, get an eagle

http://www.fws.gov/
Hit “reload”, get a tiger

Hit “reload” again, get a mountain

function random_imglink() {

    var ry=Math.floor(Math.random(1)*myimages.length);

    if (ry==0) 
        ry=1

    document.write('<a href=""'+imagelinks[ry]+'"'><img src="'+myimages[ry]+'" border="0" alt="The Open Spaces Blog. A Talk on the Wild Side. Click to Read"/></a>')
}
Actually, the fws.gov example was super easy; most changes are much harder to trace.

Mohamed Aturban, unpublished, memento:
Temporal violations: reconstructing pages that never existed on the live web
(examples below are transient; sometimes you get the 1st image, sometimes the 2nd image)

(a) Downloaded on November 16, 2017. Its hash ends in “4465eb88c7”.
  embedded in umich.edu memento, archived in perma.cc
  2nd image is compressed (12209 vs. 19448 bytes); 2nd image modified in 2017-03, but replayed in a 2017-01 page

(b) Downloaded on December 25, 2017. Its hash ends in “021e7b224b”.
  (c) Comparing images (a) and (b) using [9] (mismatched pixels in pink).

(embedded in copybogger.com memento, archived in archive.org
  2nd image modified in 2017-12, but replayed in a 2017-11 page; blackout for privacy)

Temporal violations: https://ws-dl.blogspot.com/2015/12/2015-12-08-evaluating-temporal.html
CNI Fall 2018 Membership Meeting, 2018-12-11,
@phonedude_mln, @WebSciDL
1 WARC file, 2 Wayback Machines, 3 Browsers = 6 different replays

http://wayback.archive-it.org/all/20130106140348/http://www.harvard.edu/
see also. https://ws-dl.blogspot.com/2016/12/2016-12-20-archiving-pages-with.html
Experiment Design
Sample 16k+ Mementos from 17 Web Archives

<table>
<thead>
<tr>
<th>Archive</th>
<th>URI-Ms</th>
</tr>
</thead>
<tbody>
<tr>
<td>perma-archives.org</td>
<td>182</td>
</tr>
<tr>
<td>bibalex.org</td>
<td>199</td>
</tr>
<tr>
<td>webarchive.org.uk</td>
<td>349</td>
</tr>
<tr>
<td>bac-lac.gc.ca</td>
<td>351</td>
</tr>
<tr>
<td>proni.gov.uk</td>
<td>469</td>
</tr>
<tr>
<td>digar.ee</td>
<td>488</td>
</tr>
<tr>
<td>webharvest.gov</td>
<td>712</td>
</tr>
<tr>
<td>internetmemory.org</td>
<td>979</td>
</tr>
<tr>
<td>nationalarchives.gov.uk</td>
<td>994</td>
</tr>
<tr>
<td>stanford.edu</td>
<td>1222</td>
</tr>
<tr>
<td>archive-it.org</td>
<td>1383</td>
</tr>
<tr>
<td>archive.is</td>
<td>1396</td>
</tr>
<tr>
<td>web.archive.org</td>
<td>1566</td>
</tr>
<tr>
<td>arquivo.pt</td>
<td>1569</td>
</tr>
<tr>
<td>webcitation.org</td>
<td>1585</td>
</tr>
<tr>
<td>vefsafn.is</td>
<td>1589</td>
</tr>
<tr>
<td>loc.gov</td>
<td>1594</td>
</tr>
<tr>
<td>Total</td>
<td>16627</td>
</tr>
</tbody>
</table>
Periodically Replay Each *Archived* Page

35 times, from Nov. 2017 – Oct. 2018

For each replay, we download both the rewritten version and the “raw” version (where possible).

Above example: http://perma-archives.org/warc/20170101182813/http://umich.edu/
Periodically Replay Each *Archived* Page

35 times, from Nov. 2017 – Oct. 2018

For each replay, we download both the rewritten version and the “raw” version (where possible).

Above example: http://perma-archives.org/warc/20170101182813/http://umich.edu/

Partial archive outage because of security / maintenance upgrade
Periodically Replay Each *Archived* Page

35 times, from Nov. 2017 – Oct. 2018

For each replay, we download both the rewritten version and the “raw” version (where possible).

Above example: http://perma-archives.org/warc/20170101182813/http://umich.edu/
In 11 months, 11% of the URLs Disappeared or Changed

820 were renamed & required manual rediscovery
979 disappeared & have not yet been rediscovered
europarchive.org became internetmemory.org

URI-Ms like this:
collection.europarchive.org/nli/20130117165443/http://bbc.co.uk/news/

changed domains and became like this:
collections.internetmemory.org/nli/20130117165443/http://bbc.co.uk/news/

europarchive.org
europarchive.org is now spam
internetmemory.org is now down
979 pages lost

curl -I collection.europarchive.org/nli/20130117165443/http://bbc.co.uk/news/
HTTP/1.1 301 Moved Permanently
Date: Mon, 10 Dec 2018 04:30:50 GMT
Server: Apache
Expires: Mon, 10 Dec 2018 05:30:50 GMT
Cache-Control: max-age=3600
Location: http://europarchive.org
Connection: close
Content-Type: text/html; charset=UTF-8

curl -I collections.internetmemory.org/nli/20130117165443/http://bbc.co.uk/news/
HTTP/1.1 403 Forbidden
Date: Mon, 10 Dec 2018 04:31:51 GMT
Server: Varnish
X-Varnish: 71167297
Content-Type: text/html; charset=utf-8
Retry-After: 5
Content-Length: 252
Connection: keep-alive
Telling humans your domain is about to change is nice, but please tell robots too…

See: https://tools.ietf.org/id/draft-wilde-sunset-header-03.html
webarchive.proni.gov.uk now uses Archive-It

The top-level site webarchive.proni.gov.uk still exists, but deep links to URI-Ms are now 404.

469 pages required manual rediscovery.

HTTP/1.1 404 Not Found
Date: Mon, 10 Dec 2018 05:15:56 GMT
Server: Apache/2.4.18 (Ubuntu)
Content-Type: text/html; charset=iso-8859-1

HTTP/1.1 200 OK
...
www.collectionscanada.gc.ca became webarchive.bac-lac.gc.ca:8080
(no, really – port 8080)

And deep links to URI-Ms now redirect to the top of the new site, which means 351 pages required manual rediscovery:

$ curl -IL http://www.collectionscanada.gc.ca/webarchives/20061027192435/http://www.state.gov/ HTTP/1.0 302 Found
...
HTTP/1.1 302 Found
Location: http://webarchive.bac-lac.gc.ca/?lang=en
Date: Mon, 10 Dec 2018 04:46:24 GMT
...
HTTP/1.1 200
Date: Mon, 10 Dec 2018 04:46:24 GMT
7 out of 8 pages produced > 1 hash over 11 months

<table>
<thead>
<tr>
<th>Archive Name</th>
<th>URI-Ms</th>
<th>URI-Ms with at least two hashes</th>
</tr>
</thead>
<tbody>
<tr>
<td>webarchive.loc.gov</td>
<td>1,594</td>
<td>1,235 (77.47%)</td>
</tr>
<tr>
<td>vefsafn.is</td>
<td>1,589</td>
<td>1,133 (71.30%)</td>
</tr>
<tr>
<td>webcitation.org</td>
<td>1,585</td>
<td>981 (61.89%)</td>
</tr>
<tr>
<td>arquivo.pt</td>
<td>1,569</td>
<td>1,563 (99.61%)</td>
</tr>
<tr>
<td>archive.org</td>
<td>1,566</td>
<td>1,430 (91.31%)</td>
</tr>
<tr>
<td>archive.is</td>
<td>1,396</td>
<td>1,364 (97.70%)</td>
</tr>
<tr>
<td>archive-it.org</td>
<td>1,383</td>
<td>1,383 (100%)</td>
</tr>
<tr>
<td>swap.stanford.edu</td>
<td>1,222</td>
<td>1,005 (82.24%)</td>
</tr>
<tr>
<td>nationalarchives.gov.uk</td>
<td>994</td>
<td>978 (98.39%)</td>
</tr>
<tr>
<td>internetmemory.org</td>
<td>979</td>
<td>979 (100%)</td>
</tr>
<tr>
<td>webharvest.gov</td>
<td>712</td>
<td>712 (100%)</td>
</tr>
<tr>
<td>digar.ee</td>
<td>488</td>
<td>308 (63.11%)</td>
</tr>
<tr>
<td>proni.gov.uk</td>
<td>469</td>
<td>469 (100%)</td>
</tr>
<tr>
<td>bac-lac.gc.ca</td>
<td>351</td>
<td>351 (100.0%)</td>
</tr>
<tr>
<td>webarchive.org.uk</td>
<td>349</td>
<td>348 (99.71%)</td>
</tr>
<tr>
<td>archive.bibalex.org</td>
<td>199</td>
<td>199 (100%)</td>
</tr>
<tr>
<td>perma-archives.org</td>
<td>182</td>
<td>182 (100%)</td>
</tr>
</tbody>
</table>

You cannot replay twice the same archived page
(apologies to Heraclitus)
More Archived Pages Changed Every Time Than Never Changed

Never changed:
2007 URI-Ms (1 in 8)

Always changed:
2773 URI-Ms (1 in 6)
A metaphor for replaying archived web pages

https://www.youtube.com/watch?v=ekO3Z3XWa0Q
https://en.wikipedia.org/wiki/Monty_Python_and_the_Holy_Grail

CNI Fall 2018 Membership Meeting, 2018-12-11,
@phonedude_mln, @WebSciDL
King of Swamp Castle: live web/ground truth
Guard: archival replay

$ echo "Make sure the prince doesn't leave this room until I come and get him." | md5
57facbb2734d36cb823f4230cc07b888

$ echo "Not to leave the room even if you come and get him." | md5
3ba0a2359d63f43cbe9e11fb5a179b8d

$ echo "Until you come and get him, we're not to enter the room." | md5
ade3539aaa8a6d8724193e9a37f3ca6d

$ echo "We don't need to do anything apart from just stop him entering the room." | md5
ea812f5b997aa42a8f293bd1ee536fd0

$ echo "Oh yes, we'll keep him in here, obviously. But if he had to leave, and we went with him..." | md5
55d184b77d99eed6367535ef3c05d7aa

$ echo "Oh, yes of course. I thought you meant him! You know it seemed a bit daft me having to guard him when he's a guard." | md5

https://www.youtube.com/watch?v=ekO3Z3XWa0Q
https://en.wikipedia.org/wiki/Monty_Python_and_the_Holy_Grail
Archival replay & blockchain: building a castle in a swamp

• Fixity checks only work when it’s clear what to hash
  – Hash only the root HTML and modifications are possible via embedded resources (false negatives)
  – Recursively hash all embedded resources and you’ll rarely get the same hash (false positives)

• Replay is working as designed, it’s not something that will be “fixed”
  – we need server-side support for auditing, and archive-aware hashing functions

• There is increasing incentive to attack existing archives and create networks of fake archives