Difficulties of Timestamping Archived Web Pages

Mohamed Aturban,
Michael L. Nelson, and
Michele C. Weigle

Presented by Peter Foytik

Aturban, Mohamed, Michael L. Nelson, and Michele C. Weigle. "Difficulties of timestamping archived web pages." *arXiv preprint arXiv:1712.03140* (2017).

High Level Topic of Paper

- Web archived sites are run by many organizations
- Everyone has some archived sources that they trust (maybe some they don’t)
  - We often take for granted this trust
- What types of vulnerabilities exist with web archival systems?
- Can Blockchain be a solution?
- What are the list of requirements needed to achieve this solution?

Compromised Data

- Centrally stored and controlled data provides a large incentive
  - Holders of data can alter/use/sell/ the data and are required to maintain responsibility
  - Attackers can gain access easier if the reward is centrally controlled
  - Governments gain access if there is a central authority they can enforce policy on
- Access to data can:
  - Control narrative
  - Erase evidence of history
  - Cause mistrust or confusion
  - Sell or take out for a profit

Information Security "CIA" Triad

- Confidentiality
- Integrity
- Availability

Information Security "CIA" Triad

- Confidentiality
- Integrity
- Availability

Background of Wayback and Memento

- The paper covers the background
  - on the use of the Wayback Machine internet archive
  - and the value and use of the Memento HTTP protocol extension
- 4 URI teminology used:
  1. **URI-R**: identity of an original resource from the live Web
  2. **URI-M**: identity of an archived version (memento) of the original resource at a particular point in time
  3. **URI-T**: a resource (TimeMap) that provides a list of mementos (URI-Ms) for a particular original resource
  4. **URI-G**: a resource (TimeGate) that supports content negotiation based on datetime to access prior versions of an original resource

Types of Attacks/Vulnerabilities

- Four Types of Vulnerabilities of the Wayback Machine are described referenced from Lerner et al.
    1. Archive-Escapes
    2. Same-Origin Escapes
    3. Archive-Escapes + Same-Origin Escapes
    4. Anachronism-Injection

Lerner, A., Kohno, T., Roesner, F.: Rewriting history: Changing the archived web from the present. In: Proceedings of the 16th ACM conference on Computer and Communications Security (CCS) (2017)

Archive-Escapes

- URL rewriting occurs for archived web pages
  - All URL's are re-written to point to archived version of the links
- When URL's are incorrectly re-written or not re-written
  - users snapshot view of the archive might be wrong
  - containing some archived content and some live content
- If the link is to the live web "Escaping the archive" then it could point to malicious scripts or content

Example Archive-Escape

- Image could be incorrectly referenced in an archive
  - Top image shows the correct image
  - Bottom image is of the archive with the link pointing to an incorrect image pointing outside the archive

Same-Origin Escapes

- Archives save content that comes from all domains of the web page at the time
- Additional domains such as sources of advertisements can also be collected
- In live web pages iframes prevent third parties from accessing or modifying data
- The policy of preventing the cross-origin access is called the same-origin policy
- Same-origin is ineffective in the archival context
  - In archives, code within embed iframes are now executable from a single domain

Archive-Escapes + Same-Origin Escapes

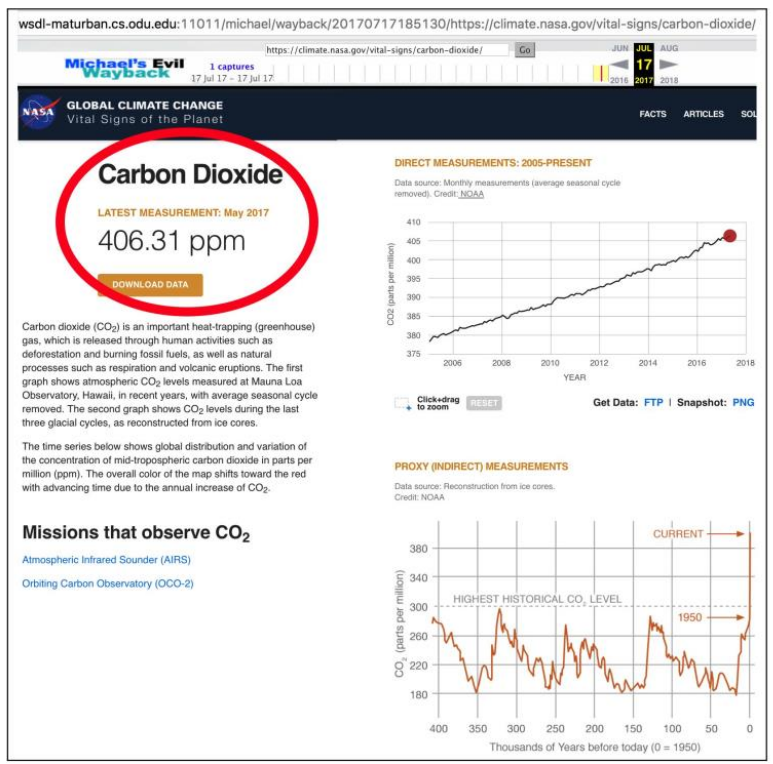
- Combination of Archive-Escapes and Same-Origin Escapes
- The attacker first utilizes the same-origin vulnerability in the iframe at the time of archival
- Then the attacker can redirect the archival link through the iframe manipulation to outside links making an archive-escape

Anachronism-Injection

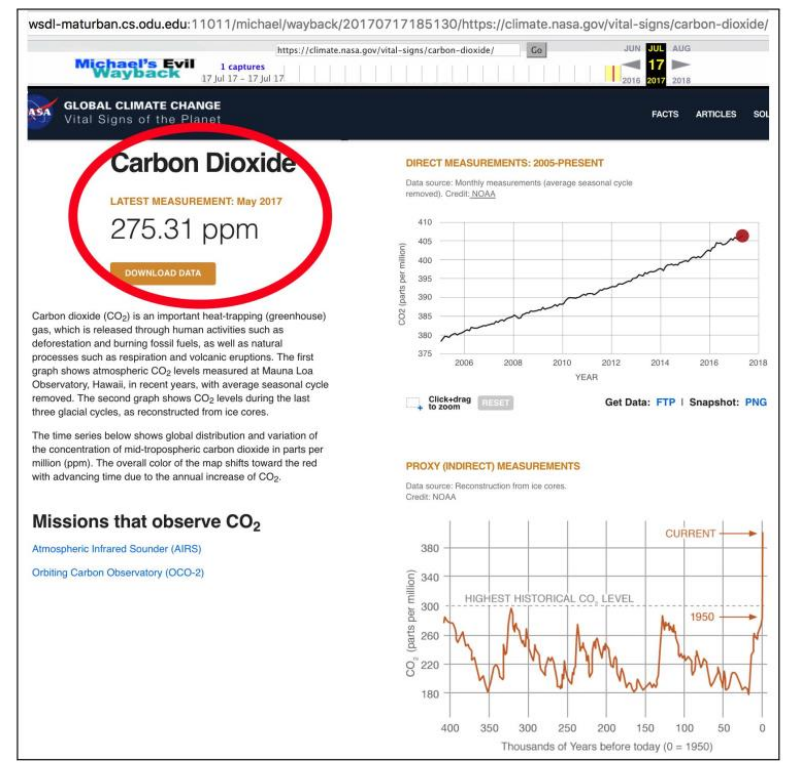
- When client browsers are redirected to timestamps very far in time from the original page
- Wayback machine will attempt to reproduce a snapshot in time with all web elements
- when time of archive not available, attempt to utilize a version in time with the smallest delta in time

The Evil Wayback Machine

- How can we independently verify the archived website we are looking at?
- Can we do this without requiring the trust of a 3rd party?
  - 3rd party includes:
    - Trusted authority to say it is true
    - Many redundant sources to verify truth



(a) Accessing the archived page in August 2017 (CO_2 was 406.31 ppm)

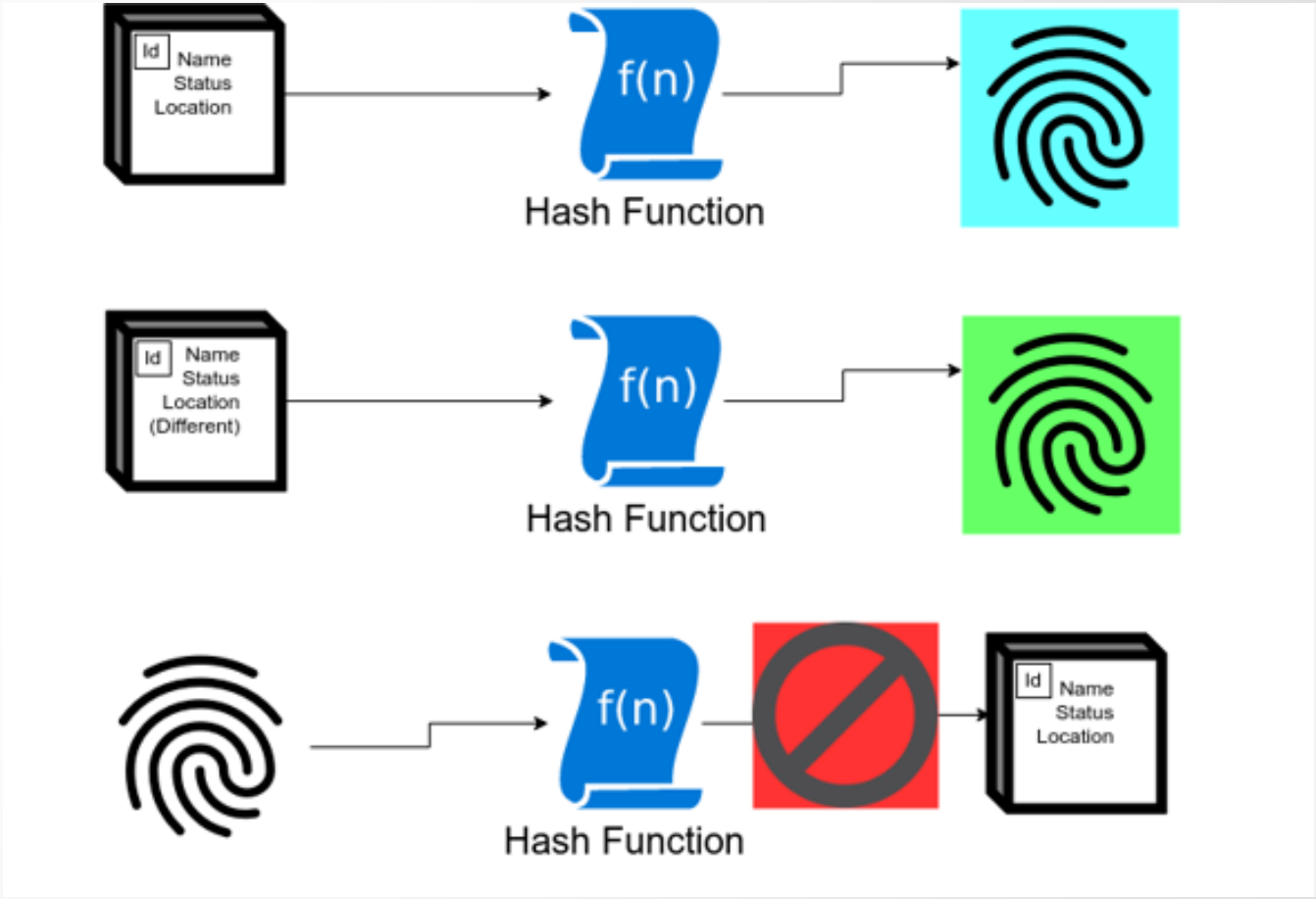


(b) Accessing the same archived page in October 2017 (CO_2 became 270.31 ppm)

Cryptography and Blockchain Background

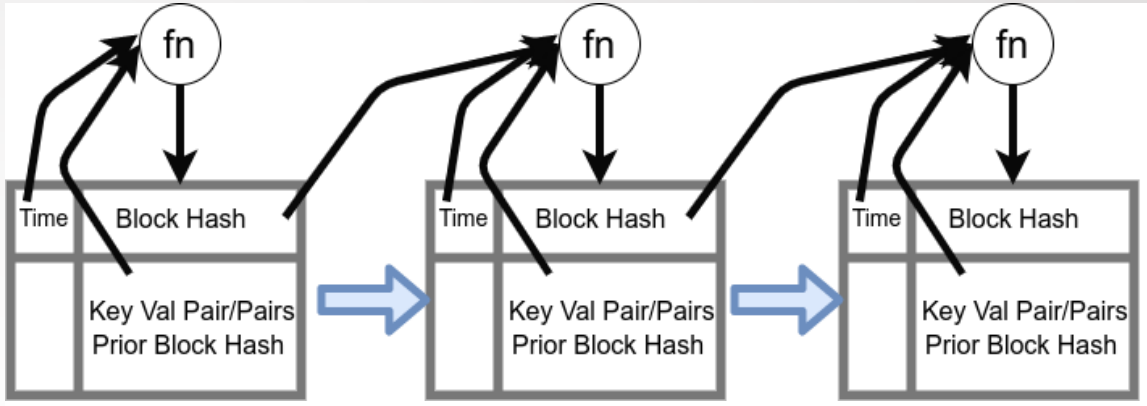
Hash Function Background

- Publicly known function
- One to one representation
- Not Reversable
- Compressed fixed length output
- Integrity is everything
  - The state of being unimpaired; soundness.
- Some Uses:
  - Checksum
  - Check digits
  - Fingerprints

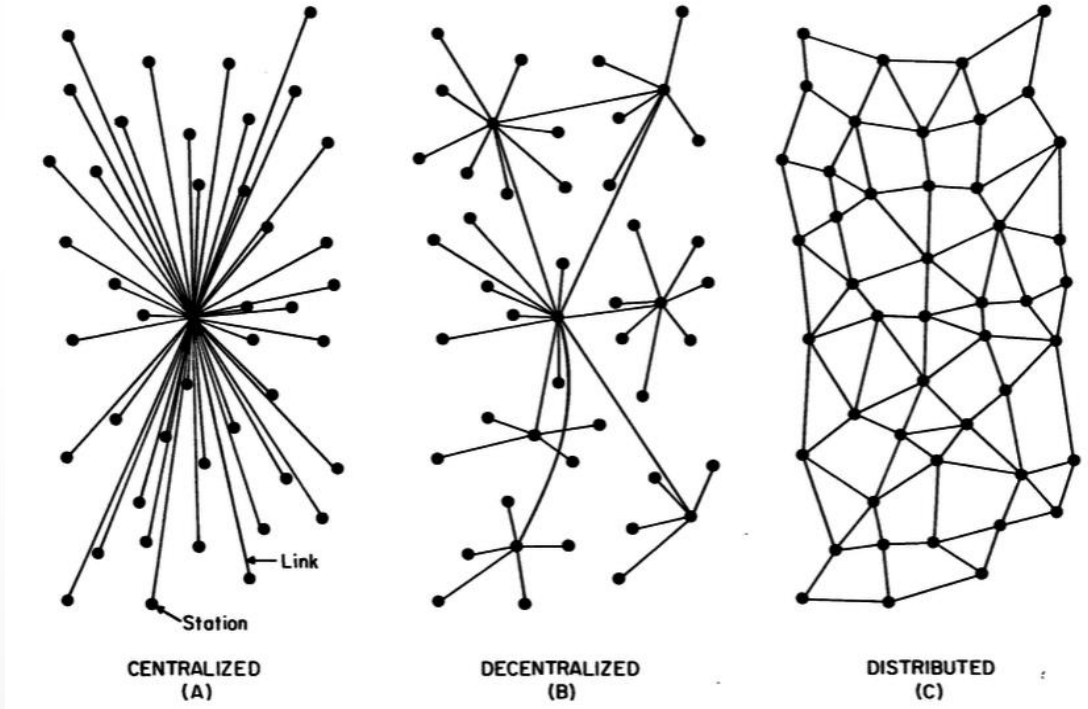


Blockchain Background

- Simply a ledger
  - Records in time
  - Tracks states or identities
- Merkle Tree/ Hash Tree
  - Hash of Hash
- Centralized vs. Decentralized
  - Self Soverignty
- Requires a consensus protocol to specify the order of transactions



Foytik, Peter, and Sachin S. Shetty. "Blockchain Evaluation Platform." *Blockchain for Distributed Systems Security* (2019): 275-310.



Baran, P. (1964). On Distributed Communications, Memorandum RM-3420-PR.

Bitcoin Blockchain and its Consensus

- Really the only "at scale" production blockchain system

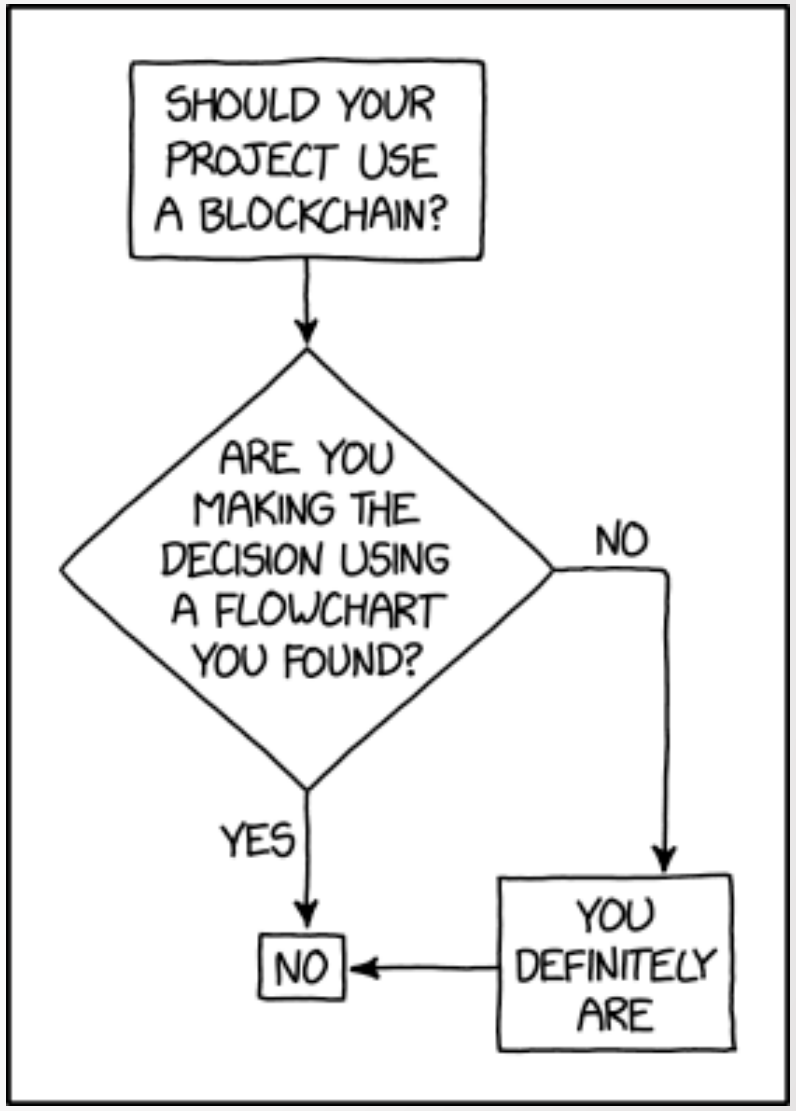
- Transaction Order is determined by a computational race
  - Who can guess the best hash value first! (proof of work)
    - "Best hash" gets harder based on the hash power of the network
    - Difficulty adjusts in order to keep block rate near 10 minute intervals
  - New transactions build on the longest chain that the miner sees
  - Inevitably longest chains will always succeed
    - Short chain transactions will need to re-work to add to the longest chain
    - Important point to how Nakamoto consensus works

Nakamoto, S.: Bitcoin: A peer-to-peer electronic cash system (2008), https://bitcoin.org/bitcoin.pdf

Blockchain Use

- Blockchains are often talked about as "the new solution to everything"
- Blockchains **do not**:
  - Add privacy
  - Add security
  - Add confidentiality
  - Add performance
  - Act as a new database

- Blockchain technology brings one aspect to the tech toolbox
  - **<u>Decentralized Integrity</u>**



"Blockchains are like grappling hooks, in that it's extremely cool when you encounter a problem for which they're the right solution, but it happens way too rarely in real life." https://xkcd.com/2267/

Existing Solutions to these problems

- OriginStamp
  - Allows users to submit plain text, a hash value, or any file format
  - The data is stored in the user's browser
  - The hash is then sent to the OriginStamp server
  - Once a day OriginStamp transacts a hash of all the hashes it has received to the bitcoin network
  - The user can then verify the text or file is authentic by comparing it to the hash stored on the bitcoin ledger (query'ed through OriginStamp API or their website)
- Others include
  - Chainpoint, Tangible.io, Proof of Existence, and OpenTimeStamps

Papers Questions and Exploration

Web Archiving today is a system that is for the most part many independent centralized servers

- We currently trust the hosts of the centralized servers

Can a cryptographic protocol that allows us to prove historical content without the need to trust a third party be built or utilized?

What requirements would the archiving community have for a system that cryptographically verifies the web content?

- **8 Requirements are derived and presented**

Archive Requirements for Blockchains

Requirement 1: Repeatable Hash Values

Requirement 1: Repeatable hash values

If we download a memento $URI\text{-}M_x$ at time t_n (denoted as $URI\text{-}M_x@t_n$), download the same memento at time t_m (denoted as $URI\text{-}M_x@t_m$), and apply a hash function H on the content of $URI\text{-}M_x@t_n$ and $URI\text{-}M_x@t_m$, then $H(URI\text{-}M_x@t_n) = H(URI\text{-}M_x@t_m)$

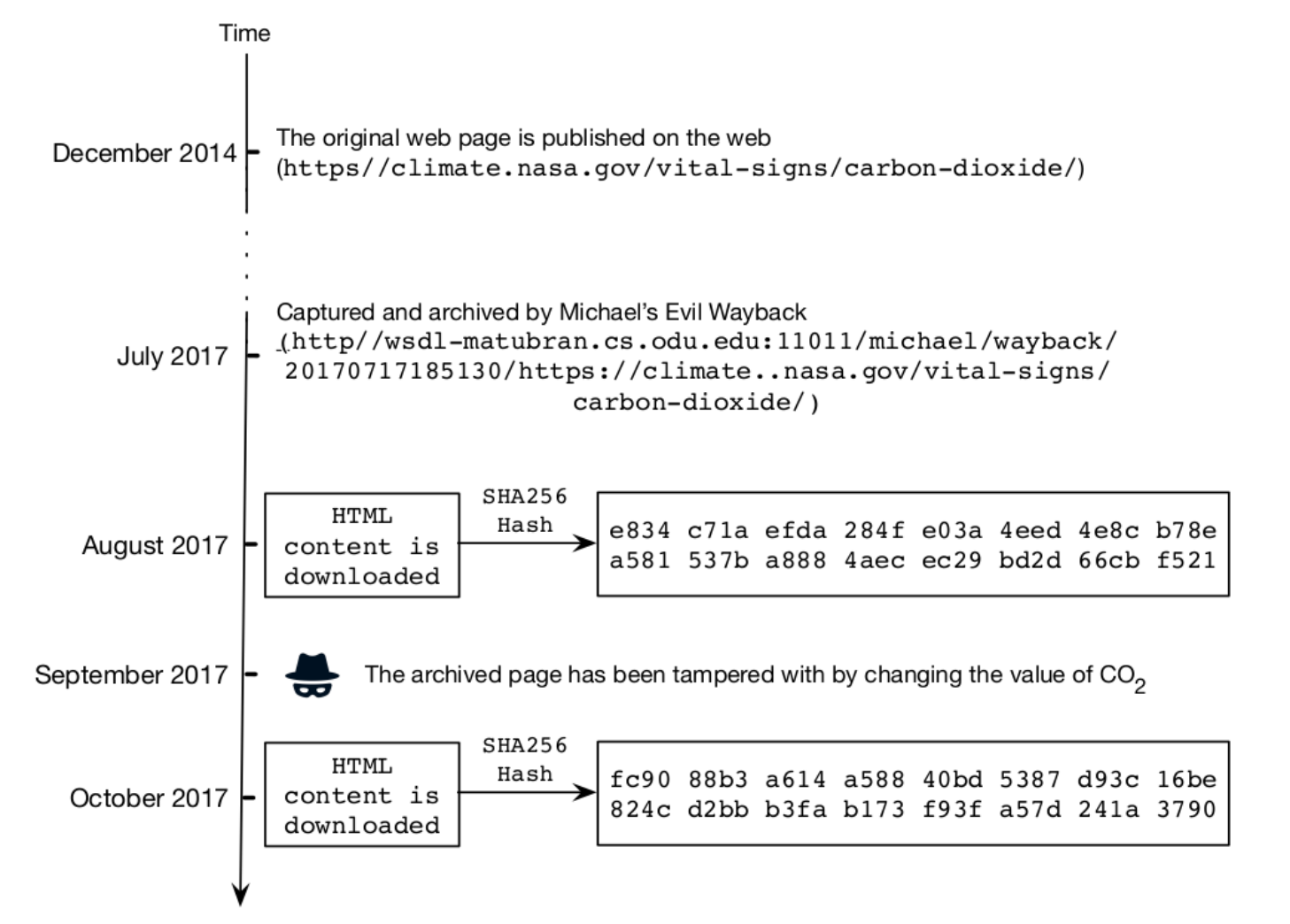
Example Requirement 1: success

- Hash functions can quickly produce values for the raw html content of an archive
  - Appending the id_ tag at the end of the date provides results of raw html from web.archive.org
- Hash functions show that grabbing the same raw html of an archive at different times produces the same Hash value
- Below is an example of cnn.com archive for july 24, 2013 and its associated md5 hash value
  - The hash was generated at different times producing the same hash value (highlighted)

```
(base) pfoytik@pfoytik-Latitude-5490:~$ curl --silent http://web
.archive.org/web/20130724144801id_/http://www.cnn.com/ | md5sum
9063723f72651e2f8aa5046281a5be78  -
(base) pfoytik@pfoytik-Latitude-5490:~$ curl --silent http://web
.archive.org/web/20130724144801id_/http://www.cnn.com/ | md5sum
9063723f72651e2f8aa5046281a5be78  -
```

Example Requirement 1: unsuccess

- For HTML content only the hash value can be obtained
- If changes are made to the archive a completely different hash function is generated
- Easily shows that a change has been made

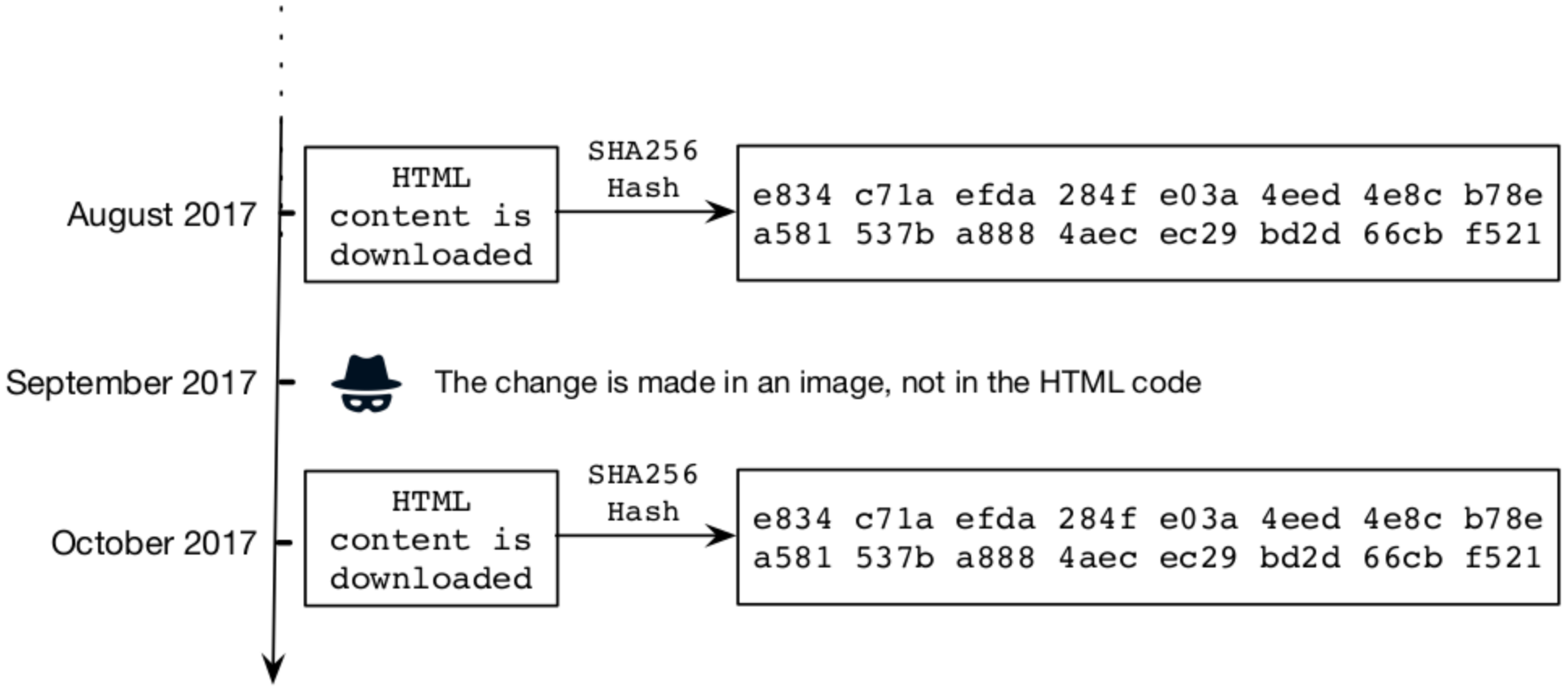


Requirement 2: Hash a Composite Memento

Requirement 2: Hash a composite memento

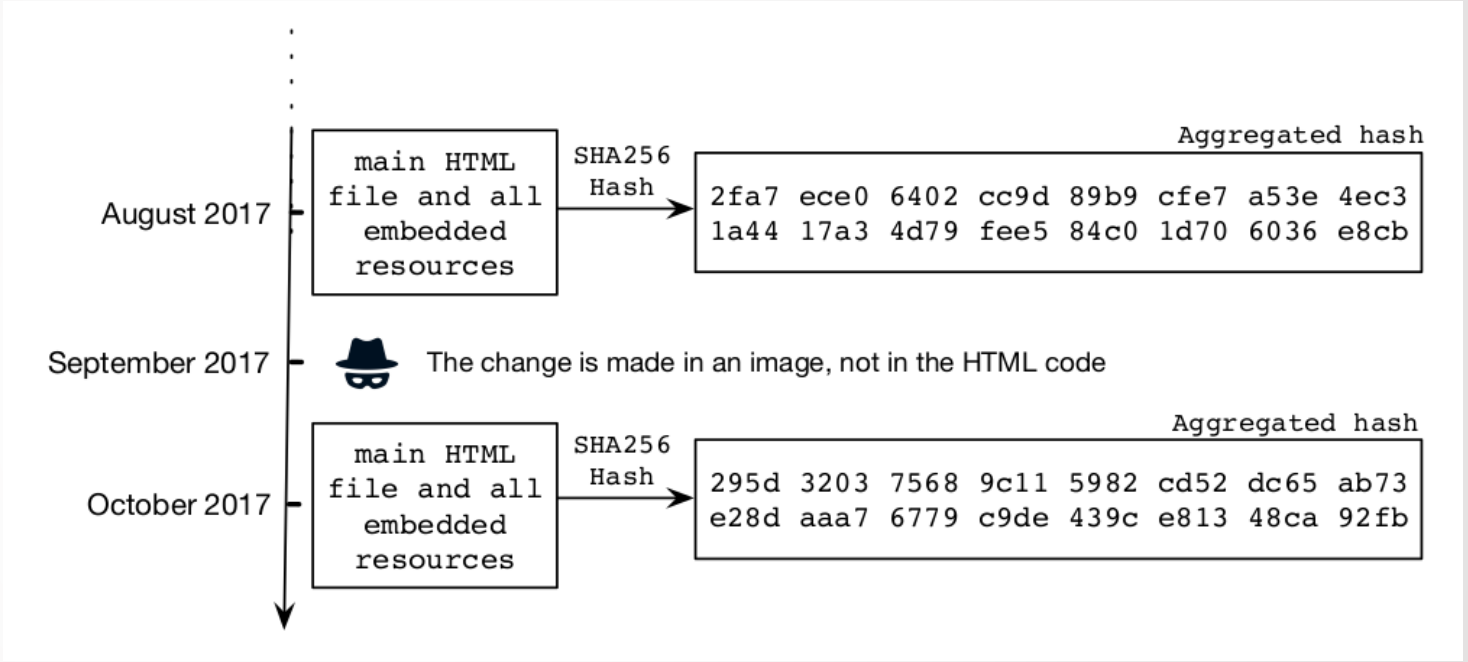
We should hash a composite memento. In most cases this would include hashing the main HTML file as well as other embedded resources in the memento, such as images, style sheets, JavaScript files, iframes, and others.

Example Requirement 2: unsuccess



Example Requirement 2: success

- An aggregated hash is utilized
  - Each element in the html is hashed from its source
  - The hash values are then combined in a single hash value
    - A Hash of the Hash values



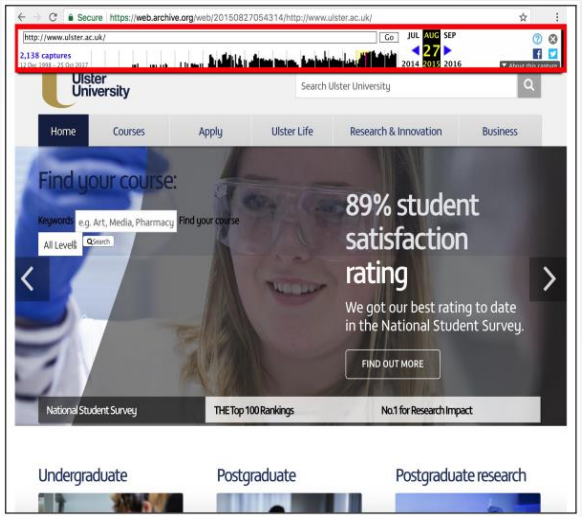
Requirement 3: Avoid Archive-Specific Resources

Requirement 3: Avoid archive-specific resources

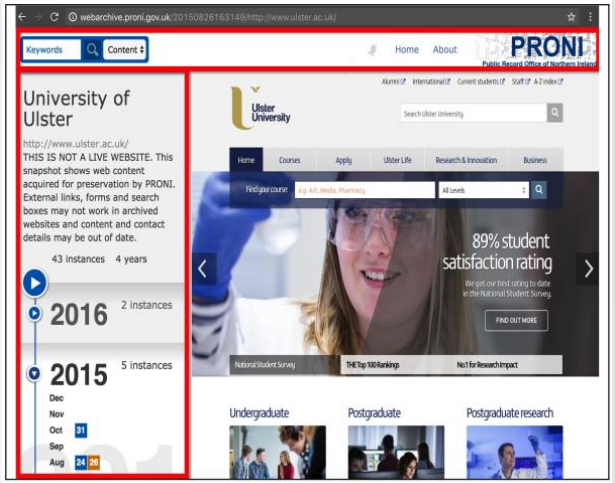
Resources added by archives are not part of the original content and should not be included in the hash calculation.

Example Requirement 3

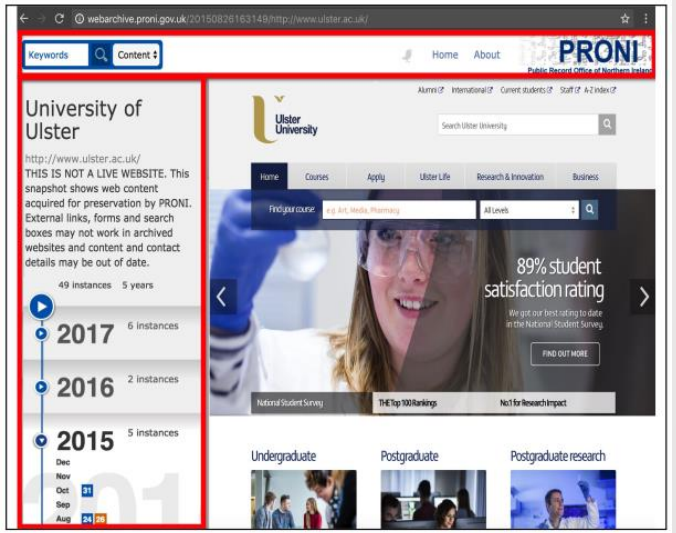
- Archival systems provide user interfaces and banners that should not be part of the hash
- The hash should only represent the content that needs to have the integrity check



(a) A Memento from Internet Archive



(b) From Proni Archive accessed in 2016



(c) Same memento accessed in 2017

Requirement 4: No Resources From the Live Web

Requirement 4: No resources from the live web

No resource located on the live web should be part of the hashing process.

Example Requirement 4: archive-escape

- Some URIs are not rewritten
  - Produced dynamically
  - Events triggered by client-side JavaScript
  - These are often retrieved from the live web

Requirement 5: Avoid Content Served From the Cache

Requirement 5: Avoid content served from the cache

We should avoid considering content returned from a cache as this does not reflect the current content of the archive.

Example Requirement 5: issue

- Wayback machines https response header indicates cacheing
  - "X-Page-Cache" indicates if the content is cached
    - X-Page-Cache: Hit
    - X-Page-Cache: Miss
  - Cache hit can produce same hash value even though the actual archive might be different
  - Below shows when not using id_ tag the hash value is the same when used frequently
  - After a period hash value is different

```
% date
Mon Oct  2 01:15:18 EDT 2017
% curl --silent http://web.archive.org/web/20130724144801/htt
p://www.cnn.com/ | md5
477b6d923cbb7bf9675a0d2feb37afd3

% date
Mon Oct  2 01:16:29 EDT 2017
% curl --silent http://web.archive.org/web/20130724144801/htt
p://www.cnn.com/ | md5
477b6d923cbb7bf9675a0d2feb37afd3

% date
Mon Oct  2 01:19:31 EDT 2017
% curl --silent http://web.archive.org/web/20130724144801/htt
p://www.cnn.com/ | md5
477b6d923cbb7bf9675a0d2feb37afd3

% date
Mon Oct  2 02:10:24 EDT 2017
% curl --silent http://web.archive.org/web/20130724144801/htt
p://www.cnn.com/ | md5
dda6a9bf091d412cbdc2226ce3eb1059
```

```
(base) pfoytik@pfoytik-Latitude-5490:~$ curl --silent http://web
.archive.org/web/20130724144801/http://www.cnn.com/ | md5sum
6d3f703c9009bf54c9ff4abb62ad1e17  -
(base) pfoytik@pfoytik-Latitude-5490:~$ curl --silent http://web
.archive.org/web/20130724144801/http://www.cnn.com/ | md5sum
6d3f703c9009bf54c9ff4abb62ad1e17  -
(base) pfoytik@pfoytik-Latitude-5490:~$ curl --silent http://web
.archive.org/web/20130724144801/http://www.cnn.com/ | md5sum
6d3f703c9009bf54c9ff4abb62ad1e17  -
(base) pfoytik@pfoytik-Latitude-5490:~$ curl --silent http://web
.archive.org/web/20130724144801/http://www.cnn.com/ | md5sum
e648a736a1ba16c11ee8b6e0a0266d68  -
```

Requirement 6: Be Aware of the Effect of Changing TimeMaps

Requirement 6: Be aware of the effect of changing TimeMaps

Changing TimeMaps could affect the computation of hashes. It might be necessary to estimate when a memento becomes stable within the archive to avoid issues of having different hashes.

Example Requirement 6: details

- Archives attempt to match archived elements to the closest time
  - Date and time is encoded in the archive links
  - The archived content for the associated elements can change

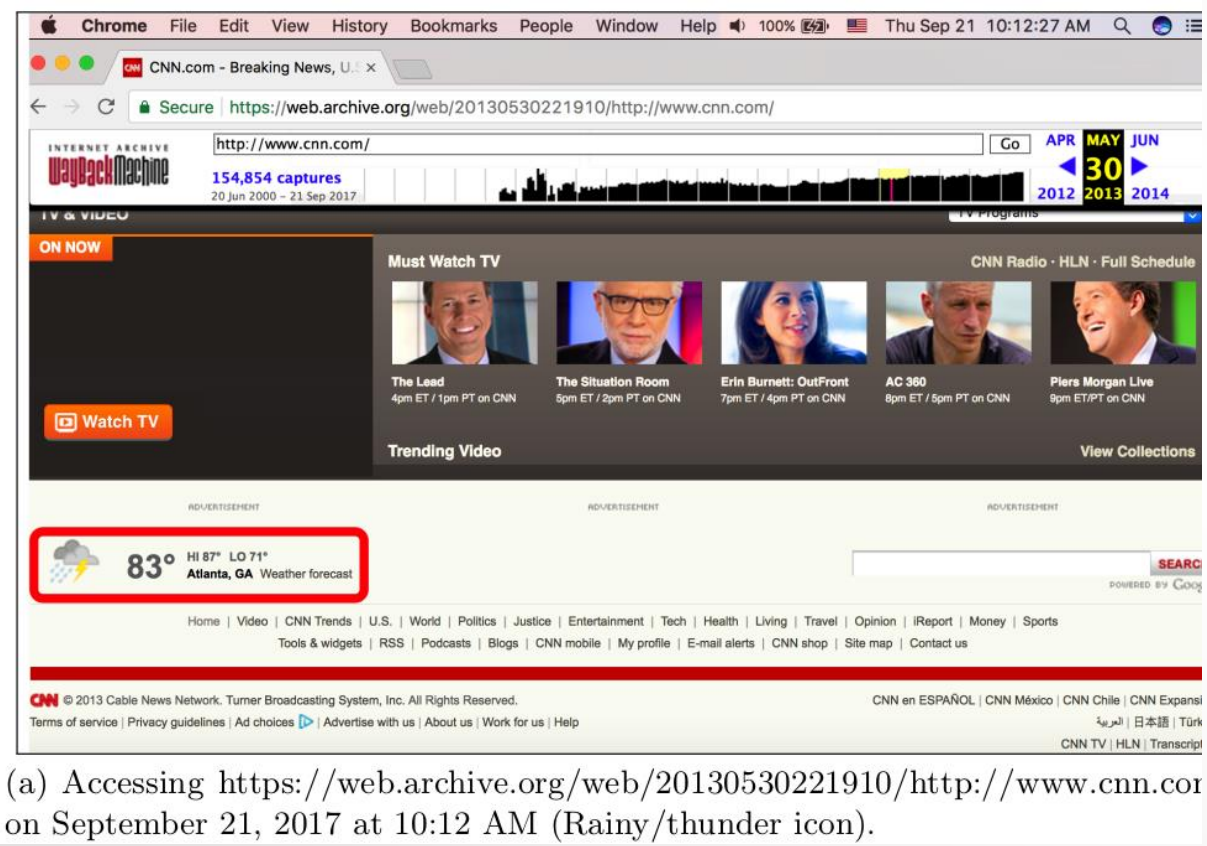
```
http://web.archive.org/web/timemap/link/http://ichef.bbci.co.uk/
wwhp/144/cpsprodpb/730D/production/_97235492_p05brd0w.jpg

https://web.archive.org/web/20170807231028im_/http://ichef.bbci.
co.uk/wwhp/144/cpsprodpb/730D/production/_97235492_p05brd0w.jpg

https://web.archive.org/web/20170807230527im_/http://ichef.bbci.
co.uk/wwhp/144/cpsprodpb/730D/production/_97235492_p05brd0w.jpg
```

Requirement 7: Avoid Using Dynamic Content in Hash Calculations

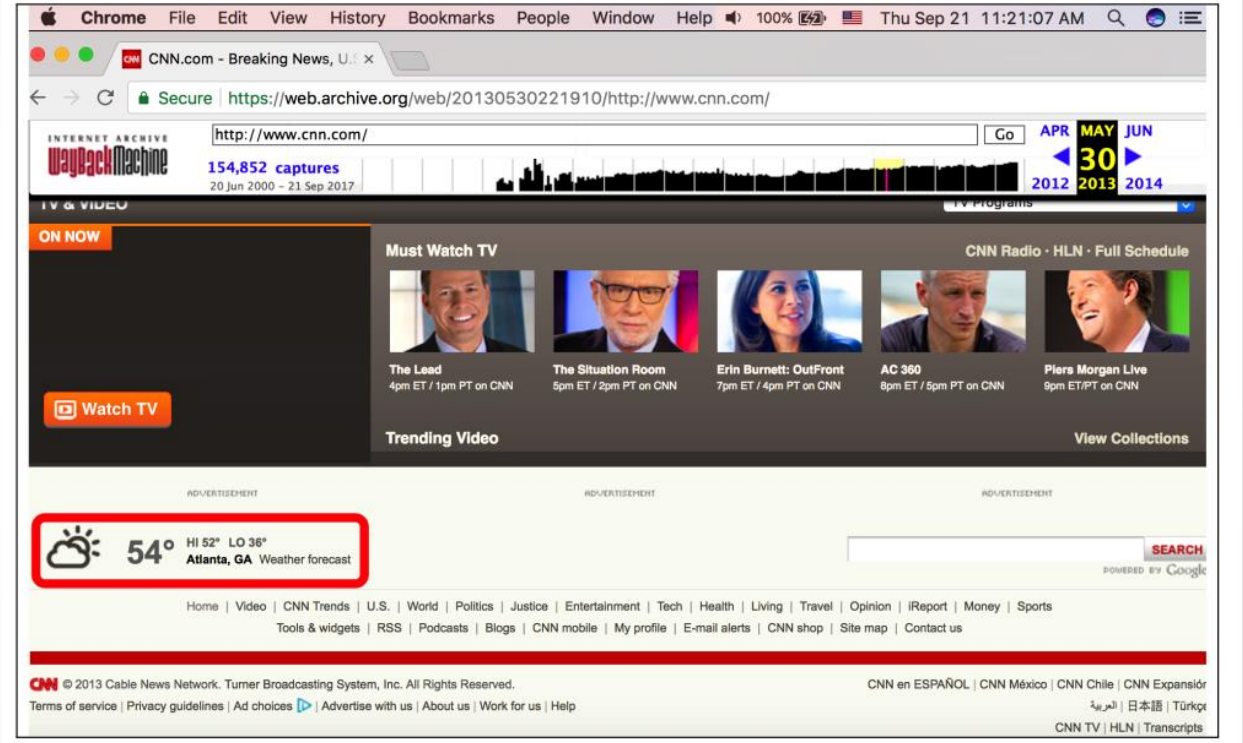
Requirement 7: Avoid using dynamic content in hash calculations

Any resources discovered to have randomly generated values should not be a part of the computation of hashes.

Example Requirement 7: Archive-Escape



(a) Accessing https://web.archive.org/web/20130530221910/http://www.cnn.com on September 21, 2017 at 10:12 AM (Rainy/thunder icon).



(b) Reloading the same memento at a different time may produce a different icon (Cloudy icon).

Requirement 8: Include HTTP Response Headers

Requirement 8: Include HTTP Response headers

Important HTTP Response headers should be included in the hash computation.

Example Requirement 8: discussion

- The Headers are important to verify the integrity of the archive itself
- Original state of the memento
- Not so much the webpage that is being archived
- This is important because it can help identify changes to the memento
- For situations where content type has changed
    1. where images have been converted to different formats
    2. "X-Page-Cache" has changed
    3. HTTP Response header "Location"

Conclusions

- Importance of timestamping archived web pages
- State of the art timestamping services in blockchain-based networks do not allow users to submit URIs of web pages in order to establish trusted timestamps.
- Difficulties exist in timestamping archived web pages (I.e. mementos)
- Archived pages will not be directly timestamped in the blockchain, instead a hash value calculated on the content of the memento is the data to be timestamped

Future/Ongoing Work

- Work is ongoing utilizing the requirements to generate hash records of Mementos
  - Archive assisted archival fixity verification framework
  - Fixity: The process of record keeping (hash files) and checking integrity of source making sure it has not been altered or corrupted
- Utilizing the memento manifest, hash representations of the original archive are recorded and stored
- The Fixity records are stored separately in multiple online archives

Aturban, Mohamed, et al. "Archive assisted archival fixity verification framework." *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. IEEE, 2019.

Finally…

- Blockchains do not strive to prove authenticity, only integrity
  - Garbage in…. Garbage out….
    - All that integrity will give you is certainty that the garbage out is the the same as the garbage in…
  - This is still very valuable but needs to be used in the right way
  - It will not solve all problems
- It will help independently verify that (non-changing) contents have not been altered since being archived
- In web archives there are many pieces
  - Each of these pieces need to be checked for integrity
  - As these pieces are combined the pieces can be checked individually as or combined