Readying Web Archives to Consume and Leverage Web Bundles

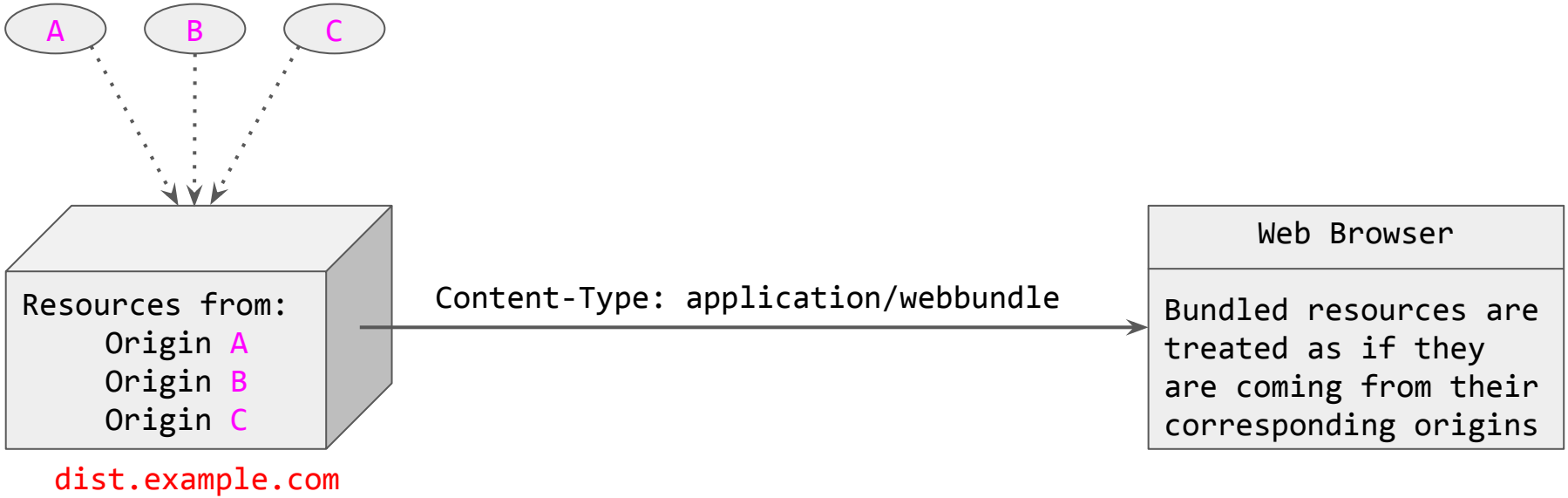
Sawood Alam[1], Michele C. Weigle[2], Michael L. Nelson[2], Martin Klein[3], and Herbert Van de Sompel[4]

[1]Internet Archive, San Francisco, California, USA
[2]Old Dominion University, Norfolk, Virginia, USA
[3]Los Alamos National Laboratory, New Mexico, USA
[4]Data Archiving and Networked Services, Netherlands

@ibnesayeed

IIPC Web Archiving Conference, June 15, 2021

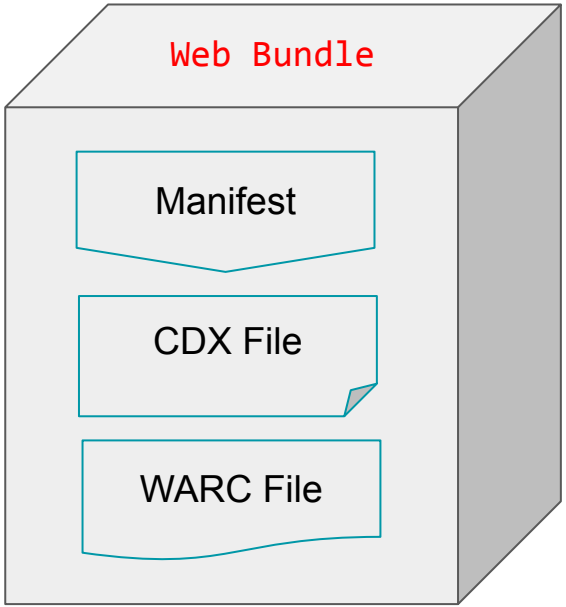
Web Bundles



Bundled HTTP Exchanges may or may not be signed by their corresponding origins

https://github.com/WICG/webpackage#explainers

Web Bundles in Terms of WARC and CDX



- Uses CBOR format for HTTP Response Message storage
- Includes an entrypoint URI
- Lossy (e.g., HTTP headers with the same name are concatenated)
- No provision for provenance information

Web Bundles are optimized for performance, not for the long-term preservation with provenance

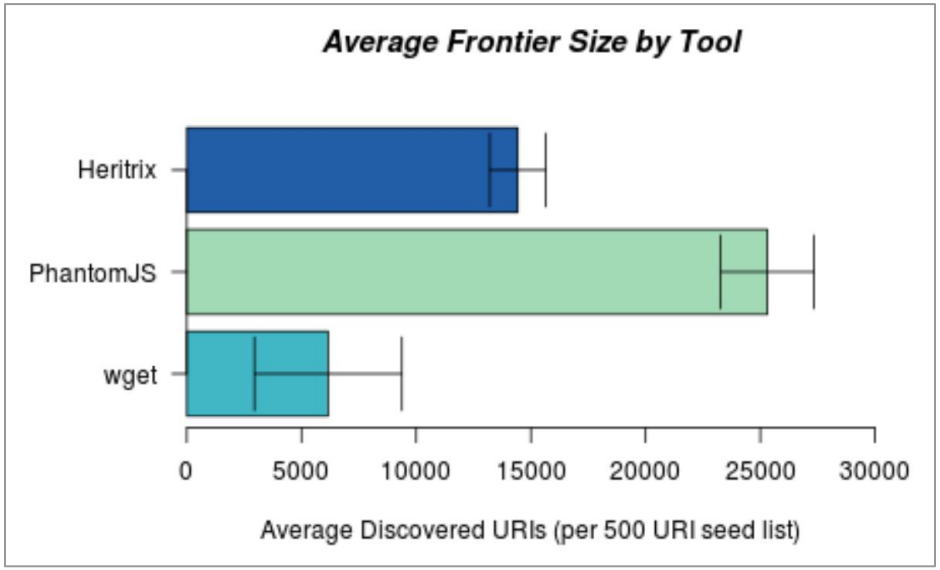
https://www.slideshare.net/ibnesayeed/web-archive-warc-file-format

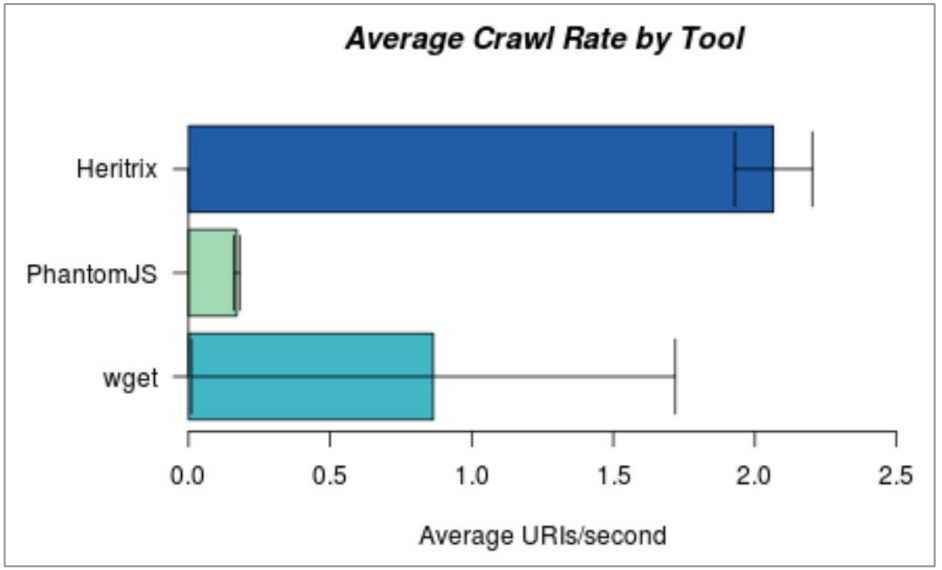
Web Packaging Use Cases

- Essential
  - Offline installation
  - Offline browsing
  - Save and share a web page
  - Privacy-preserving prefetch
- Nice-to-have
  - Packaged Web Publications
  - Avoiding Censorship
  - Third-party security review
  - Building packages from multiple libraries
  - Cross-CDN Serving
  - Pre-installed applications
  - Protecting Users from a Compromised Frontend
  - Installation from a self-extracting executable
  - Packages in version control
  - Subresource bundling
  - **Archival**

https://datatracker.ietf.org/doc/html/draft-yasskin-wpack-use-cases-01.txt#section-2.2.11

Crawling JavaScript-driven Deferred Representations





PhantomJS discovers 75% more resources than Heritrix, but crawls 12 times slower

Web Bundles have the potential to enable efficient and coherent crawling

https://arxiv.org/abs/1508.02315

Request, Decompose, and Index Bundled Resources

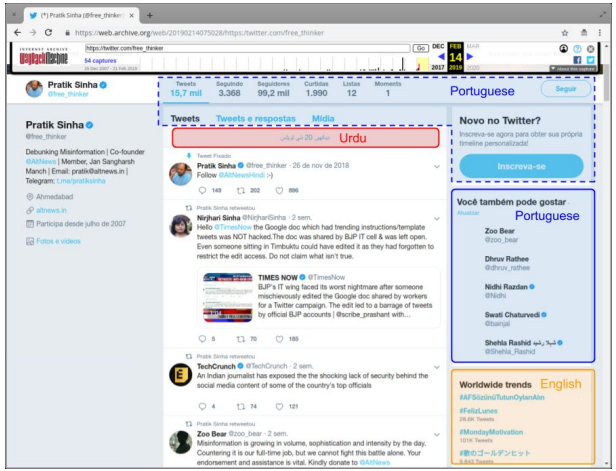
- Crawlers can content negotiate to prefer Web Bundles, when present
  - Server has likely bundled resources that would be difficult to discover
  - A single transaction downloads multiple resources (efficient politeness)
- Media type "`application/webbundle`" needs to be decomposed for preservation in WARC records and indexing

Servers will likely continue to serve non-bundled resources for years to come

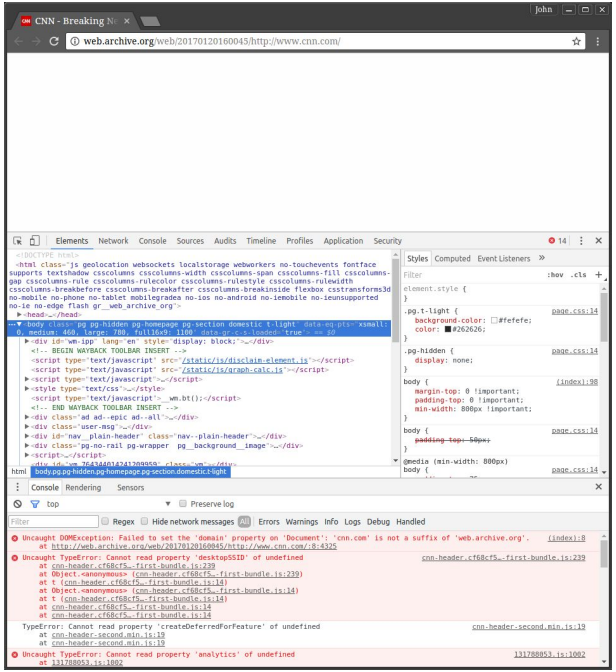
https://developer.mozilla.org/en-US/docs/Web/API/Service_Worker_API

Mementos That Never Existed on the Live Web

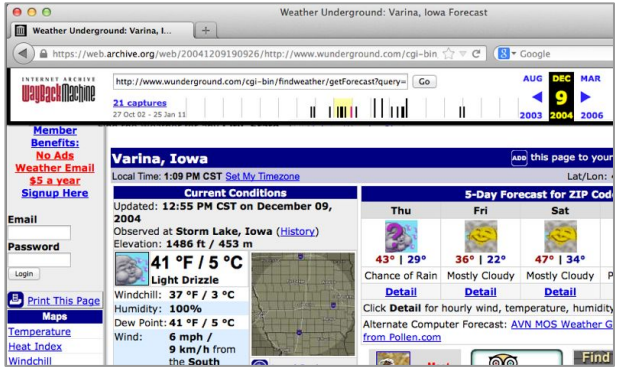
Cookie Violations



Origin Violations



Temporal Violations

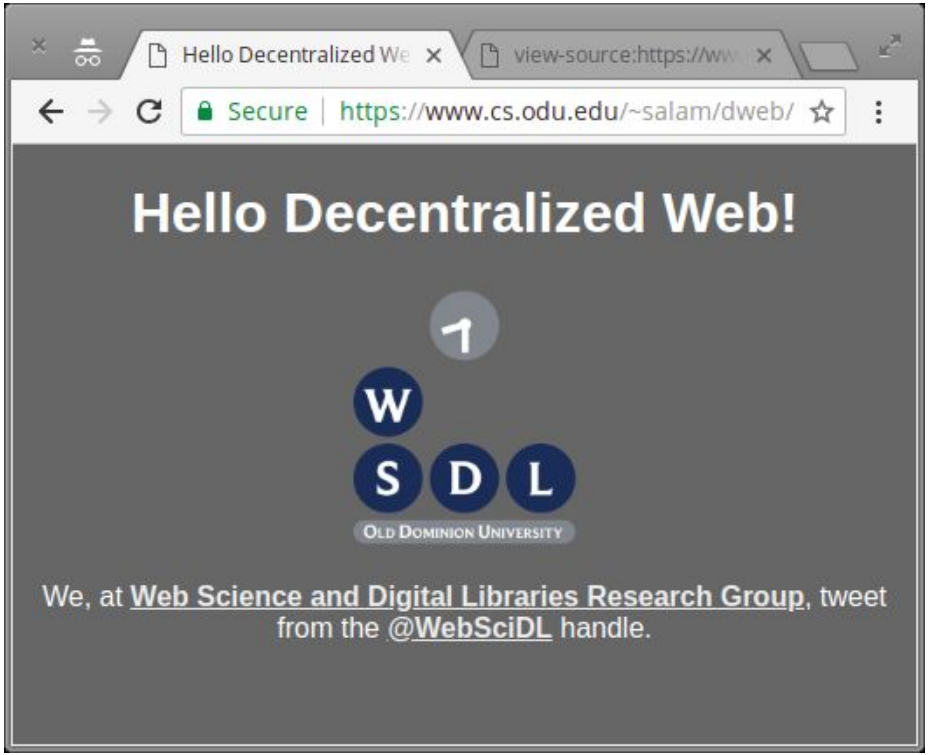


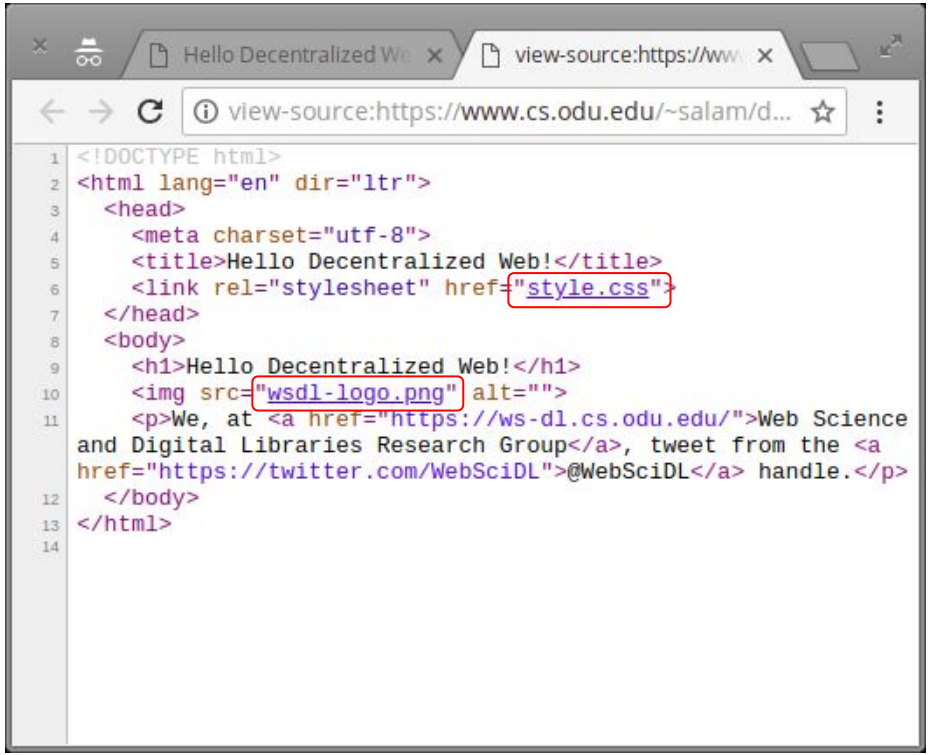
Live leakage (Zombies)

Web Packaging has the potential to eliminate these issues

https://ws-dl.blogspot.com/2012/10/2012-10-10-zombies-in-archives.html
https://ws-dl.blogspot.com/2015/12/2015-12-08-evaluating-temporal.html
https://ws-dl.blogspot.com/2017/01/2017-01-20-cnncom-has-been-unarchivable.html
https://ws-dl.blogspot.com/2019/03/2019-03-18-cookie-violations-cause.html

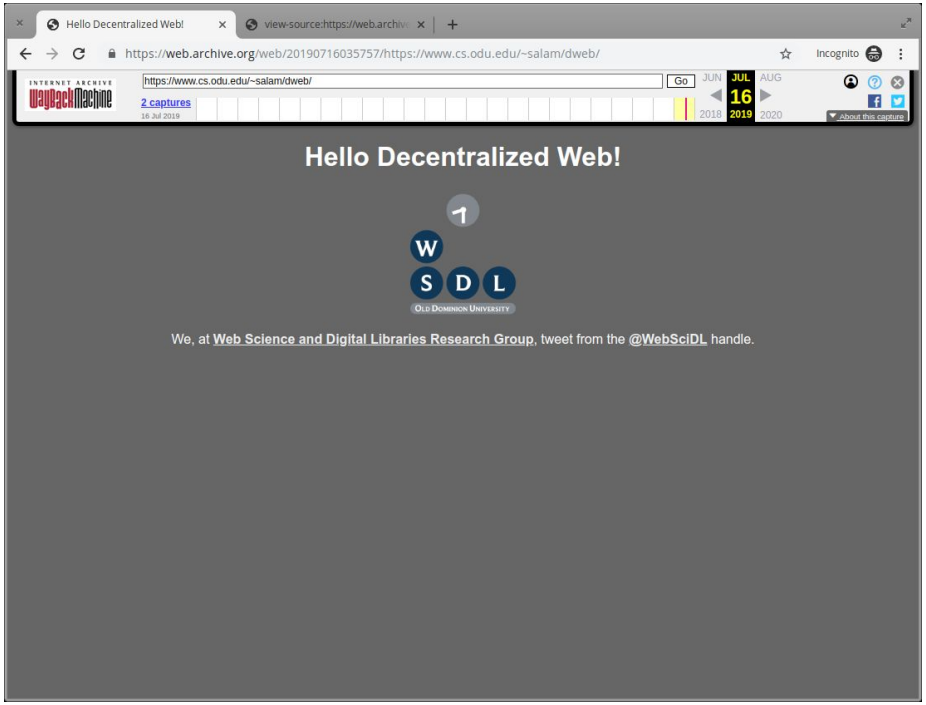
An HTML Page With External Style Sheet and Image

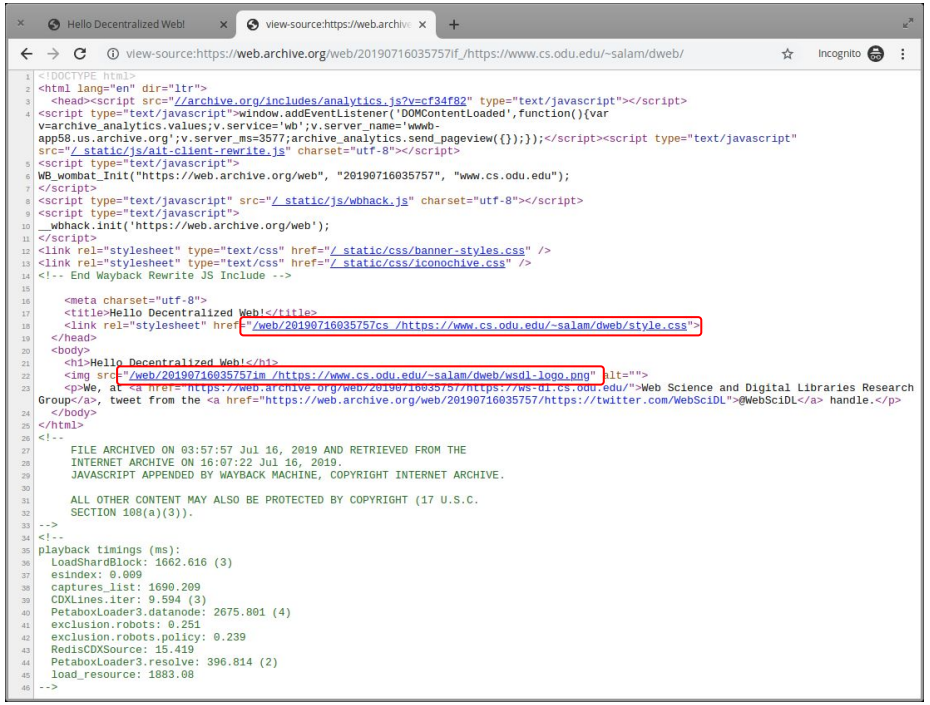




https://www.cs.odu.edu/~salam/dweb/

Archival Replay: Server-side Rewriting

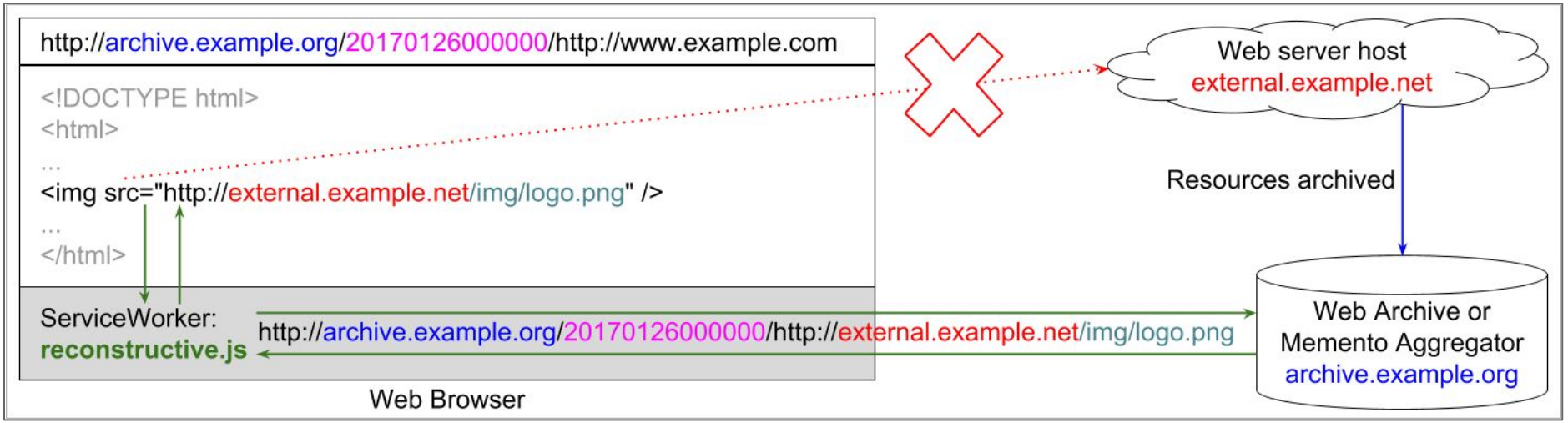




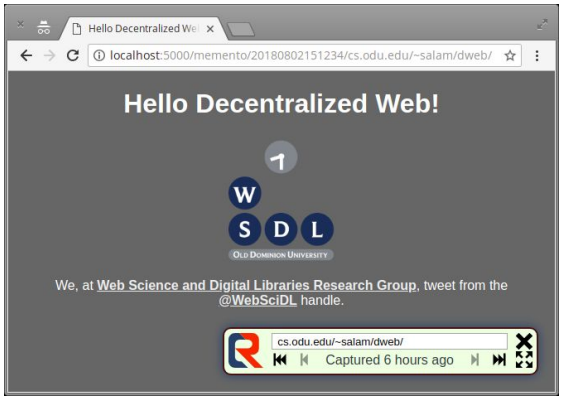
URI references (href, src, and srcset etc.) are rewritten to point to their archived versions at a nearby time

https://web.archive.org/web/20190716035757/https://www.cs.odu.edu/~salam/dweb/

Archival Replay: Client-side Rerouting

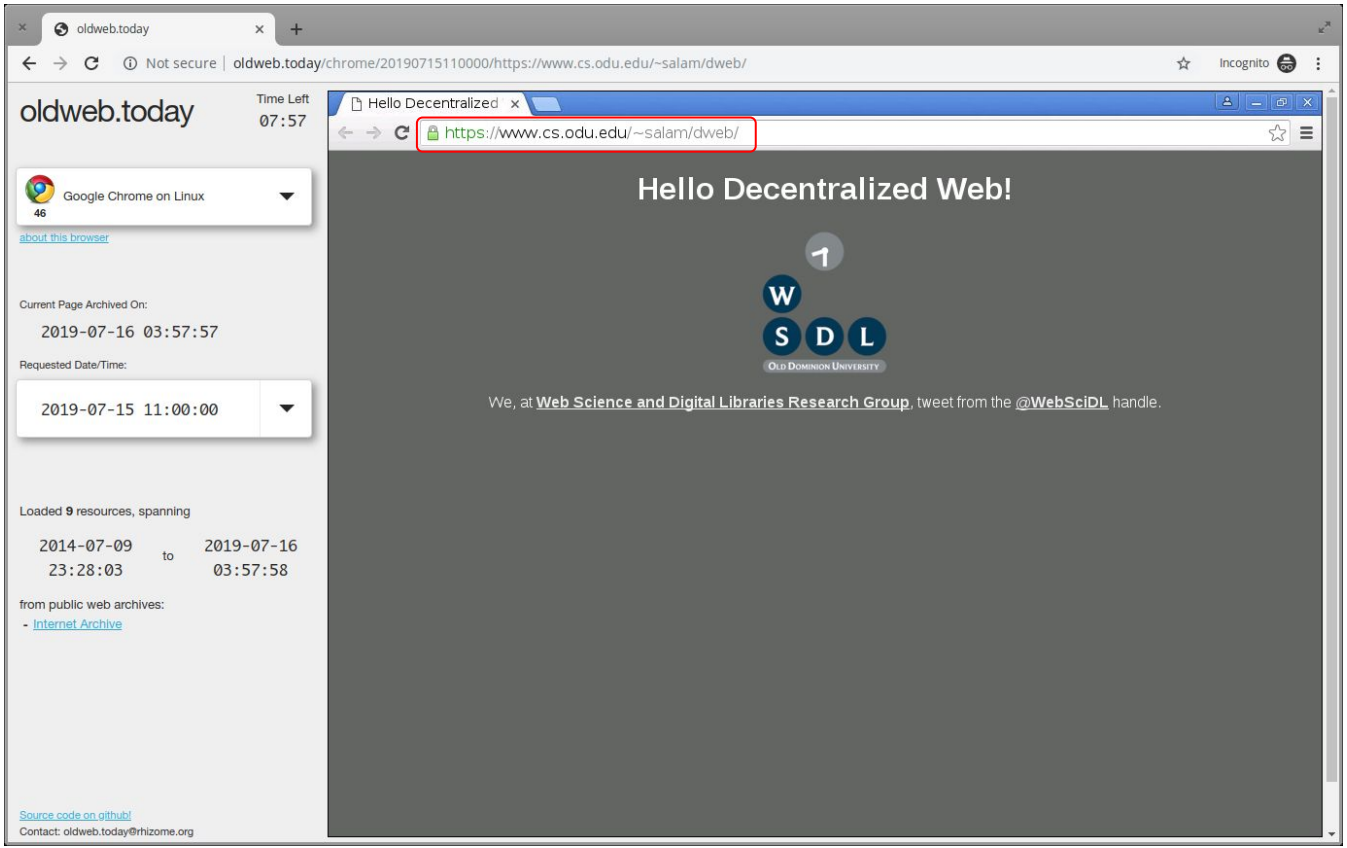


- Avoids zombies (live-leakage)
- Adds an unobtrusive archival banner (using Custom HTML Element)



https://oduwsdl.github.io/Reconstructive/
http://ws-dl.blogspot.com/2018/01/2018-01-08-introducing-reconstructive.html

Archival Replay: Proxy-based Rerouting

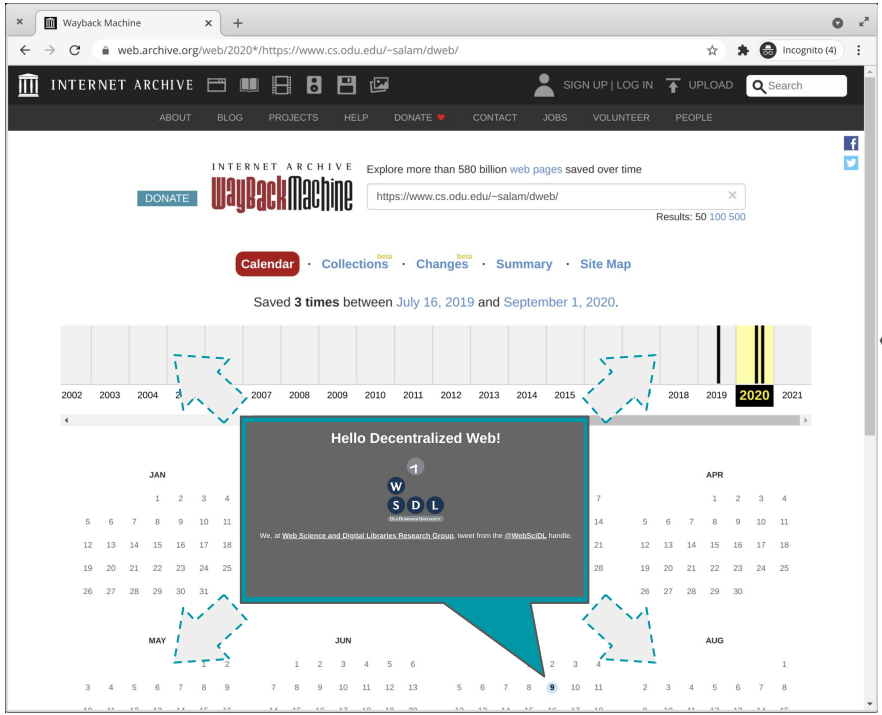


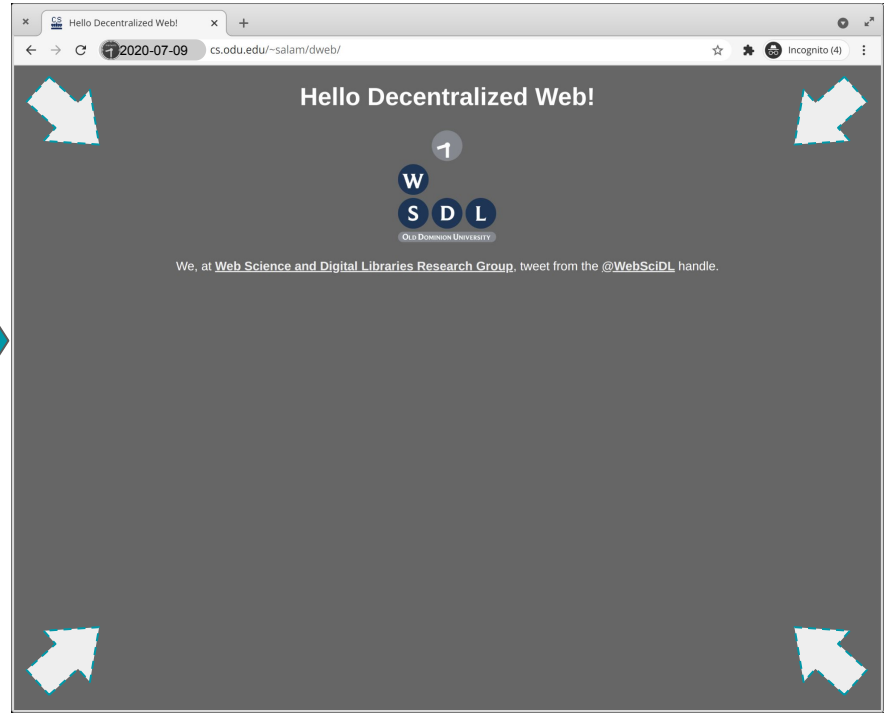
- A web browser runs on a remote host
- Configured to use a web archive proxy
- Accessed via VNC
- Does not scale well

Web Bundles can enable archival replay from original URIs locally without a replay proxy

http://oldweb.today/chrome/20190715110000/https://www.cs.odu.edu/~salam/dweb/

Bundled Archival Replay in Portals





Portal is an experimental HTML element, similar to iframe, but it allows seamless transition to become the main context when activated

https://web.dev/hands-on-portals/

Memento TimeGate

```
$ curl -I https://www.w3.org/wiki/Main_Page
HTTP/2 200
date: Tue, 16 Jul 2019 03:16:01 GMT
link: <https://www.w3.org/wiki/Main_Page>; rel="original latest-version",
      <https://www.w3.org/wiki/Special:TimeGate/Main_Page>; rel="timegate",
      <https://www.w3.org/wiki/Special:TimeMap/Main_Page>; rel="timemap"; type="application/link-format"; from="Thu, 01 Jan 1970 00:00:00 GMT"; until="Fri, 16 Nov 2018 19:10:23 GMT",
      <https://www.w3.org/wiki/index.php?title=Main_Page&oldid=30366>; rel="first memento"; datetime="Thu, 01 Jan 1970 00:00:00 GMT",
      <https://www.w3.org/wiki/index.php?title=Main_Page&oldid=108148>; rel="last memento"; datetime="Fri, 16 Nov 2018 19:10:23 GMT"
content-language: en
vary: Accept-Encoding,Cookie
cache-control: s-maxage=18000, must-revalidate, max-age=0
last-modified: Mon, 15 Jul 2019 22:16:01 GMT
content-type: text/html; charset=UTF-8

$ curl -I -H "Accept-Datetime: Sat, 20 Dec 2014 12:30:00 GMT" https://www.w3.org/wiki/Special:TimeGate/Main_Page
HTTP/2 302
date: Tue, 16 Jul 2019 03:16:21 GMT
vary: Accept-Encoding,Cookie,Accept-Datetime
location: https://www.w3.org/wiki/index.php?title=Main_Page&oldid=80125
link: <https://www.w3.org/wiki/Special:TimeMap/Main_Page>; rel="timemap"; type="application/link-format"; from="Thu, 01 Jan 1970 00:00:00 GMT"; until="Fri, 16 Nov 2018 19:10:23 GMT",
      <https://www.w3.org/wiki/index.php?title=Main_Page&oldid=30366>; rel="first memento"; datetime="Thu, 01 Jan 1970 00:00:00 GMT",
      <https://www.w3.org/wiki/index.php?title=Main_Page&oldid=108148>; rel="last memento"; datetime="Fri, 16 Nov 2018 19:10:23 GMT",
      <https://www.w3.org/wiki/Main_Page>; rel="original latest-version"
content-type: text/html; charset=UTF-8

$ curl -I "https://www.w3.org/wiki/index.php?title=Main_Page&oldid=80125"
HTTP/2 200
date: Tue, 16 Jul 2019 03:38:35 GMT
x-content-type-options: nosniff
memento-datetime: Sat, 20 Dec 2014 11:34:08 GMT
link: <https://www.w3.org/wiki/Main_Page>; rel="original latest-version",<https://www.w3.org/wiki/Special:TimeGate/Main_Page>;
rel="timegate",<https://www.w3.org/wiki/Special:TimeMap/Main_Page>; rel="timemap"; type="application/link-format"; from="Thu, 01 Jan 1970 00:00:00 GMT"; until="Fri, 16 Nov 2018
19:10:23 GMT",<https://www.w3.org/wiki/index.php?title=Main_Page&oldid=30366>; rel="first memento"; datetime="Thu, 01 Jan 1970 00:00:00
GMT",<https://www.w3.org/wiki/index.php?title=Main_Page&oldid=108148>; rel="last memento"; datetime="Fri, 16 Nov 2018 19:10:23 GMT"
content-language: en
vary: Accept-Encoding,Cookie
expires: Thu, 01 Jan 1970 00:00:00 GMT
cache-control: private, must-revalidate, max-age=0
content-type: text/html; charset=UTF-8
```

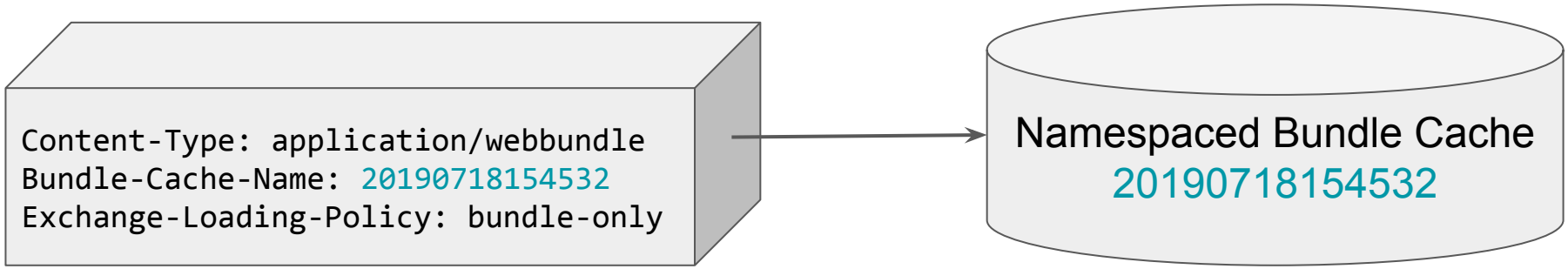
Currently, Web Bundles prefer the most recent version of a resource

Web archives provide third-party generic TimeGate resources

Native TimeGate support in Exchange Loading would be great

https://tools.ietf.org/html/rfc7089

Namespaced Cache for Temporal Coherence



- Distributor provides a custom cache namespace with Bundles to stash them in
- More than one related Bundles can have the same namespace
- A configurable policy determines the behavior of cache utilization for Loading
- Archive origins configure Loading policy to only load resources from bundles they delivered to prevent loading zombie resources

Enable sandboxing via namespaced caches for security and coherence

https://developer.mozilla.org/en-US/docs/Web/API/Cache

Catch-All Route to Prevent Zombies

- If a needed resources is not in a bundle or any caches of the browser, it causes live-leakage
- This allows publishers to create partial Web Bundles for distribution while serving resources like analytics, trackers, ads, etc. live
- A fallback catch-all route would be helpful in preventing Zombies in archives, but give too much power to distributors

Introduce wildcard routes or Service Workers for namespaced caches to handle missing resources

https://developer.mozilla.org/en-US/docs/Web/API/Service_Worker_API

Web Archives Currently Lack Technical Means to Prove Fixity and Non-repudiation

- Joy Ann Reid claimed copies of her blog in the Internet Archive has been hacked
- The Internet Archive publicly denied it
- We investigated the matter using multiple web archives and concluded Reid's claim to be very unlikely

There is strong need of verifiable web archiving

https://twitter.com/Jamie_Maz/status/936349041264414721
http://blog.archive.org/2018/04/24/addressing-recent-claims-of-manipulated-blog-posts-in-the-wayback-machine/
https://ws-dl.blogspot.com/2018/04/2018-04-24-why-we-need-multiple-web.html
https://arxiv.org/abs/1905.12565

Temporal Validation of Signed Exchanges

- Signed exchanges contain "`Date`" response header
- An archive returns "`Memento-Datetime`" header when delivering an HTTP Bundle
- Validate signature in the context of that historical time
  - The difference in the two times should be within an acceptable margin
- Utilize means like Certificate Transparency Logs to establish temporal validation

A means to establish that the signature would have been "`temporally valid`" at a given time in the past

http://www.certificate-transparency.org/

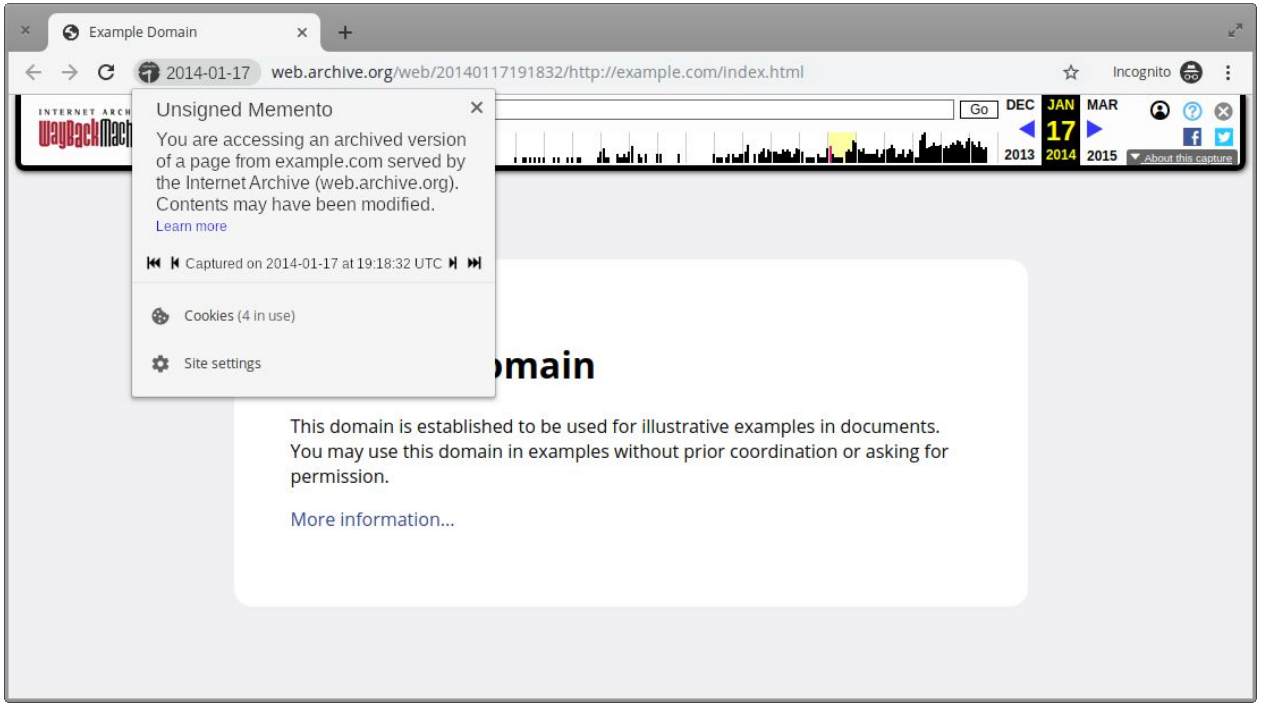
Archival Replay With Unsigned Exchanges

- The origin will likely change (specs are not stable yet)
- Loses the benefits of replaying archived pages as the original domain
- Benefits from faster replay and better caching of popular mementos

Archives can still create Web Bundles with rewritten URIs to replay in the traditional fashion

http://www.certificate-transparency.org/

Native Memento Support and Acknowledgement

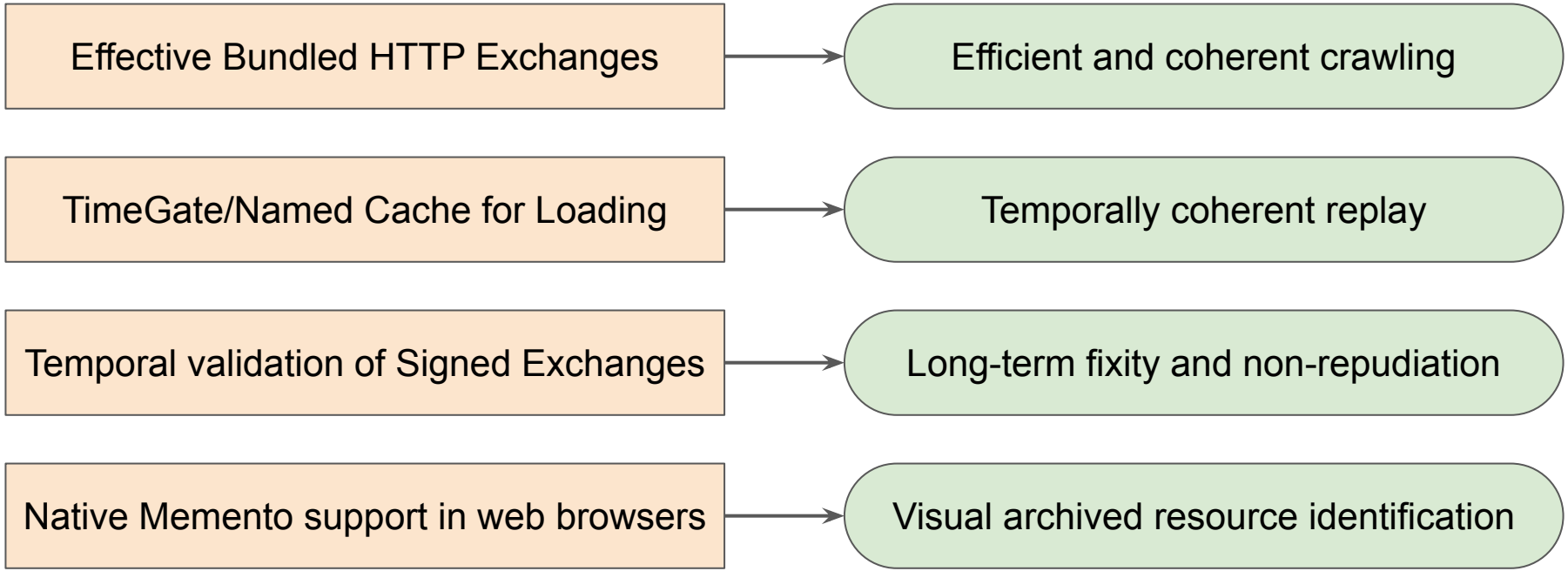


- Visually acknowledge presence of "`Memento-Datetime`" header
- Surface archival metadata from "`Link`" header
- Enable necessary security features for the archived web which is read-only
- Prevent live-leakage

https://docs.google.com/presentation/d/1YAQI_1sPH25ZdAiEPhE5cqj8zMomLBKDQWzPS5Avw5s/edit
https://www.slideshare.net/mweigle/enabling-personal-use-of-web-archives

Conclusions

Web Packaging has a unique opportunity to devise a technology that supports web archiving and provides a much needed capability to verify the integrity of archived web resources



https://arxiv.org/abs/1906.07104