# Readying Web Archives to Consume and Leverage Web Bundles

**Sawood Alam<sup>1</sup>**, Michele C. Weigle<sup>2</sup>, Michael L. Nelson<sup>2</sup>, Martin Klein<sup>3</sup>, and Herbert Van de Sompel<sup>4</sup>

A R C H I V E

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

@ibnesayeed

DATA Archiving and Networked Services

IIPC Web Archiving Conference, June 15, 2021

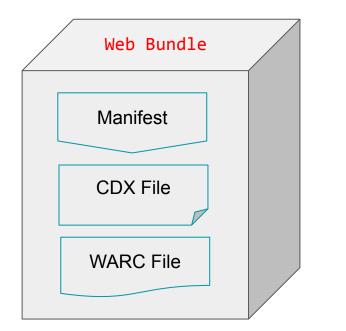
#### Web Bundles



dist.example.com

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

### 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

### Web Packaging Use Cases

#### Essential

- Offline installation
- Offline browsing
- Save and share a web page
- Privacy-preserving prefetch

#### • Nice-to-have

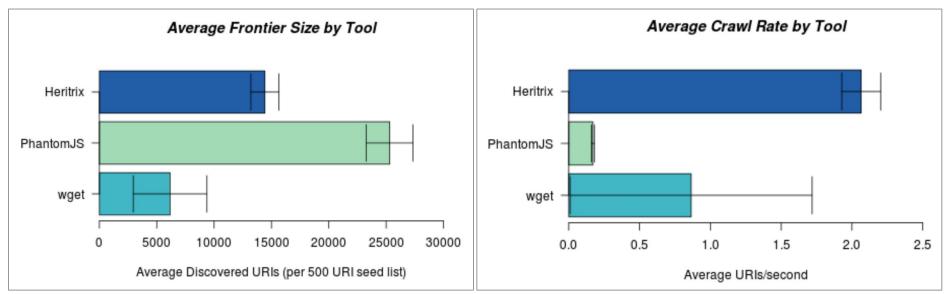
- Packaged Web Publications
- Avoiding Censorship
- Third-party security review
- Building packages from multiple libraries
- Cross-CDN Serving
- Pre-installed applications
- Protecting Users from a Compromised Frontend
- Installation from a self-extracting executable
- Packages in version control
- Subresource bundling
- Archival

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

@ibnesayeed

4

## 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

#### 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

**Origin Violations** 

CNN - Breaking No ×

#### Cookie Violations

https://witter.com/tree_think	· · · · · · · · · · · · · · · · · · ·	DEC FEB MAR
Pratik Sinha O	Tweets Seguindo Seguindores Curtidas Listas Moments 15,7 mil 3.368 99,2 mil 1.990 12 1	Portuguese Segur
Pratik Sinha 📀	Tweets Tweets e respostas Midia	dzoo bear
Nee_Interver Hebunking Misinformation   Co-founder NANews   Member, Jan Sangharsh anch   Email: pratikthatnews.in   elegram: Line/pratiksinha ® Ahmedabad		
y Antinadad Ο Johnson II Ε) Partopa dinda juha da 2007 Ξζ Fatos e Videos	C1 Pratk Selfa Intervetou Wind Selfa Selfa Charlow The Google doc which had hending instructions/hemplain Hell of Threads was BOO had been by BLP TC all & was intrope Evens someone stilling in Timukatic ouch have defail as they had longotter restruct the ddi access. Do not claim what and true.	
	BP-IF trays facel its work of threadbar BP-IF trays facel its work nightness after someone machineously edited the doogle doc shared by workers for a "fitter company. The edit do a barring of threats by official BJP accounts   discribe_preshert with	Nidhi Razdan O
	O 5 €2,70 ☉ 185	Shehla Rashid لا شہلا رخب O OShehla, Rashid
	Pask Enhancement     TechCrunch • 2 sem.     An indian journalist has exposed the the shocking lack of security behind the social media content of some of the country's top officials	Worldwide trends English
	Q 4 12 74 (V 121	#FelizLunes
	11 Pratk Ends where ends of the second se	26.5K Tweets #MondayMotivation 101k Tweets #数のゴールデンヒット 9.843 context

Web Packaging has the potential to eliminate these issues

#### C () web.archive.org/web/20170120160045/http://www.cnn.com ÷ 🖟 🗐 Elements Network Console Sources Audits Timeline Profiles Application Security 014 : X Styles Computed Event Listeners -ntml class— js geotocation websockets localstorage webworkers no-touenevents fontrace upports textshadow csscolumns csscolumns-width csscolumns-span csscolumns-fill csscolumns ap csscolumns-rule csscolumns-rulecolor csscolumns-rulestyle csscolumns-rulewidth :hoy .cls + columos-breakbefore cascolumos-breakafter cascolumos-breakinside flexbox castransforms mobile no-phone no-tablet mobilegradea no-ios no-android no-iemobile no-ieunsupp ie no-edge flash gr web archive org" .pg.t-light ( page.css:14 background-color: #fefefe color: ##2626263 <!-- BEGIN WAYBACK TOOLBAR INSERT --> .pg-hidden ( page.css:14 <script type='text/javascript' src='/static/is/disclain-element.is'></script-</pre> display: none: <script type="text/javascript" src="/static/is/graph-calc.is"></script></script> P<script type="text/javascript">\_</script;</pre> body { (index):98 >>=<style type="text/css">=</style> margin-top: 0 !important; script type='text/javascript'> wm.bt();</script: padding-top: 0 !important; ... END MAYRACK TOOL BAR INSERT .... min-width: 800px !important ▶ div class="user-msg">\_</div> body f page.css:14 padding top: 50ox >>div class='pg-no-rail pg-wrapper pg background image'>..</div</pre> Omedia (min\_width: 880nx) page.css:14 Console Rendering Sensors O T top T Preserve loc Regex Hide network messages Frors Warnings Info Logs Debug Handled Uncaught DDMException: Failed to set the 'domain' property on 'Document': 'cnn.com' is not a suffix of 'web.archive.org'. (index):8 at http://web.archive.org/web/20170120160845/http://www.cnn.com/:8:432 caught TypeError: Cannot read property 'desktopSSID' of undefined cnn-header.cf68cf5\_-first-bundle.js:239 TypeError: Cannot read property 'createDeferredForFeature' of undefin cnn-header-second.min.is:19 at <u>cnn-header-second.min.js:19</u> at <u>cnn-header-second.min.js:19</u> ht TypeError: Cannot read property 'analytics' of undefined 131788053.15:100

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

#### **Temporal Violations**

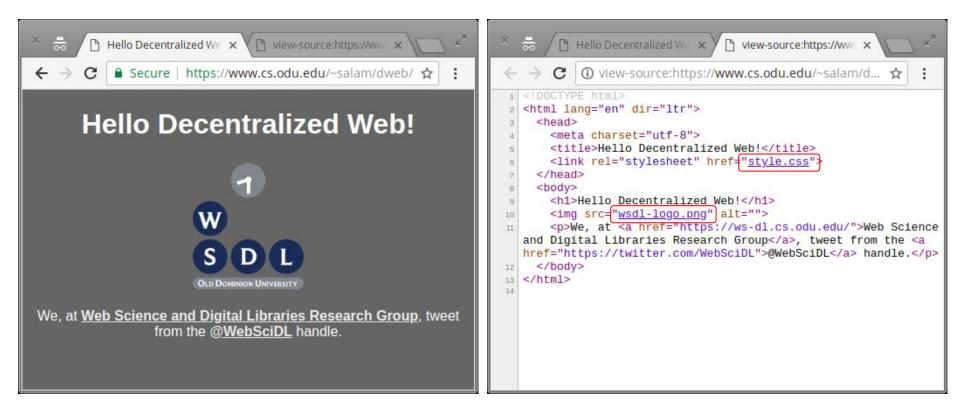


#### Live leakage (Zombies)

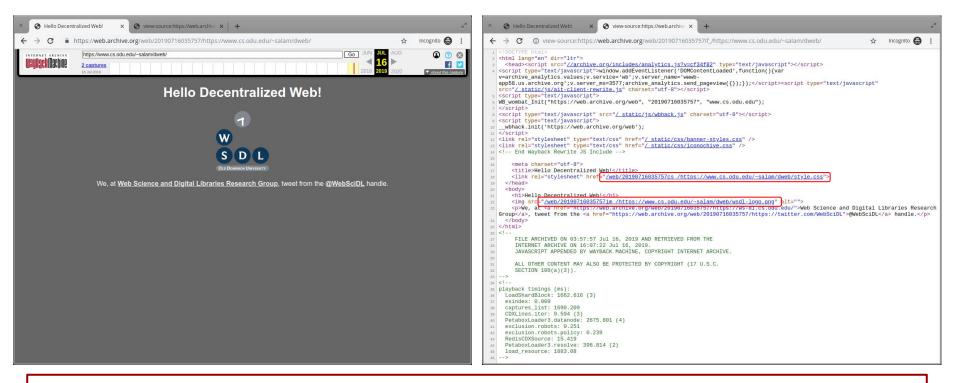


#### @ibnesayeed 7

## An HTML Page With External Style Sheet and Image



### Archival Replay: Server-side Rewriting

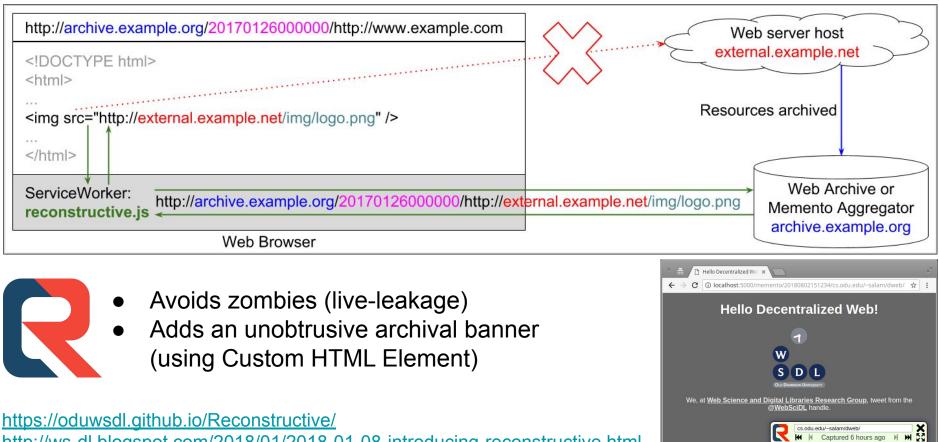


URI references (href, src, and srcset etc.) are rewritten to point to their archived versions at a nearby time

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

@ibnesayeed 9

## Archival Replay: Client-side Rerouting



http://ws-dl.blogspot.com/2018/01/2018-01-08-introducing-reconstructive.html

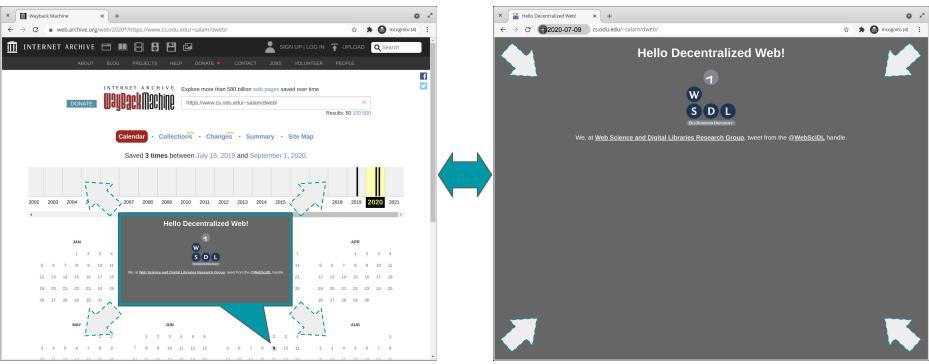
Captured 6 hours ago

## Archival Replay: Proxy-based Rerouting

×     Soldweb.today     ×     +       ←     →     C     ① Not secure   oldweb.today/	chrome/20190715110000/https://www.cs.odu.edu/~salam/dweb/	_× €	• A web browser runs
Oldweb.today       Time Left 07:57         Image: Comparison of Linux       Image: Comparison of Linux         Browt this browser       Image: Comparison of C	<ul> <li>Hello Decentralized ×</li> <li>A le and le bit ps://www.cs.odu.edu/~salam/dweb/</li> </ul>		<ul> <li>on a remote host</li> <li>Configured to use a web archive proxy</li> <li>Accessed via VNC</li> <li>Does not scale wel</li> </ul> Web Bundles can enable archival replay from original URIs locally without a replay proxy
Loaded 9 resources, spanning 2014-07-09 to 2019-07-16 23:28:03 03:57:58 from public web archives: • Internet Archive Source code on othus/ Contact: oldweb.lodgw@rhtzome.org			

@ibnesayeed <sup>11</sup>

### **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

varv: Accept-Encoding.Cookie.Accept-Datetime

location: https://www.w3.org/wiki/index.php?title=Main Page&oldid=80125

#### Currently, Web Bundles prefer the

#### most recent version of a resource

Web archives provide third-party generic **TimeGate resources** 

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

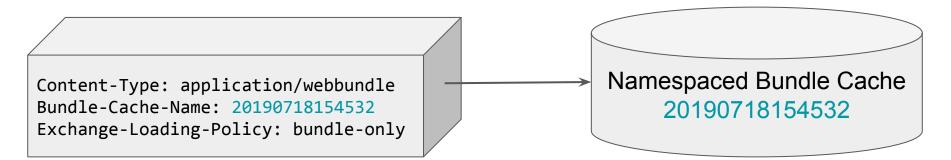
link: <https://www.w3.org/wiki/Main Page>; rel="original latest-version",<https://www.w3.org/wiki/Special:TimeGate/Main Page>; rel="timegate",<https://www.w3.org/wiki/Special:TimeMap/Main Page>; rel="timemap"; type="application/link-format"; from="Thu, 01 Jan 1970 00:00:00 GMT"; until="Fri, 16 Nov 2018 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

varv: 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

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

Native TimeGate support in Exchange Loading would be great

### 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

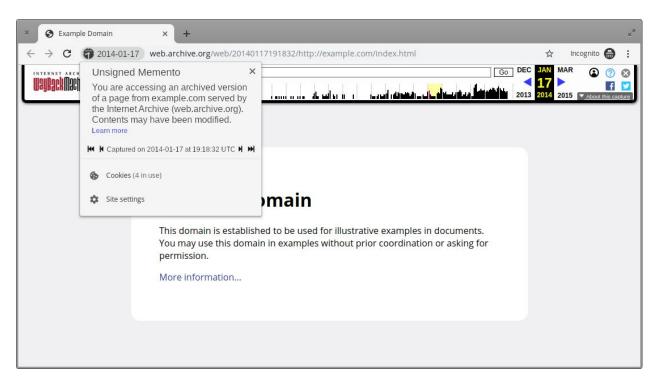
#### A means to establish that the signature would have been "temporally valid" at a given time in the past

### 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

## 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

19

https://docs.google.com/presentation/d/1YAQI\_1sPH25ZdAiEPhE5cqj8zMomLBKDQWzPS5Avw5s/edit https://www.slideshare.net/mweigle/enabling-personal-use-of-web-archives @ibnesayeed

### 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

