ipwb

interplanetary wayback

InterPlanetary Wayback

The Next Step Towards Decentralized Web Archiving

Sawood Alam, Mat Kelly, Michele C. Weigle, Michael L. Nelson

Web Science and Digital Libraries Research Group

Old Dominion University

Norfolk, Virginia, USA

@WebSciDL

`http://github.com/oduwsdl/ipwb`

IPFS Lab Day, Decentralized Web Summit, 2018

San Francisco, CA (USA)

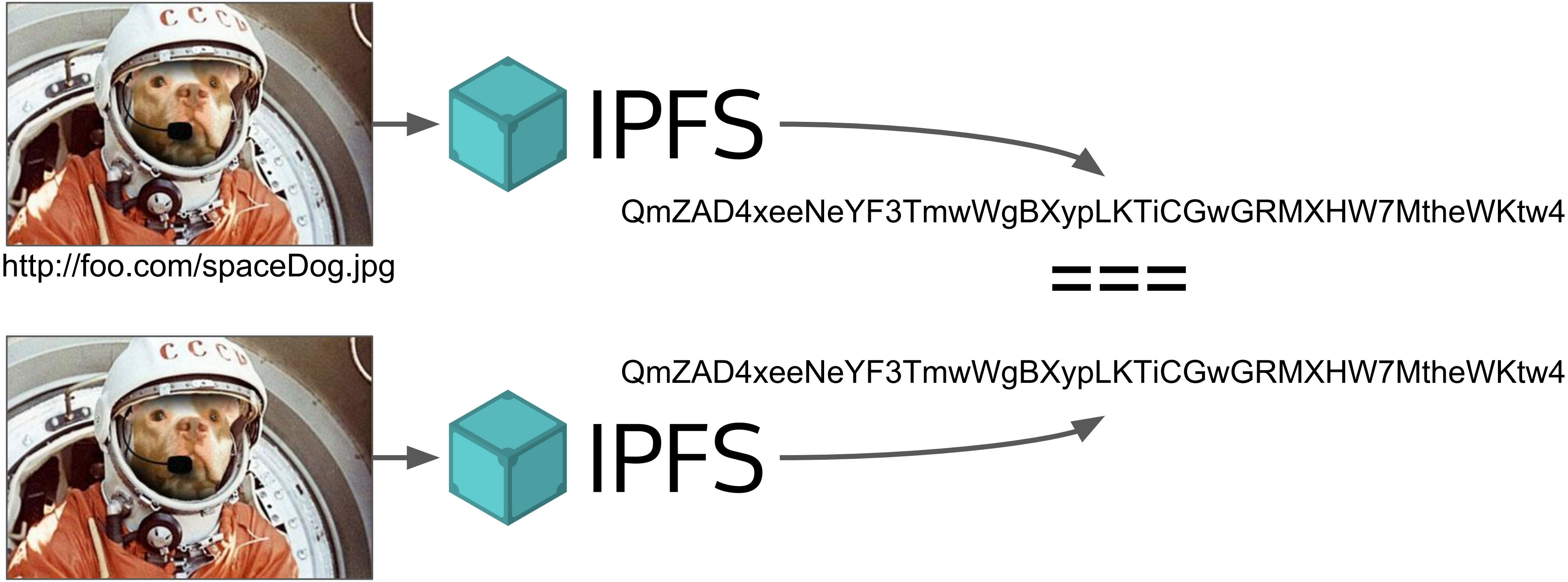
August 3, 2018

W S D L

Old Dominion University

Protocol Labs

Content Addressing



http://foo.com/spaceDog.jpg

IPFS

QmZAD4xeeNeYF3TmwWgBXypLKTiCGwGRMXHW7MtheWKtw4

===

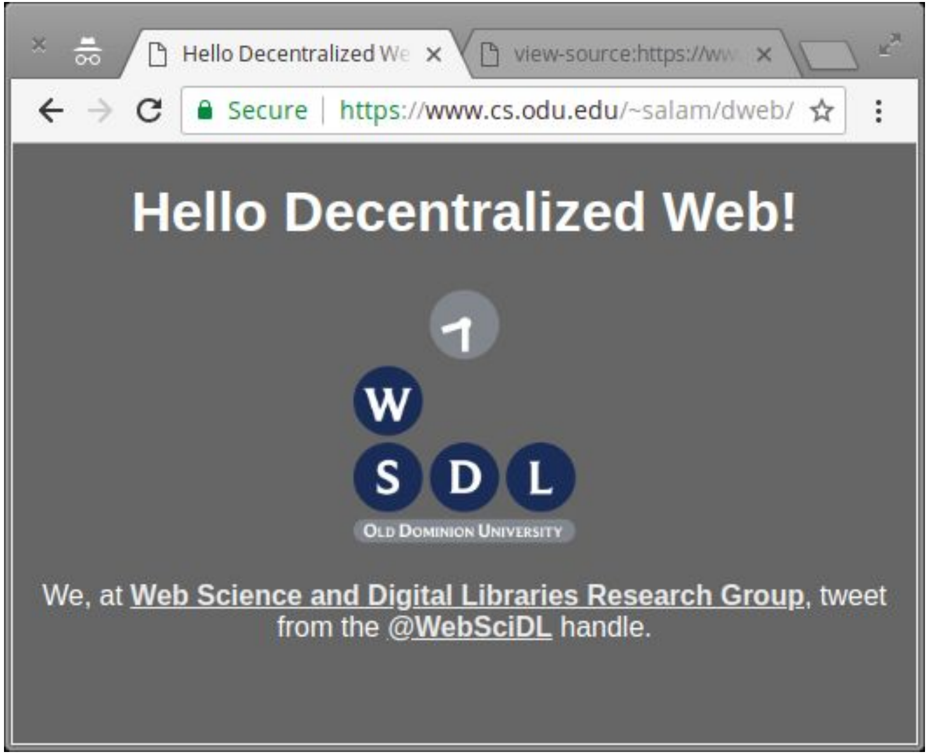
QmZAD4xeeNeYF3TmwWgBXypLKTiCGwGRMXHW7MtheWKtw4

IPFS

http://example.org/yuri.jpg

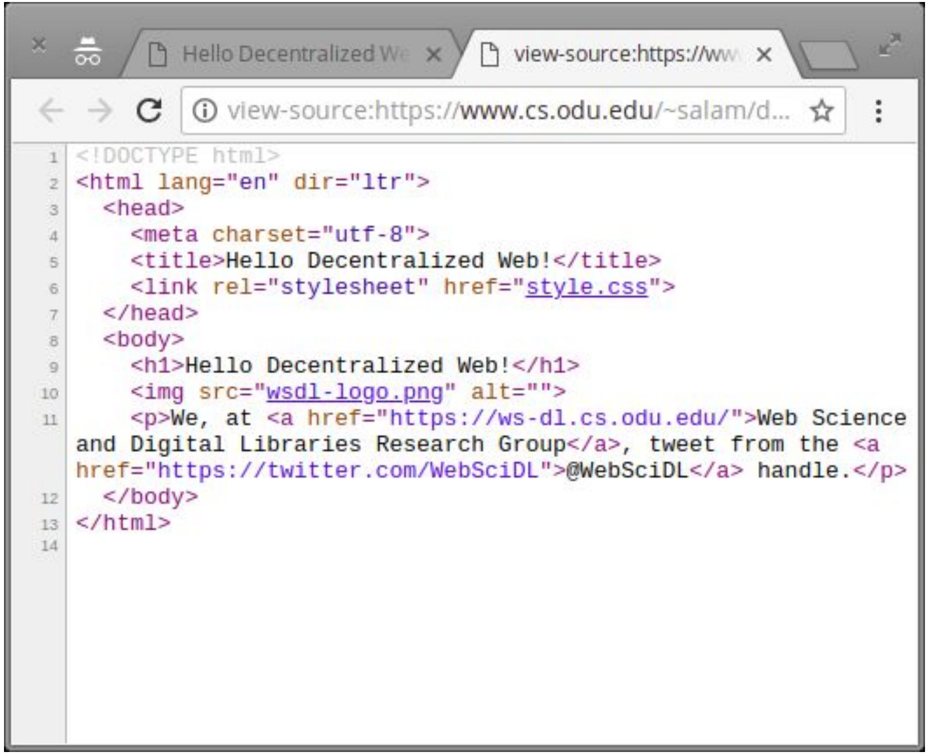
```
$ ipfs cat QmZAD4xeeNeYF3TmwWgBXypLKTiCGwGRMXHW7MtheWKtw4 > doge.jpg
```

Rendered HTML vs. Source Code



Hello Decentralized Web!

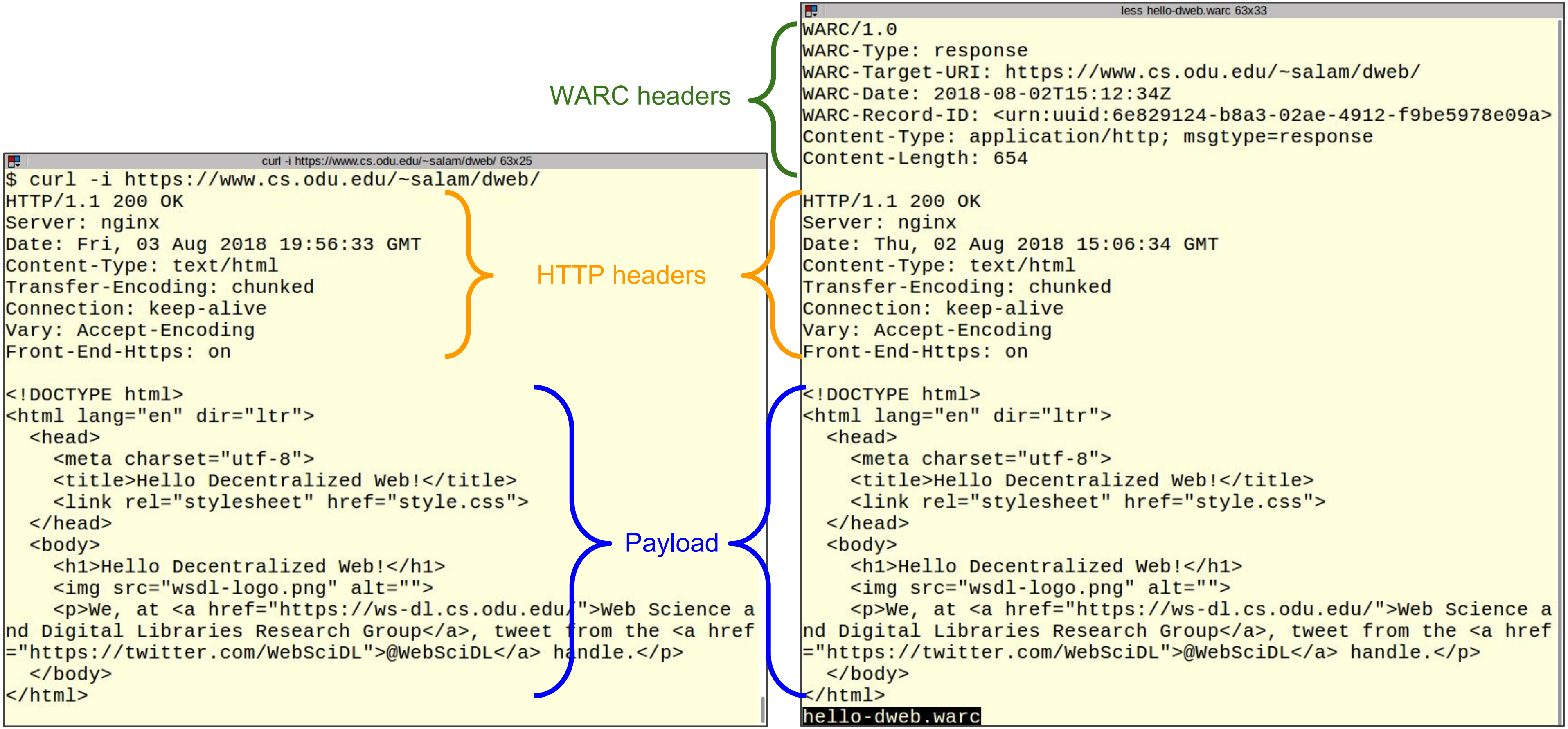
We, at Web Science and Digital Libraries Research Group, tweet from the @WebSciDL handle.



```
<!DOCTYPE html>
<html lang="en" dir="ltr">
  <head>
    <meta charset="utf-8">
    <title>Hello Decentralized Web!</title>
    <link rel="stylesheet" href="style.css">
  </head>
  <body>
    <h1>Hello Decentralized Web!</h1>
    <img src="wsdl-logo.png" alt="">
    <p>We, at <a href="https://ws-dl.cs.odu.edu/">Web Science and Digital Libraries Research Group</a>, tweet from the <a href="https://twitter.com/WebSciDL">@WebSciDL</a> handle.</p>
  </body>
</html>
```

@ibnesayeed

HTTP Response vs. WARC Record


less hello-dweb.warc 63x33
WARC headers
curl -i https://www.cs.odu.edu/~salam/dweb/ 63x25
HTTP headers
Payload

```
$ curl -i https://www.cs.odu.edu/~salam/dweb/
HTTP/1.1 200 OK
Server: nginx
Date: Fri, 03 Aug 2018 19:56:33 GMT
Content-Type: text/html
Transfer-Encoding: chunked
Connection: keep-alive
Vary: Accept-Encoding
Front-End-Https: on

<!DOCTYPE html>
<html lang="en" dir="ltr">
  <head>
    <meta charset="utf-8">
    <title>Hello Decentralized Web!</title>
    <link rel="stylesheet" href="style.css">
  </head>
  <body>
    <h1>Hello Decentralized Web!</h1>
    <img src="wsdl-logo.png" alt="">
    <p>We, at <a href="https://ws-dl.cs.odu.edu/">Web Science a
nd Digital Libraries Research Group</a>, tweet from the <a href
="https://twitter.com/WebSciDL">@WebSciDL</a> handle.</p>
  </body>
</html>
```

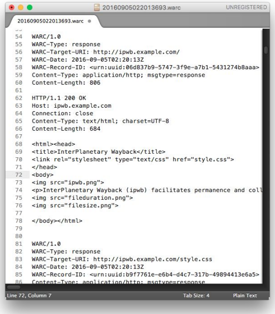
```
WARC/1.0
WARC-Type: response
WARC-Target-URI: https://www.cs.odu.edu/~salam/dweb/
WARC-Date: 2018-08-02T15:12:34Z
WARC-Record-ID: <urn:uuid:6e829124-b8a3-02ae-4912-f9be5978e09a>
Content-Type: application/http; msgtype=response
Content-Length: 654

HTTP/1.1 200 OK
Server: nginx
Date: Thu, 02 Aug 2018 15:06:34 GMT
Content-Type: text/html
Transfer-Encoding: chunked
Connection: keep-alive
Vary: Accept-Encoding
Front-End-Https: on

<!DOCTYPE html>
<html lang="en" dir="ltr">
  <head>
    <meta charset="utf-8">
    <title>Hello Decentralized Web!</title>
    <link rel="stylesheet" href="style.css">
  </head>
  <body>
    <h1>Hello Decentralized Web!</h1>
    <img src="wsdl-logo.png" alt="">
    <p>We, at <a href="https://ws-dl.cs.odu.edu/">Web Science a
nd Digital Libraries Research Group</a>, tweet from the <a href
="https://twitter.com/WebSciDL">@WebSciDL</a> handle.</p>
  </body>
</html>
hello-dweb.warc
```

@ibnesayeed

How Wayback Machine Works?

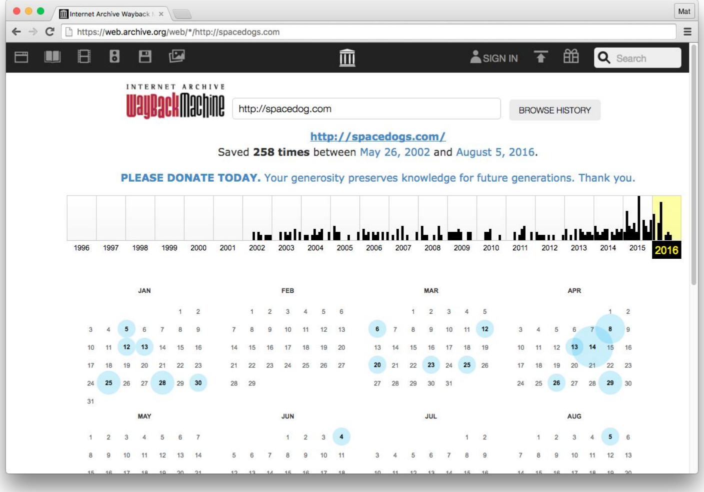
Archival Indexer

processes



```
WARC/1.0
WARC-Type: response
WARC-Target-URI: http://ipwb.example.com/
WARC-Date: 2016-09-05T02:20:13Z
WARC-Record-ID: <urn:uuid:06d837b9-5747-3f9e-a7b1-5431274b8aaa>
Content-Type: application/http; msgtype=response
Content-Length: 806

HTTP/1.1 200 OK
Host: ipwb.example.com
Connection: close
Content-Type: text/html; charset=UTF-8
Content-Length: 684

<html><head>
<title>InterPlanetary Wayback</title>
<link rel="stylesheet" type="text/css" href="style.css">
</head>
<body>
<img src="ipwb.png">
<p>InterPlanetary Wayback (ipwb) facilitates permanence and coll
<img src="fileduration.png">
<img src="filesize.png">

</body></html>


WARC/1.0
WARC-Type: response
WARC-Target-URI: http://ipwb.example.com/style.css
WARC-Date: 2016-09-05T02:20:13Z
WARC-Record-ID: <urn:uuid:b9f7761e-e6b4-d4c7-317b-49894413e6a5>
Content-Type: application/http; msgtype=response
```

Present WARC content to user



https://web.archive.org/web/*/http://spacedogs.com

http://spacedog.com

http://spacedogs.com/

Saved **258 times** between May 26, 2002 and August 5, 2016.

PLEASE DONATE TODAY. Your generosity preserves knowledge for future generations. Thank you.

Outputs references

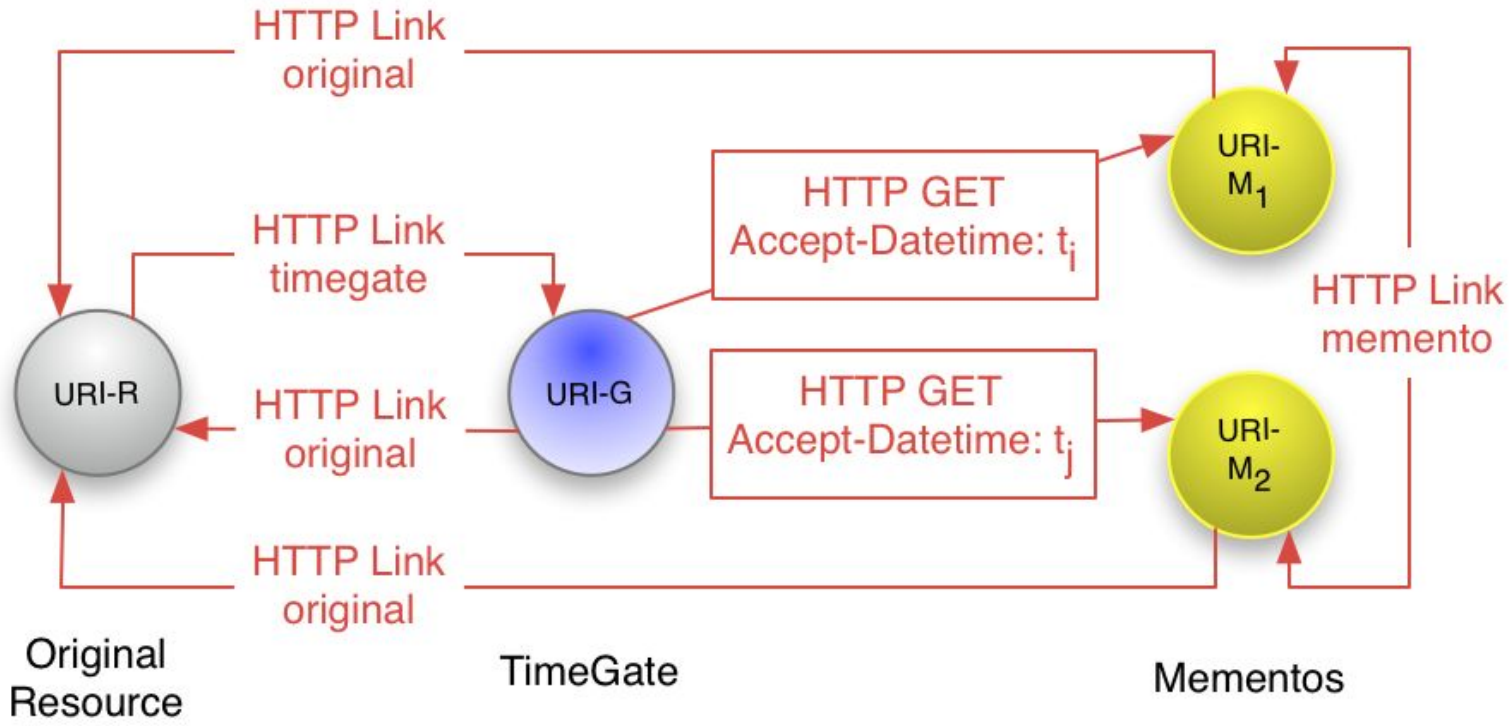
read archived content

Archival Index (e.g., CDXJ)

reads (file, offset)

Replay Engine

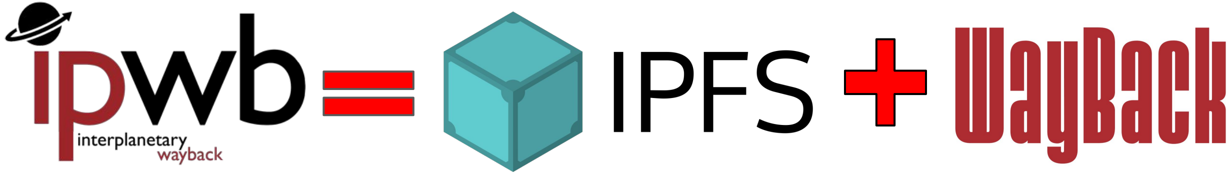
Memento: Time Dimension to the Web



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

@ibnesayeed

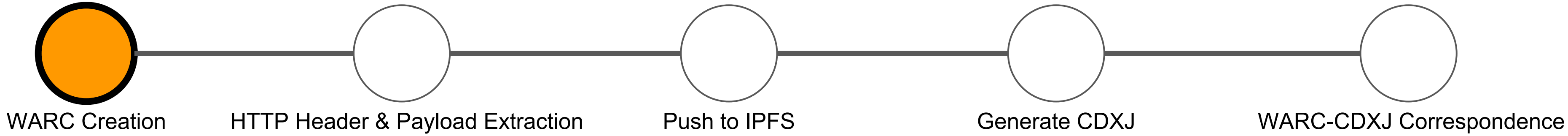
Why IPWB?



- Persistence of archived web dependent on resilience of organizations
- Availability of data is subject to censorship
- Redundancy in web archive files of exact duplicate content
- Lack of public participation in web archiving
- Discoverability issue of small web archives

@ibnesayeed

Indexing



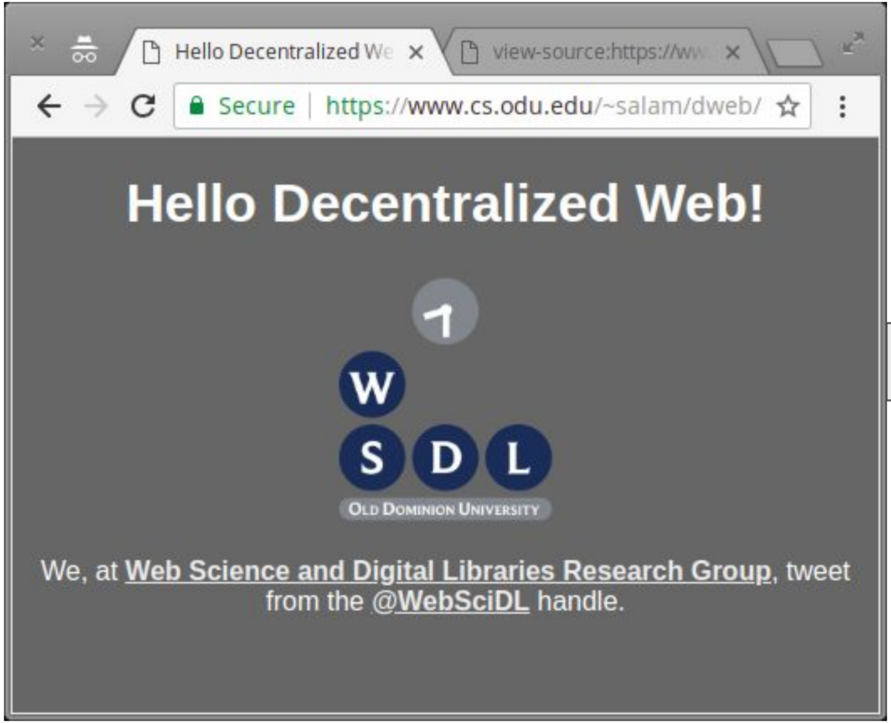
WARC Creation

HTTP Header & Payload Extraction

Push to IPFS

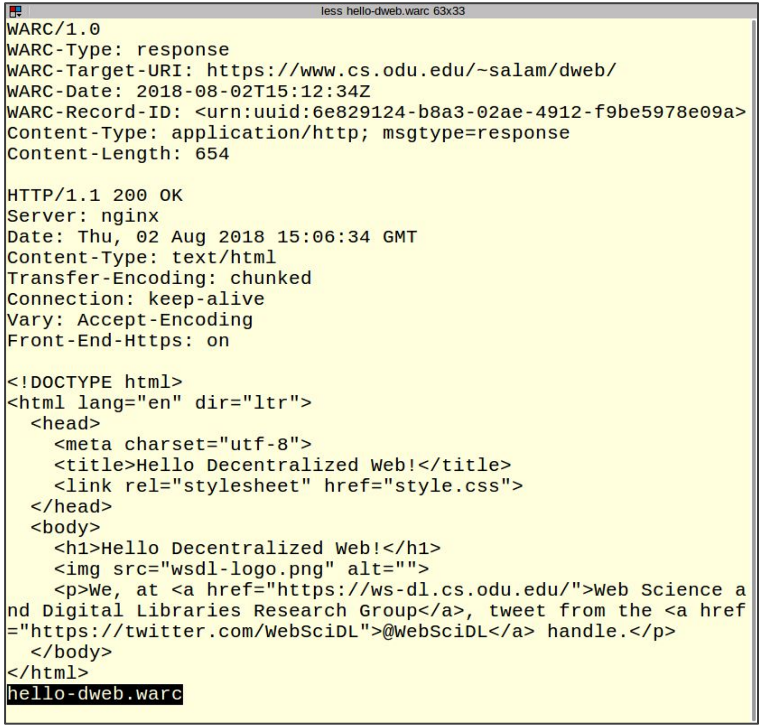
Generate CDXJ

WARC-CDXJ Correspondence

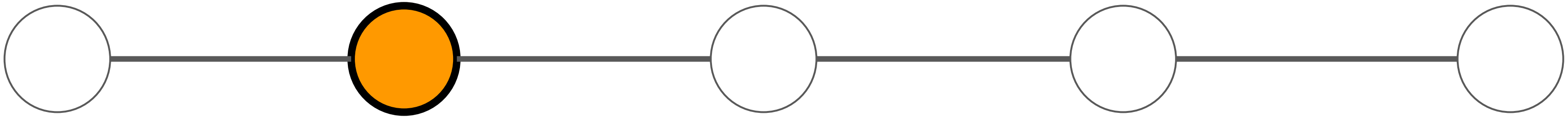


Hello Decentralized Web!

We, at Web Science and Digital Libraries Research Group, tweet from the @WebSciDL handle.



```
WARC/1.0
WARC-Type: response
WARC-Target-URI: https://www.cs.odu.edu/~salam/dweb/
WARC-Date: 2018-08-02T15:12:34Z
WARC-Record-ID: <urn:uuid:6e829124-b8a3-02ae-4912-f9be5978e09a>
Content-Type: application/http; msgtype=response
Content-Length: 654

HTTP/1.1 200 OK
Server: nginx
Date: Thu, 02 Aug 2018 15:06:34 GMT
Content-Type: text/html
Transfer-Encoding: chunked
Connection: keep-alive
Vary: Accept-Encoding
Front-End-Https: on

<!DOCTYPE html>
<html lang="en" dir="ltr">
  <head>
    <meta charset="utf-8">
    <title>Hello Decentralized Web!</title>
    <link rel="stylesheet" href="style.css">
  </head>
  <body>
    <h1>Hello Decentralized Web!</h1>
    <img src="wsdl-logo.png" alt="">
    <p>We, at <a href="https://ws-dl.cs.odu.edu/">Web Science a
nd Digital Libraries Research Group</a>, tweet from the <a href
="https://twitter.com/WebSciDL">@WebSciDL</a> handle.</p>
  </body>
</html>
hello-dweb.warc
```



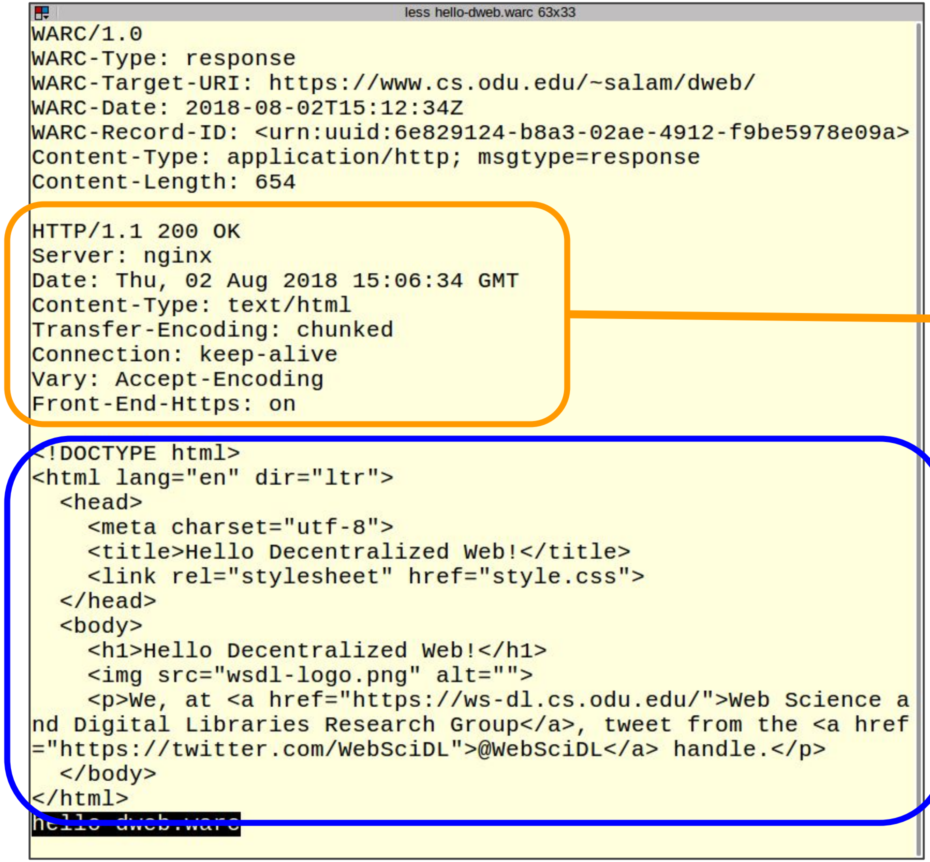
WARC Creation

HTTP Header & Payload Extraction

Push to IPFS

Generate CDXJ

WARC-CDXJ Correspondence



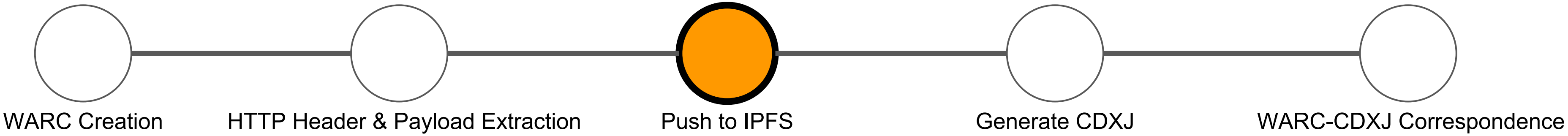
less hello-dweb.warc 63x33

```
WARC/1.0
WARC-Type: response
WARC-Target-URI: https://www.cs.odu.edu/~salam/dweb/
WARC-Date: 2018-08-02T15:12:34Z
WARC-Record-ID: <urn:uuid:6e829124-b8a3-02ae-4912-f9be5978e09a>
Content-Type: application/http; msgtype=response
Content-Length: 654

HTTP/1.1 200 OK
Server: nginx
Date: Thu, 02 Aug 2018 15:06:34 GMT
Content-Type: text/html
Transfer-Encoding: chunked
Connection: keep-alive
Vary: Accept-Encoding
Front-End-Https: on

<!DOCTYPE html>
<html lang="en" dir="ltr">
  <head>
    <meta charset="utf-8">
    <title>Hello Decentralized Web!</title>
    <link rel="stylesheet" href="style.css">
  </head>
  <body>
    <h1>Hello Decentralized Web!</h1>
    <img src="wsdl-logo.png" alt="">
    <p>We, at <a href="https://ws-dl.cs.odu.edu/">Web Science a
nd Digital Libraries Research Group</a>, tweet from the <a href
="https://twitter.com/WebSciDL">@WebSciDL</a> handle.</p>
  </body>
</html>
hello-dweb.warc
```

HTTP HEADER BLOCK

HTTP PAYLOAD BLOCK

@ibnesayeed



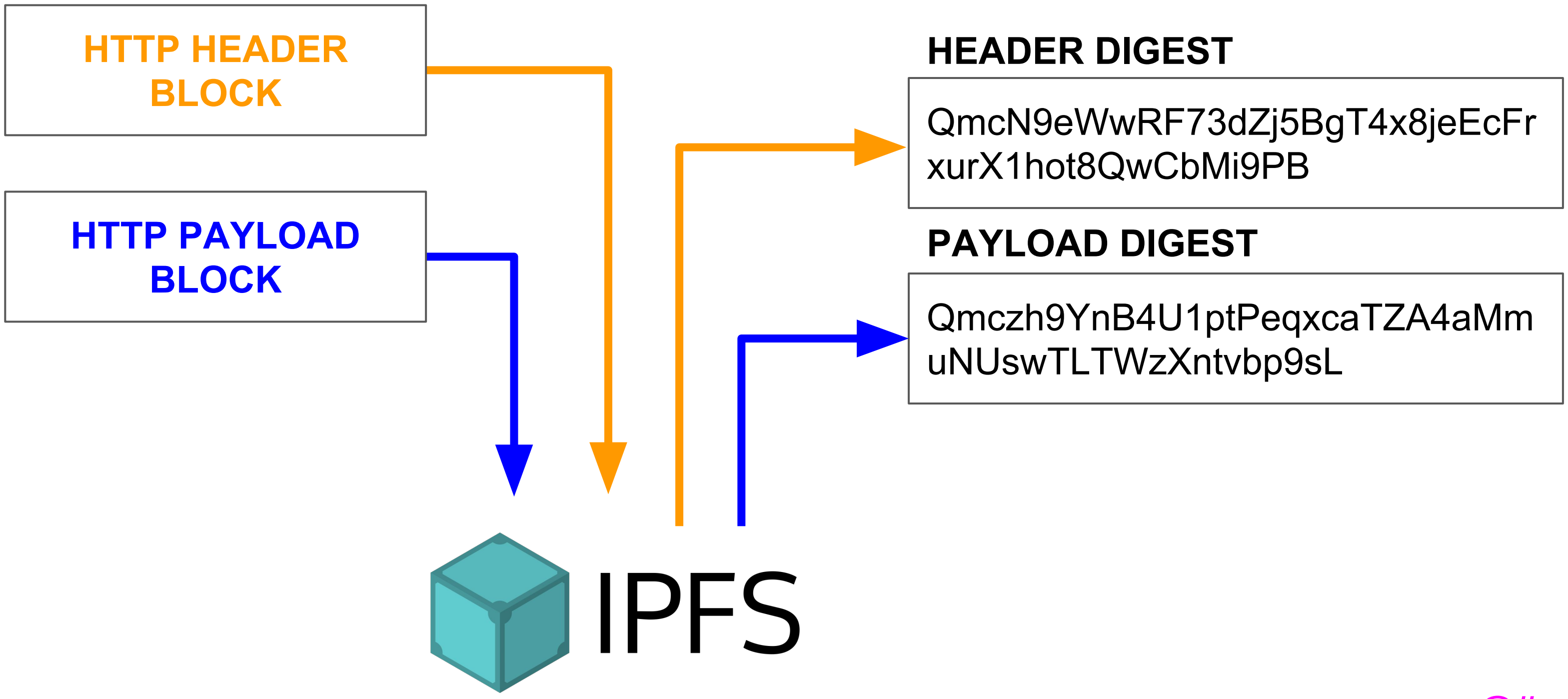
WARC Creation

HTTP Header & Payload Extraction

Push to IPFS

Generate CDXJ

WARC-CDXJ Correspondence



HTTP HEADER BLOCK

HTTP PAYLOAD BLOCK

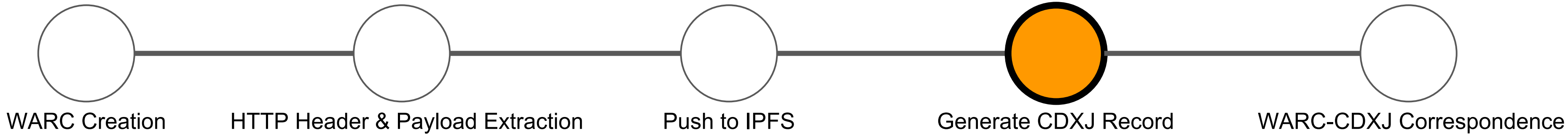
HEADER DIGEST

QmcN9eWwRF73dZj5BgT4x8jeEcFr
xurX1hot8QwCbMi9PB

PAYLOAD DIGEST

Qmczh9YnB4U1ptPeqxcaTZA4aMm
uNUswTLTWzXntvbp9sL

@ibnesayeed



WARC Creation

HTTP Header & Payload Extraction

Push to IPFS

Generate CDXJ Record

WARC-CDXJ Correspondence

HEADER DIGEST

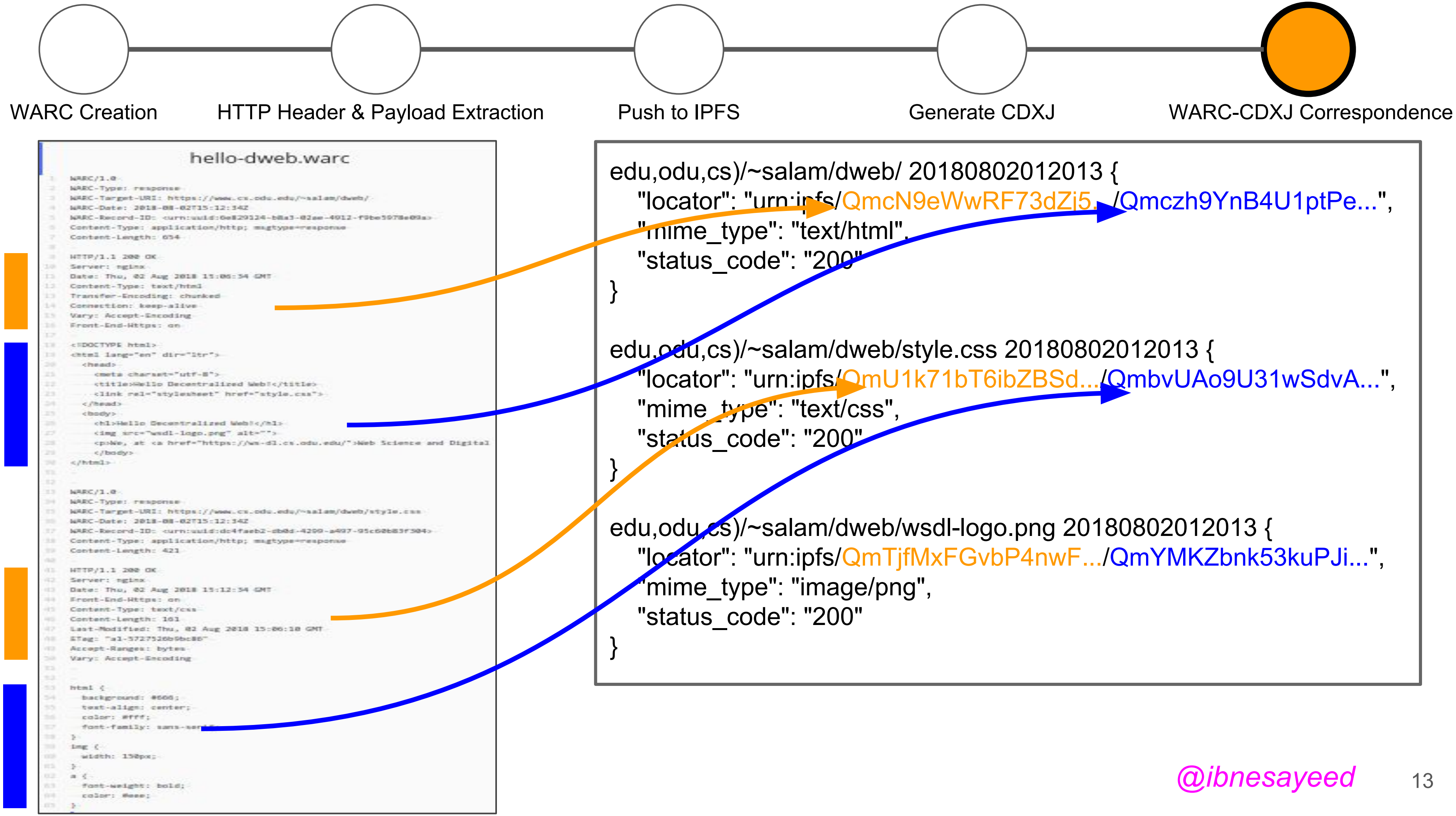
QmcN9eWwRF73dZj5BgT4x8jeEcFrxurX1hot8QwCbMi9PB

PAYLOAD DIGEST

Qmczh9YnB4U1ptPeqxcaTZA4aMmuNUswTLTWzXntvbp9sL

```
com,example,ipwb)/ 20180802012013 {
        "locator":"urn:ipfs/QmcN9eWwRF73dZj5BgT4x8jeEcFrxurX1hot8QwCbMi9PB/Qmczh9YnB4U1ptPeqxca
        TZA4aMmuNUswTLTWzXntvbp9sL",
        "mime_type": "text/html",
        "status_code": 200,
        “other_fields”: “other values...”
}
// * This is a single-line record, line breaks and indentation are added for readability only
```

CDXJ: http://ws-dl.blogspot.com/2015/09/2015-09-10-cdxj-object-resource-stream.html



WARC Creation

HTTP Header & Payload Extraction

Push to IPFS

Generate CDXJ

WARC-CDXJ Correspondence

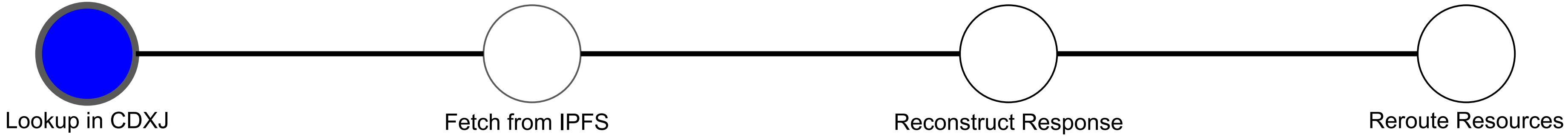
hello-dweb.warc

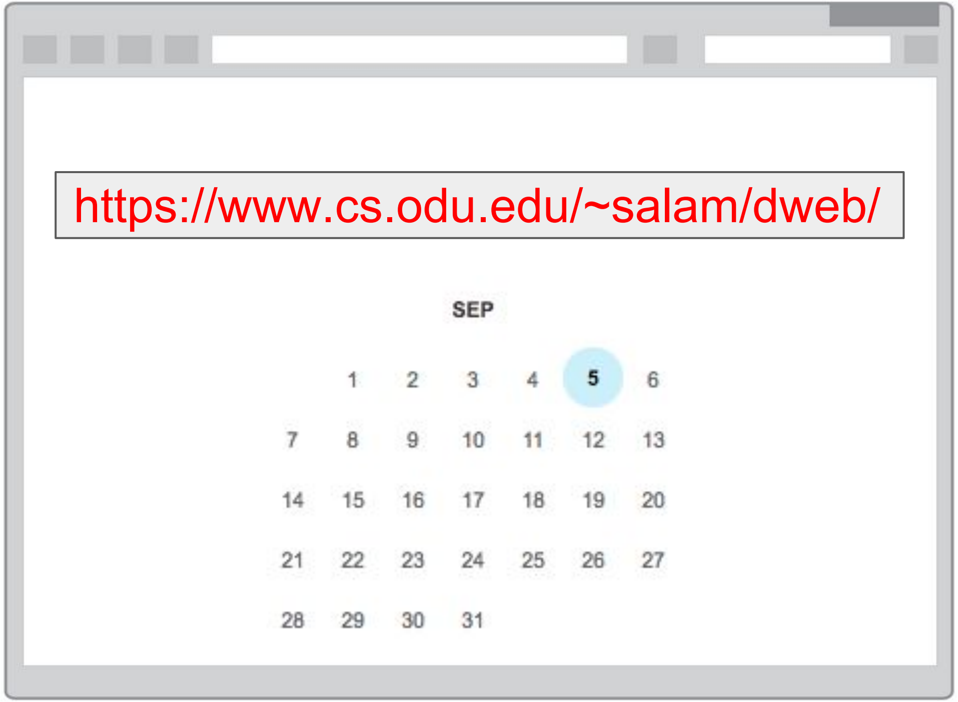
```
WARC/1.0
WARC-Type: response
WARC-Target-URI: https://www.cs.odu.edu/~salam/dweb/
WARC-Date: 2018-08-02T15:12:34Z
WARC-Record-ID: <urn:uuid:6e829124-b8a3-02ae-4012-f9be5978e09a>
Content-Type: application/http; msgtype=response
Content-Length: 654

HTTP/1.1 200 OK
Server: nginx
Date: Thu, 02 Aug 2018 15:06:34 GMT
Content-Type: text/html
Transfer-Encoding: chunked
Connection: keep-alive
Vary: Accept-Encoding
Front-End-Https: on

<!DOCTYPE html>
<html lang="en" dir="ltr">
  <head>
    <meta charset="utf-8">
    <title>Hello Decentralized Web!</title>
    <link rel="stylesheet" href="style.css">
  </head>
  <body>
    <h1>Hello Decentralized Web!</h1>
    <img src="wsdl-logo.png" alt="">
    <p>We, at <a href="https://ws-dl.cs.odu.edu/">Web Science and Digital
    </body>
</html>


WARC/1.0
WARC-Type: response
WARC-Target-URI: https://www.cs.odu.edu/~salam/dweb/style.css
WARC-Date: 2018-08-02T15:12:34Z
WARC-Record-ID: <urn:uuid:dc4faeb2-db0d-4209-a497-95c60b83f304>
Content-Type: application/http; msgtype=response
Content-Length: 421

HTTP/1.1 200 OK
Server: nginx
Date: Thu, 02 Aug 2018 15:12:34 GMT
Front-End-Https: on
Content-Type: text/css
Content-Length: 161
Last-Modified: Thu, 02 Aug 2018 15:06:10 GMT
ETag: "a1-5727526b9bc86"
Accept-Ranges: bytes
Vary: Accept-Encoding


html {
  background: #666;
  text-align: center;
  color: #fff;
  font-family: sans-serif;
}
img {
  width: 150px;
}
a {
  font-weight: bold;
  color: #eee;
}
```

```
edu,odu,cs)/~salam/dweb/ 20180802012013 {
    "locator": "urn:ipfs/QmcN9eWwRF73dZj5.../Qmczh9YnB4U1ptPe...",
    "mime_type": "text/html",
    "status_code": "200"
}

edu,odu,cs)/~salam/dweb/style.css 20180802012013 {
    "locator": "urn:ipfs/QmU1k71bT6ibZBSd.../QmbvUAo9U31wSdvA...",
    "mime_type": "text/css",
    "status_code": "200"
}

edu,odu,cs)/~salam/dweb/wsdl-logo.png 20180802012013 {
    "locator": "urn:ipfs/QmTjfMxFGvbP4nwF.../QmYMKZbnk53kuPJi...",
    "mime_type": "image/png",
    "status_code": "200"
}
```

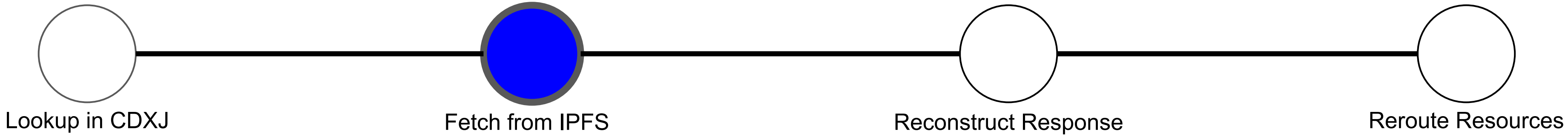
@ibnesayeed

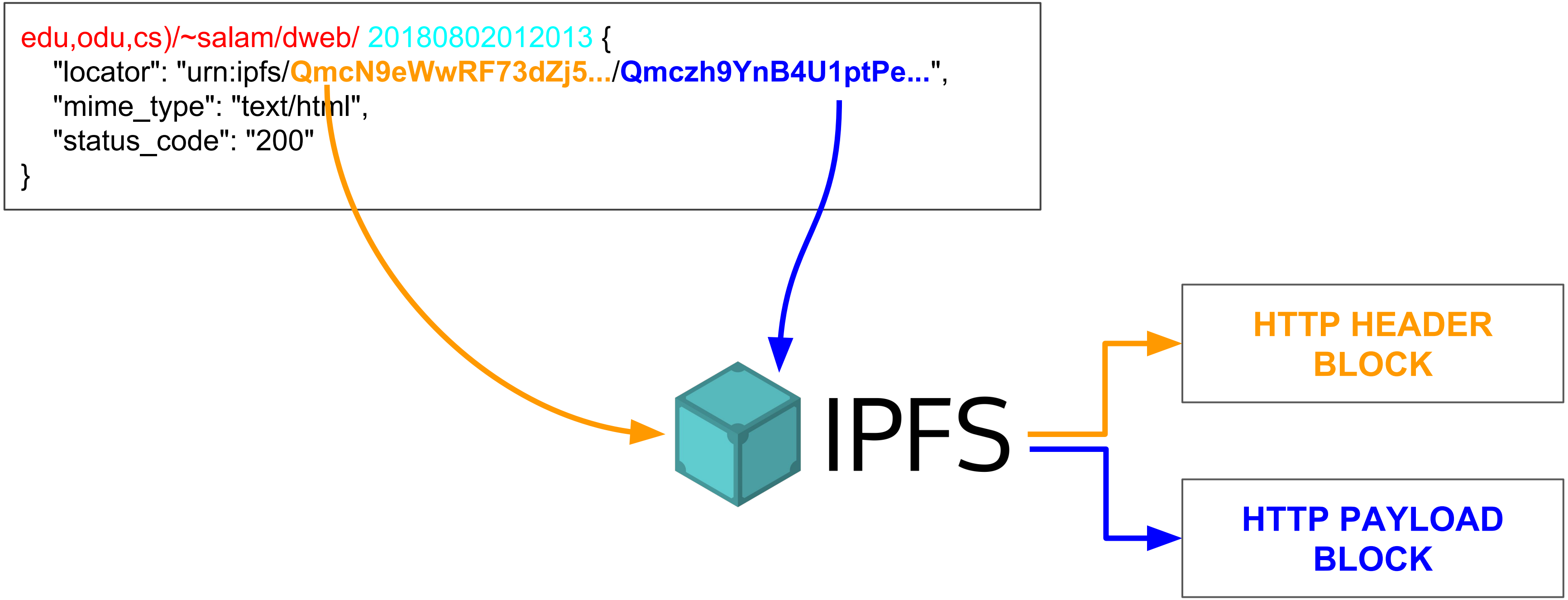
Replay



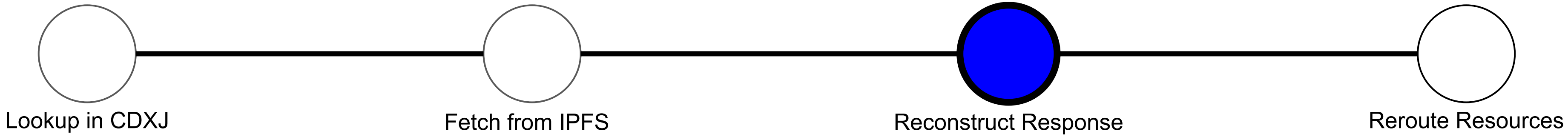

https://www.cs.odu.edu/~salam/dweb/

```
edu,odu,cs)/~salam/dweb/ 20180802012013 {
    "locator": "urn:ipfs/QmcN9eWwRF.../Qmczh9YnB4...",
    "mime_type": "text/html",
    "status_code": "200"
}

edu,odu,cs)/~salam/dweb/style.css 20180802012013 {
    "locator": "urn:ipfs/QmU1k71bT6.../QmbvUAo9U3...",
    "mime_type": "text/css",
    "status_code": "200"
}

edu,odu,cs)/~salam/dweb/wsdl-logo.png 20180802012013 {
    "locator": "urn:ipfs/QmTjfMxFGv.../QmYMKZbnk5...",
    "mime_type": "image/png",
    "status_code": "200"
}
```





```
edu,odu,cs)/~salam/dweb/ 20180802012013 {
   "locator": "urn:ipfs/QmcN9eWwRF73dZj5.../Qmczh9YnB4U1ptPe...",
   "mime_type": "text/html",
   "status_code": "200"
}
```

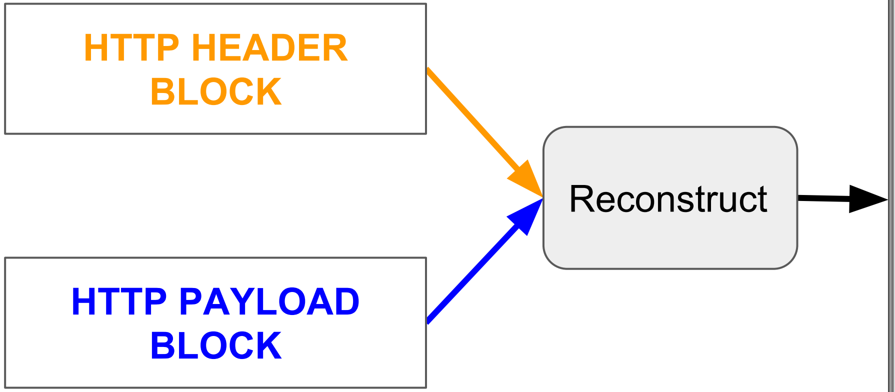


Lookup in CDXJ

Fetch from IPFS

Reconstruct Response

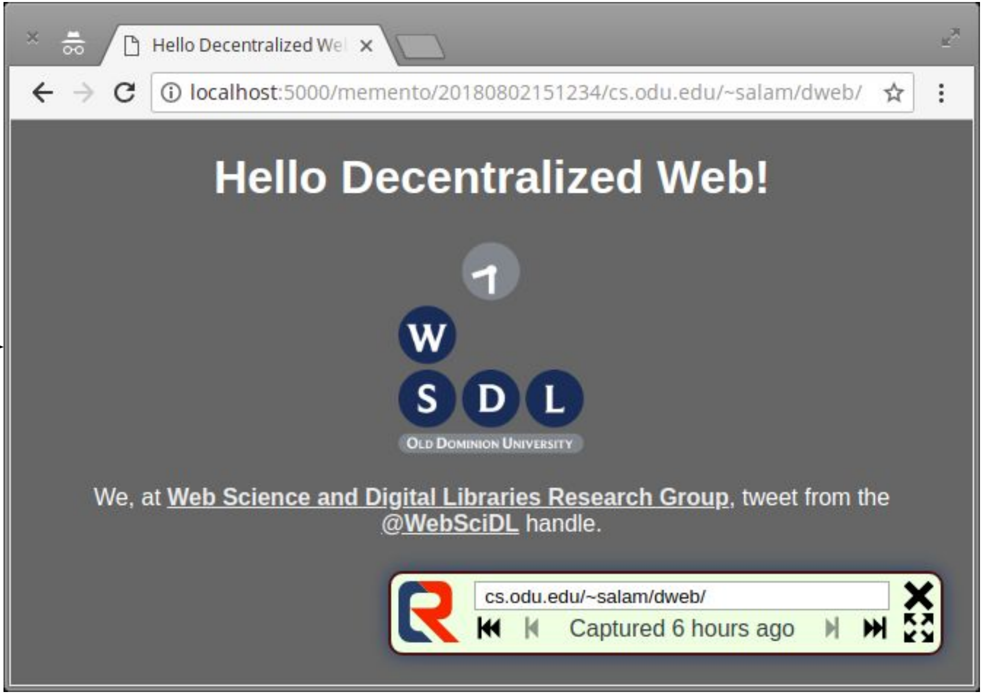
Reroute Resources



HTTP HEADER BLOCK

HTTP PAYLOAD BLOCK

Reconstruct



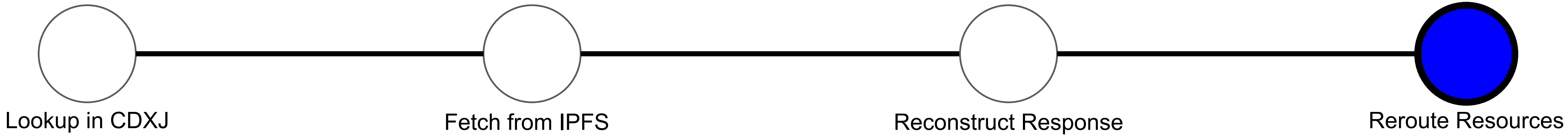
Hello Decentralized We ×
localhost:5000/memento/20180802151234/cs.odu.edu/~salam/dweb/
W
S D L
Old Dominion University

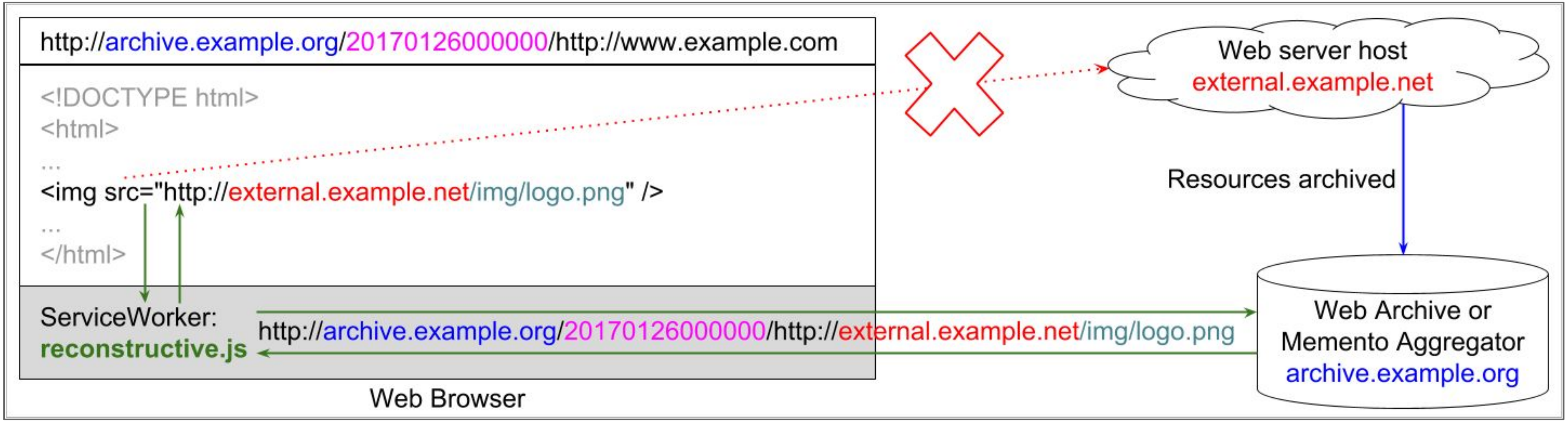
Hello Decentralized Web!

We, at Web Science and Digital Libraries Research Group, tweet from the @WebSciDL handle.

cs.odu.edu/~salam/dweb/

Captured 6 hours ago





http://archive.example.org/20170126000000/http://www.example.com

```
<!DOCTYPE html>
<html>
...
<img src="http://external.example.net/img/logo.png" />
...
</html>
```

ServiceWorker:
reconstructive.js

http://archive.example.org/20170126000000/http://external.example.net/img/logo.png

Web Browser

Web server host
external.example.net

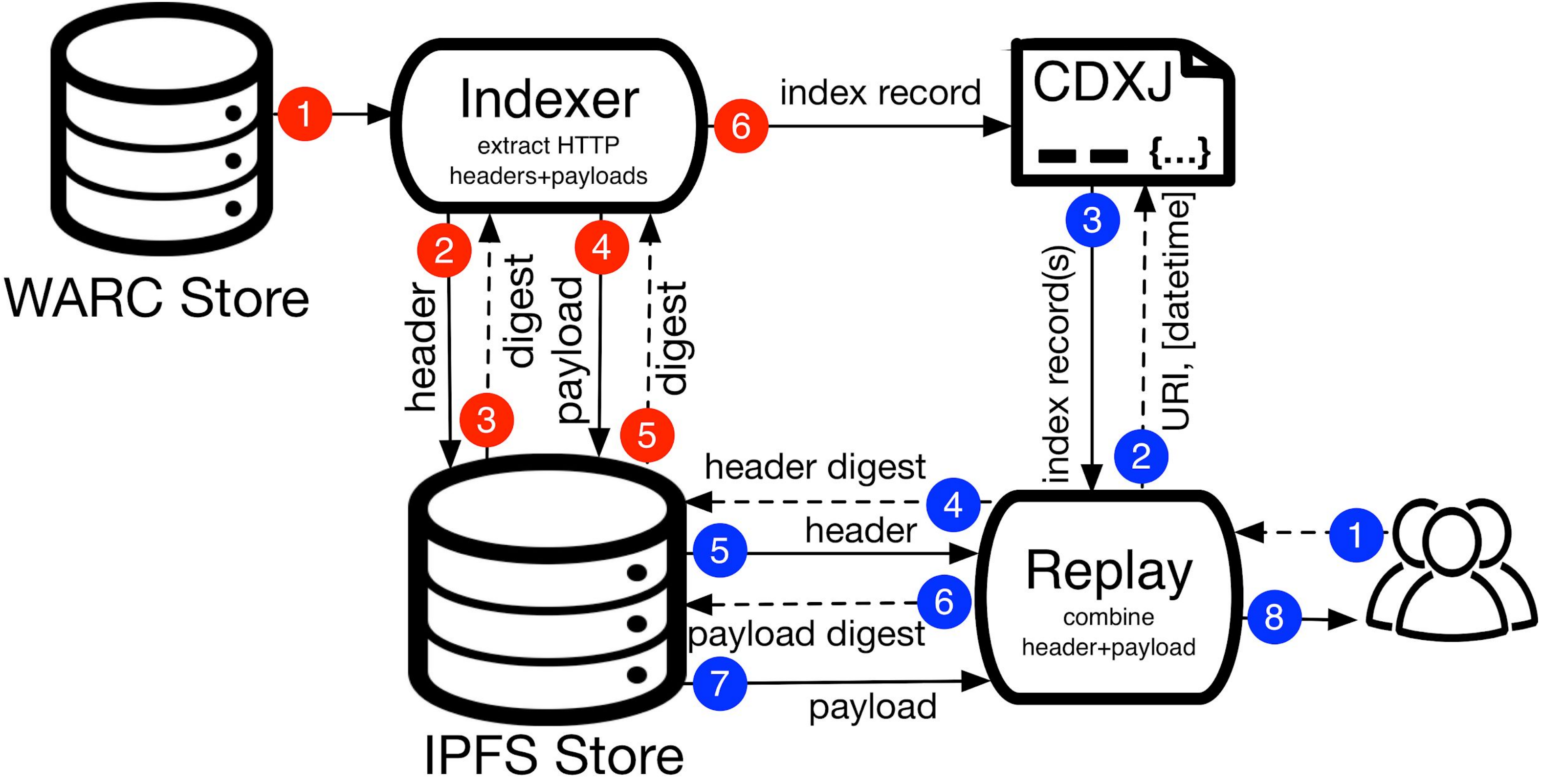
Resources archived

Web Archive or
Memento Aggregator
archive.example.org

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

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

IPWB Indexing and Replay



@ibnesayeed

Decentralization

Current Issues

- IPFS is permanent, but not persistent
- DHT-based IPNS is history-unaware
- CDXJ index, a critical piece of replay, is centralized

Persistence

- Data persistence is critical for web archiving
- A decentralized storage with sufficient replication is needed
- Memory organizations should contribute storage infrastructure
- Qri, Filecoin, IPFS-Cluster, IPFS-Sync etc. can be helpful

IPNS: InterPlanetary Naming System

URI	IPFS Hash
http://example.org/yuri.jpg	QmZAD4xeeNeYF3TmwWgBXypLKTiCGwGRMXHW7MtheWKtw4
http://example.com/style.css	QmbvUAo9U31wSdvARjvbPeVBTAwCjN1kyPhQ4ho3n8TAZo
http://example.com/logo.png	QmYMKZbnk53kuPJirahJHGevCCy2afLyePRdX38TukFUwd

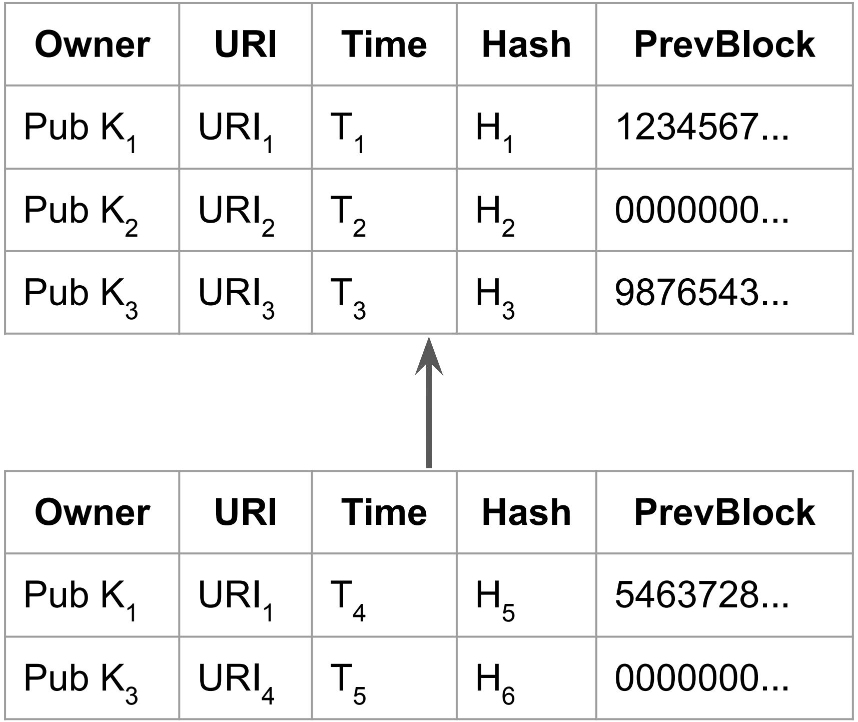
How about changes and history?

http://example.com/style.css	QmAAykNAehCjbn43wvUQUd1ovhTPjRo39VvoT8bSA1ZBwP

@ibnesayeed

IPNS Blockchain

- URI → Latest hash
- URI + DateTime → A historical hash
- URI → List historical hashes with times



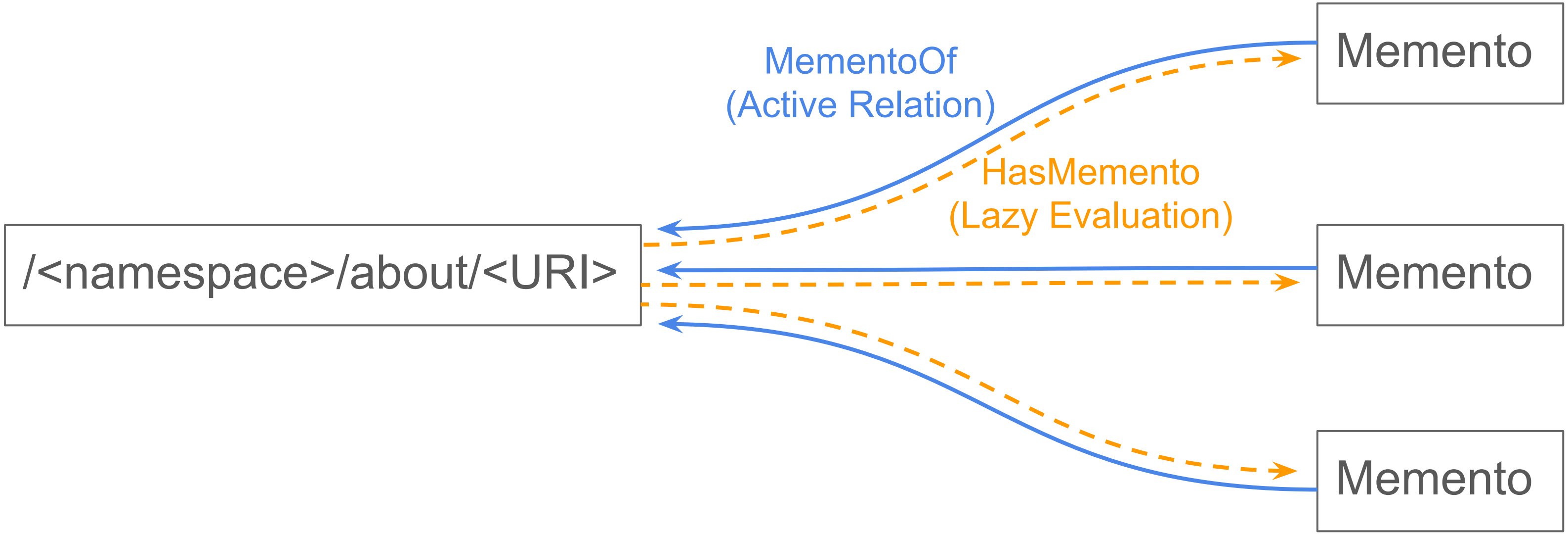
Owner	**URI**	**Time**	**Hash**	**PrevBlock**
Pub K_1	URI$_1$	T_1	H_1	1234567...
Pub K_2	URI$_2$	T_2	H_2	0000000...
Pub K_3	URI$_3$	T_3	H_3	9876543...

Owner	**URI**	**Time**	**Hash**	**PrevBlock**
Pub K_1	URI$_1$	T_4	H_5	5463728...
Pub K_3	URI$_4$	T_5	H_6	0000000...

IPNS + Blockchain + Memento

https://github.com/oduwsdl/IPNS-Blockchain

Lazy Relationship Evaluation



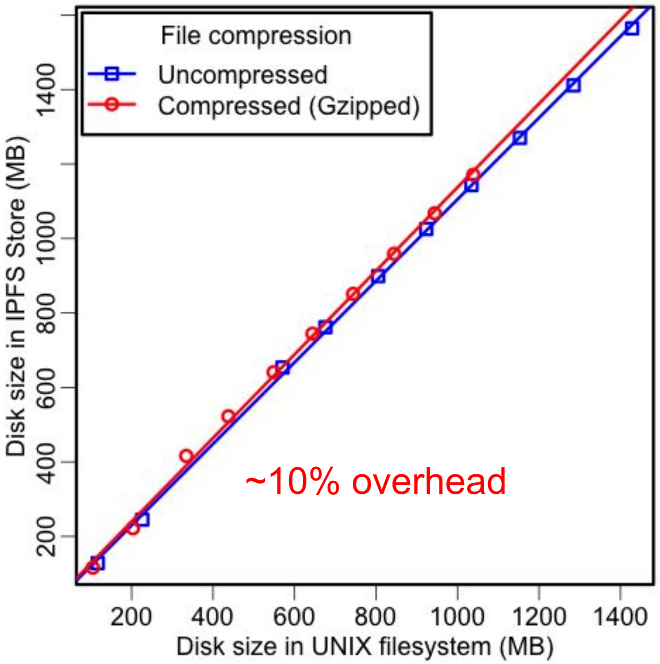
IP-LD to the rescue?

https://github.com/oduwsdl/ipwb/issues/61

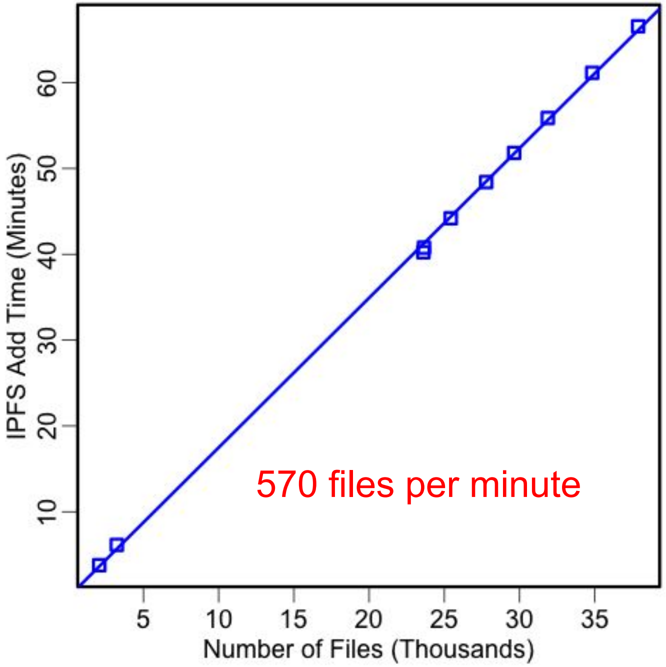
@ibnesayeed

Evaluation

Storage Space and Time Overhead



~10% overhead



570 files per minute

- Reported IPFS slowness https://github.com/ipfs/go-ipfs/issues/1216
    - Has since been fixed, but we did not evaluate again

Replay Time

- 600 requests in 222 seconds
- Slower than PyWB (which took 5.26 seconds)
- File vs. rich object based retrieval
- Never expiring cache

https://github.com/ibnesayeed/ipfsapi-concurrency-test

Future Works

- Evaluate the improved IPFS on large dataset
- Evaluate deduplication
- Implement an index-free collaborative archiving system
- Utilize IPNS to reference URI-Rs with datetime

Conclusions

- A proof of concept system to leverage a novel approach to archiving and retrieval
- Storage and time costs evaluation and qualitative analysis
- It can only work for small archives in its current state
- A path to answer "who will archive the archives?"
- More work to be done to make it a truly decentralized archiving system

InterPlanetary Wayback

The Next Step Towards Decentralized Web Archiving

<u>Sawood Alam</u>, Mat Kelly, Michele C. Weigle, Michael L. Nelson

`http://github.com/oduwsdl/ipwb`

@WebSciDL

Supported in part by Protocol Labs, AMF 11600663, and NSF IIS-1526700