



InterPlanetary Wayback

Peer-to-Peer Permanence of Web Archives

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

Old Dominion University

Web Science and Digital Libraries Research Group

Norfolk, Virginia, USA

@WebSciDL

<http://github.com/oduwsdl/ipwb>

TPDL 2016

Hannover, Germany

September 7, 2016



Background - IPFS

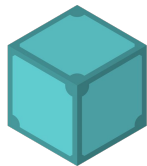


- Hypermedia distributed protocol
- IPFS entity hashes are content addressed
 - Content changes → different hash produced
 - Inherent potential for de-duplication of content
- Files accessible via HTTP: <http://ipfs.io/<hash>>
- Built on trust chains for provenance

Content addressing



<http://foo.com/spaceDog.jpg>



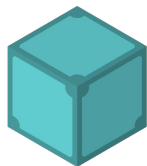
IPFS

QmZAD4xeeNeYF3TmwWgBXypLKTiCGwGRMXHW7MtheWKtw4

===



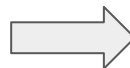
<http://example.org/yuri.jpg>



IPFS

QmZAD4xeeNeYF3TmwWgBXypLKTiCGwGRMXHW7MtheWKtw4

\$ ipfs cat QmZAD4xeeNeYF3TmwWgBXypLKTiCGwGRMXHW7MtheWKtw4 > doge.jpg



Background - WARC

WARC response record

Warc-response
header

HTTP resp header

HTTP resp payload

WARCs also contain:

- HTTP requests
- warc-info
- warc-metadata records
- etc.



uses only warc-response records

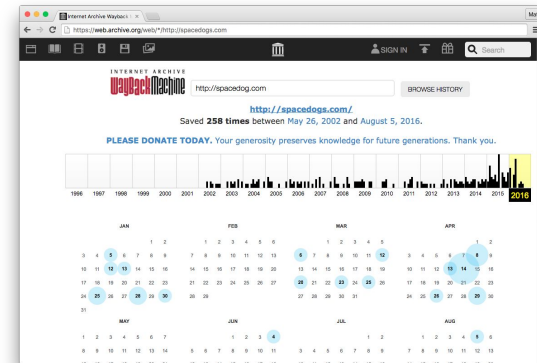
A screenshot of a web browser window displaying a WARC file named '20160905022013693.warc'. The browser's address bar shows the file path. The main content area displays the raw text of a WARC response record, with line numbers on the left. The record starts with 'WARC/1.0' and 'WARC-Type: response'. It includes a 'WARC-Target-URI' pointing to 'http://ipwb.example.com/'. The 'WARC-Date' is '2016-09-05T02:20:13Z'. The 'WARC-Record-ID' is a long UUID. The 'Content-Type' is 'application/http; msgtype=response' and the 'Content-Length' is 806. This is followed by an HTTP 1.1 200 OK status line and headers for 'Host', 'Connection', 'Content-Type', and 'Content-Length'. The body of the record is an HTML document with a title 'InterPlanetary Wayback', a link to 'style.css', and three image tags with src attributes 'ipwb.png', 'fileduration.png', and 'filesize.png'. The record ends with 'WARC/1.0' and another 'WARC-Type: response' line, followed by similar metadata and a 'Content-Type: application/http; msgtype=response' line. The browser's status bar at the bottom indicates 'Line 72, Column 7', 'Tab Size: 4', and 'Plain Text'.

The diagram illustrates the WARC playback system architecture, showing the flow of data and control between several components:

- Archival Indexer**: Processes inputs and outputs the **Archival Index (e.g., CDXJ)**.
- Archival Index (e.g., CDXJ)**: A document icon representing the index of archived content.
- Replay Engine**: Reads the index (file, offset) and outputs **Present WARC content to user**.
- Present WARC content to user**: A smiley face icon representing the user interface.
- WARC Content**: A screenshot of a WARC file showing HTTP request and response data, including headers, body, and status codes.

The flow is as follows:

- The **Archival Indexer** processes inputs and outputs the **Archival Index (e.g., CDXJ)**.
- The **Replay Engine** reads the index (file, offset) and outputs **Present WARC content to user**.
- The **Present WARC content to user** is displayed to the user (smiley face).
- The **WARC Content** is a screenshot of a WARC file showing HTTP request and response data, including headers, body, and status codes.

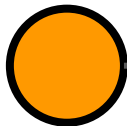


Motivation

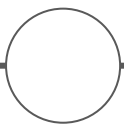


- Persistence of archived web data dependent on resilience of organization and availability of data
- Remove massive redundancy in web archive files of exact duplicate content
- Determine feasibility of pushing WARC files into IPFS

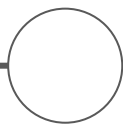
Indexing



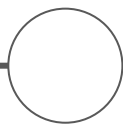
WARC Creation



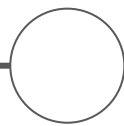
HTTP Header & Payload Extraction



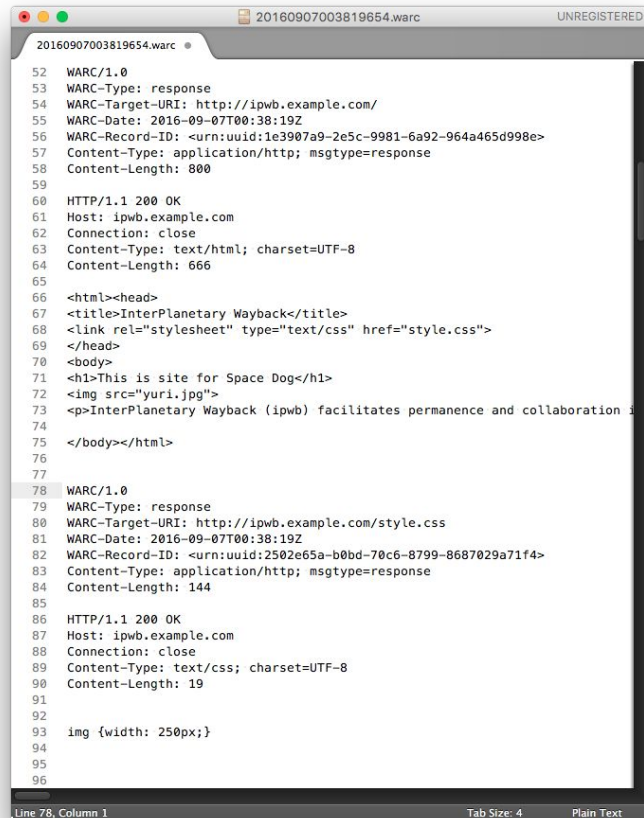
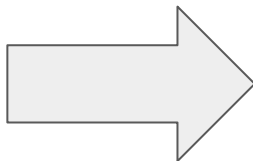
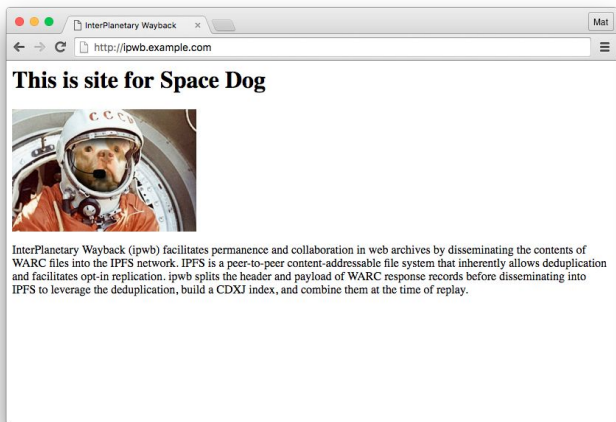
Push to IPFS

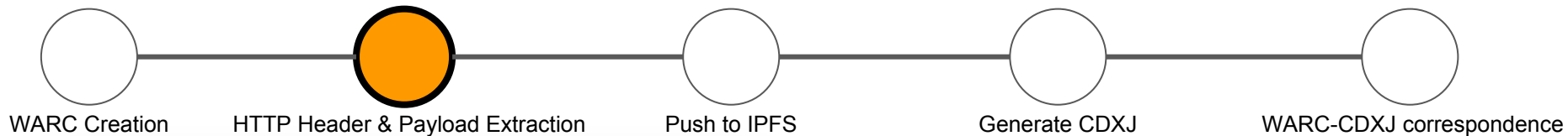


Generate CDXJ



WARC-CDXJ correspondence

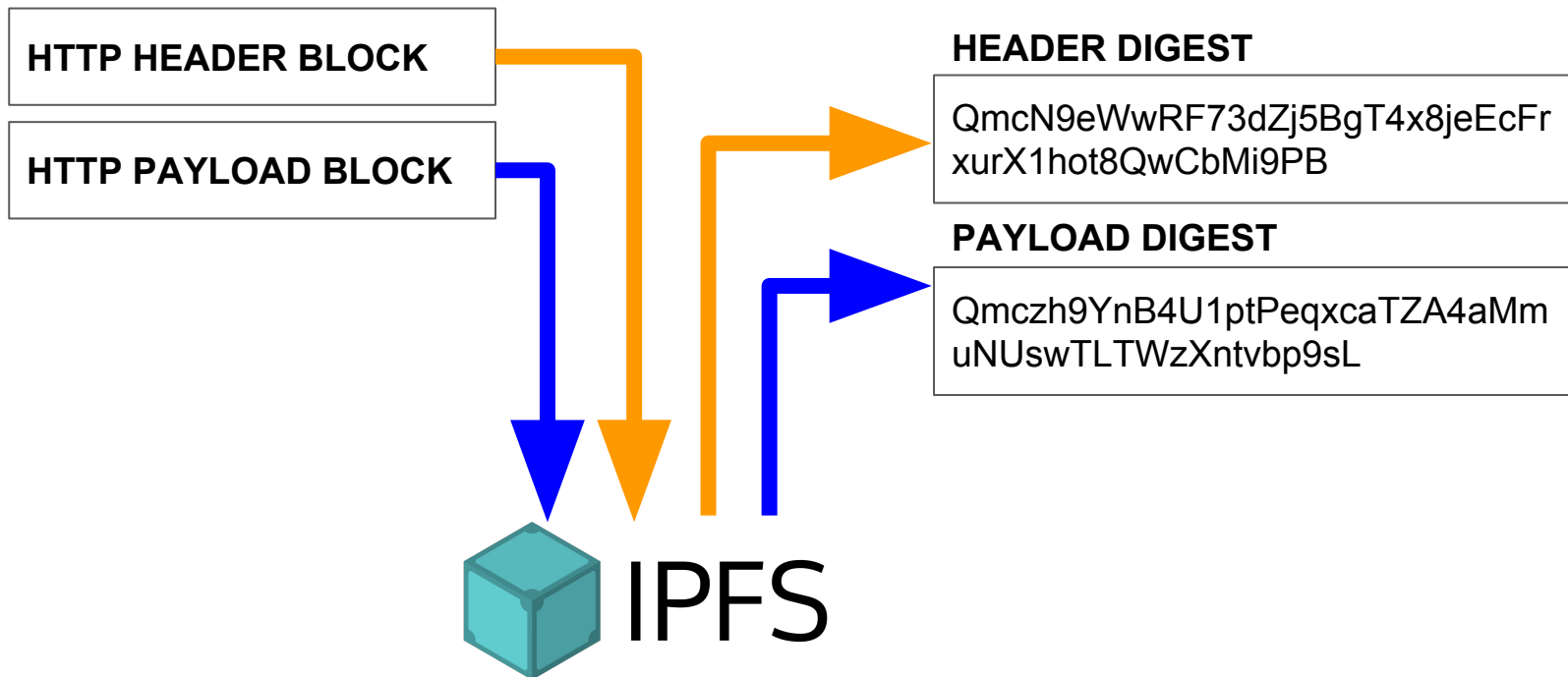
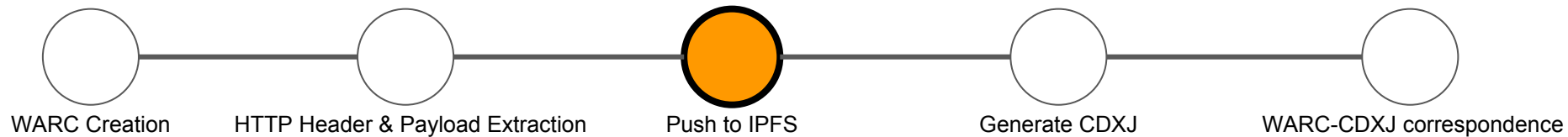


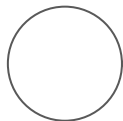


```
20160907003819654.warc x UNREGISTERED
20160907003819654.warc x
52 WARC/1.0
53 WARC-Type: response
54 WARC-Target-URI: http://ipwb.example.com/
55 WARC-Date: 2016-09-07T00:38:19Z
56 WARC-Record-ID: <urn:uuid:1e3907a9-2e5c-9981-6a92-964a465d998e>
57 Content-Type: application/http; msgtype=response
58 Content-Length: 800
59
60 HTTP/1.1 200 OK
61 Host: ipwb.example.com
62 Connection: close
63 Content-Type: text/html; charset=UTF-8
64 Content-Length: 666
65
66 <html><head>
67 <title>InterPlanetary Wayback</title>
68 <link rel="stylesheet" type="text/css" href="style.css">
69 </head>
70 <body>
71 <h1>This is site for Space Dog</h1>
72 
73 <p>InterPlanetary Wayback (ipwb) facilitates permanence and collabora
74
75 </body></html>
76
77
78 WARC/1.0
79 WARC-Type: request
80 WARC-Target-URI: http://ipwb.example.com/style.css
81 WARC-Date: 2016-09-07T00:38:19Z
82 WARC-Concurrent-To: <urn:uuid:2d315cc1-a34d-3945-c5d9-ab4c7ac13fe6>
83 WARC-Record-ID: <urn:uuid:5a1491a6-f5be-d75e-25bd-6650c69a7182>
Line 91, Column 14 Tab Size: 4 Plain Text
```

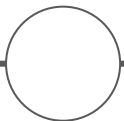
HTTP HEADER BLOCK

HTTP PAYLOAD BLOCK

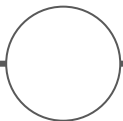




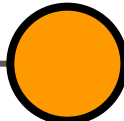
WARC Creation



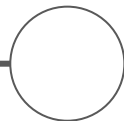
HTTP Header & Payload Extraction



Push to IPFS



Generate CDXJ Record



WARC-CDXJ correspondence

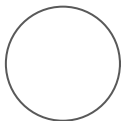
HEADER DIGEST

QmcN9eWwRF73dZj5BgT4x8jeEcFrurX1hot8QwCb
Mi9PB

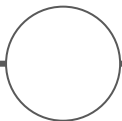
PAYLOAD DIGEST

Qmczh9YnB4U1ptPeqxcaTZA4aMmuNUswTLTWzX
ntvbp9sL

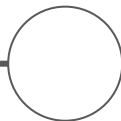
```
ipwb.example.com)/ 20160905022013 {  
  "locator": "urn:ipfs/QmcN9eWwRF73dZj5BgT4x8jeEcFrurX1hot8QwCbMi9PB/Qmczh9YnB4U1ptPeqxca  
TZA4aMmuNUswTLTWzXntvbp9sL",  
  "mime_type": "text/html",  
  "status_code": 200,  
  "other_fields": "other values..."  
}
```



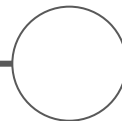
WARC Creation



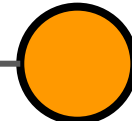
HTTP Header & Payload Extraction



Push to IPFS



Generate CDXJ

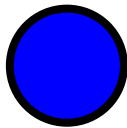


WARC-CDXJ correspondence

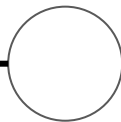
```
20160907003819654.warc
20160907003819654.warc
52 WARC/1.0
53 WARC-Type: response
54 WARC-Target-URI: http://ipwb.example.com/
55 WARC-Date: 2016-09-07T00:38:19Z
56 WARC-Record-ID: <urn:uuid:1e3907a9-2e5c-9981-6a92-964a465d998e>
57 Content-Type: application/http; msgtype=response
58 Content-Length: 800
59
60 HTTP/1.1 200 OK
61 Host: ipwb.example.com
62 Connection: close
63 Content-Type: text/html; charset=UTF-8
64 Content-Length: 666
65
66 <html><head>
67 <title>InterPlanetary Wayback</title>
68 <link rel="stylesheet" type="text/css" href="style.css">
69 </head>
70 <body>
71 <h1>This is site for Space Dog</h1>
72 
73 <p>InterPlanetary Wayback (ipwb) facilitates permanence and collaboration
74
75 </body></html>
76
77
78 WARC/1.0
79 WARC-Type: response
80 WARC-Target-URI: http://ipwb.example.com/style.css
81 WARC-Date: 2016-09-07T00:38:19Z
82 WARC-Record-ID: <urn:uuid:2502e65a-b0bd-70c6-8799-8687029a71f4>
83 Content-Type: application/http; msgtype=response
84 Content-Length: 144
85
86 HTTP/1.1 200 OK
87 Host: ipwb.example.com
88 Connection: close
89 Content-Type: text/css; charset=UTF-8
90 Content-Length: 19
91
92
93 img {width: 250px;}
94
95
96
```

```
ipwb.example.com)/ 20160905022013 {"locator":
"urn:ipfs/QmcN9eWwRF73dZj5BgT4x8jeEcFrXurX1hot8QwCbMi9PB/
Qmczh9YnB4H1ptPeqxcaTZA4aMmuNUswTLTWzXntvbp9sL",
"mime_type": "text/html", "status_code": "200"}
ipwb.example.com)/style.css 20160905022013 {"locator":
"urn:ipfs/QmU1k71bT6ibZBSdxBL35cQXwovTih8cTB4CXfrjyMfZxE/Q
mbvUAo9U31wSdvARjvbPeVBTawCjN1kvPhQ4ho3n8TAZo",
"mime_type": "text/css", "status_code": "200"}
ipwb.example.com)/ipwb.png 20160905022013 {"locator":
"urn:ipfs/QmTjFmXFGvbP4nwFoq3tNYDPW6gC99i5njqrsXSw6QRvHa/
QmYMKZbnk53kuPJirahJHGeVCCy2afLyePRdX38TukFUwd",
"mime_type": "image/png", "status_code": "200"}
ipwb.example.com)/fileduration.png 20160905022013 {"locator":
"urn:ipfs/QmaCj6LNngxwqxaLmfp1xCyxcwDt2Uzqf8gCG6bVyQppYC/
QmdgtMcGprTF8bqv7ytgMwtoi5BhRxfuvBjD6Vj2U7ohz1",
"mime_type": "image/png", "status_code": "200"}
ipwb.example.com)/filesize.png 20160905022013 {"locator":
"urn:ipfs/QmNPJrSVY31oGDooMiA18ZDNHfKLnEg3j5gRj1dFdrqmS4/
Qmb4heB8PU58nkWt6w5tBgMfpeLTkuU7iuxg9tFdoPsF1B",
"mime_type": "image/png", "status_code": "200"}
```

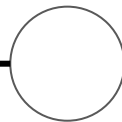
Replay



Replay reference
via CDXJ



Dereference via IPFS



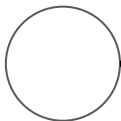
Reconstruction from IPFS

```
ipwb.example.com)/ 20160905022013 {"locator":  
  "urn:ipfs/QmcN9eWwRF73dZj5BgT4x8jeEcFrurX1  
  hot8QwCbMi9PB/Qmczh9YnB4U1ptPeqxcaTZA4a  
  MmuNUswTLTWzXntvbp9sL", "mime_type":  
  "text/html", "status_code": "200"}  
ipwb.example.com)/style.css 20160905022013  
{"locator":  
  "urn:ipfs/QmU1k71bT6ibZBSdxBL35cQXwovTih8cT  
  B4CXfrjyMfZxE/QmbvUAo9U31wSdvARjvbPeVBTA  
  wCjN1kyPhQ4ho3n8TAZo", "mime_type": "text/css",  
  "status_code": "200"}  
ipwb.example.com)/ipwb.png 20160905022013  
{"locator":  
  "urn:ipfs/QmTjfMxFGvbP4nwFoq3tNYDPW6gC99i5  
  njrqsXSw6QRvHa/QmYMKZbnk53kuPJirahJHGevC  
  Cy2afLyePRdX38TukFUwd", "mime_type":  
  "image/png", "status_code": "200"}  
...
```

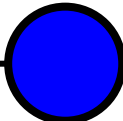
<http://ipwb.example.com>

SEP

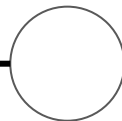
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31			



Replay reference
via CDXJ



Dereference via IPFS



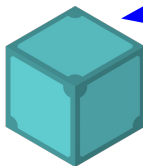
Reconstruction from IPFS

ipwb.example.com)/ 20160905022013 {"locator": "urn:ipfs/

QmcN9eWwRF73dZj5BgT4x8jeEcFrXurX1hot8QwCbMi9PB/Qmczh9YnB4U1ptPeqxcaTZA4aMmuNUswTLTWzXntvbp9sL",

"mime_type": "text/html", "status_code": "200"}

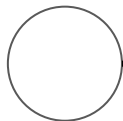
...



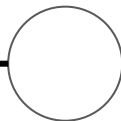
IPFS

HTTP HEADER BLOCK

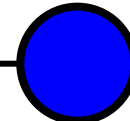
HTTP PAYLOAD BLOCK



Replay reference
via CDXJ



Dereference via IPFS

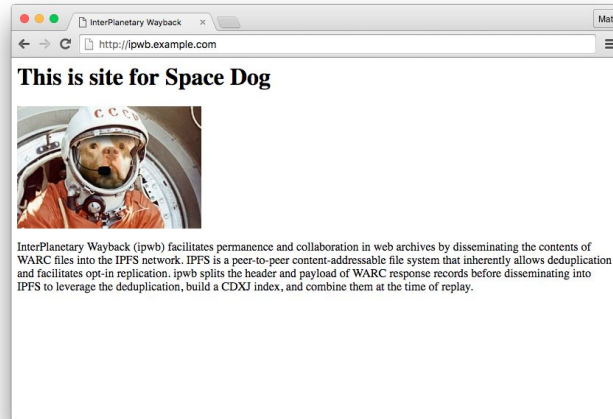


Reconstruction from IPFS

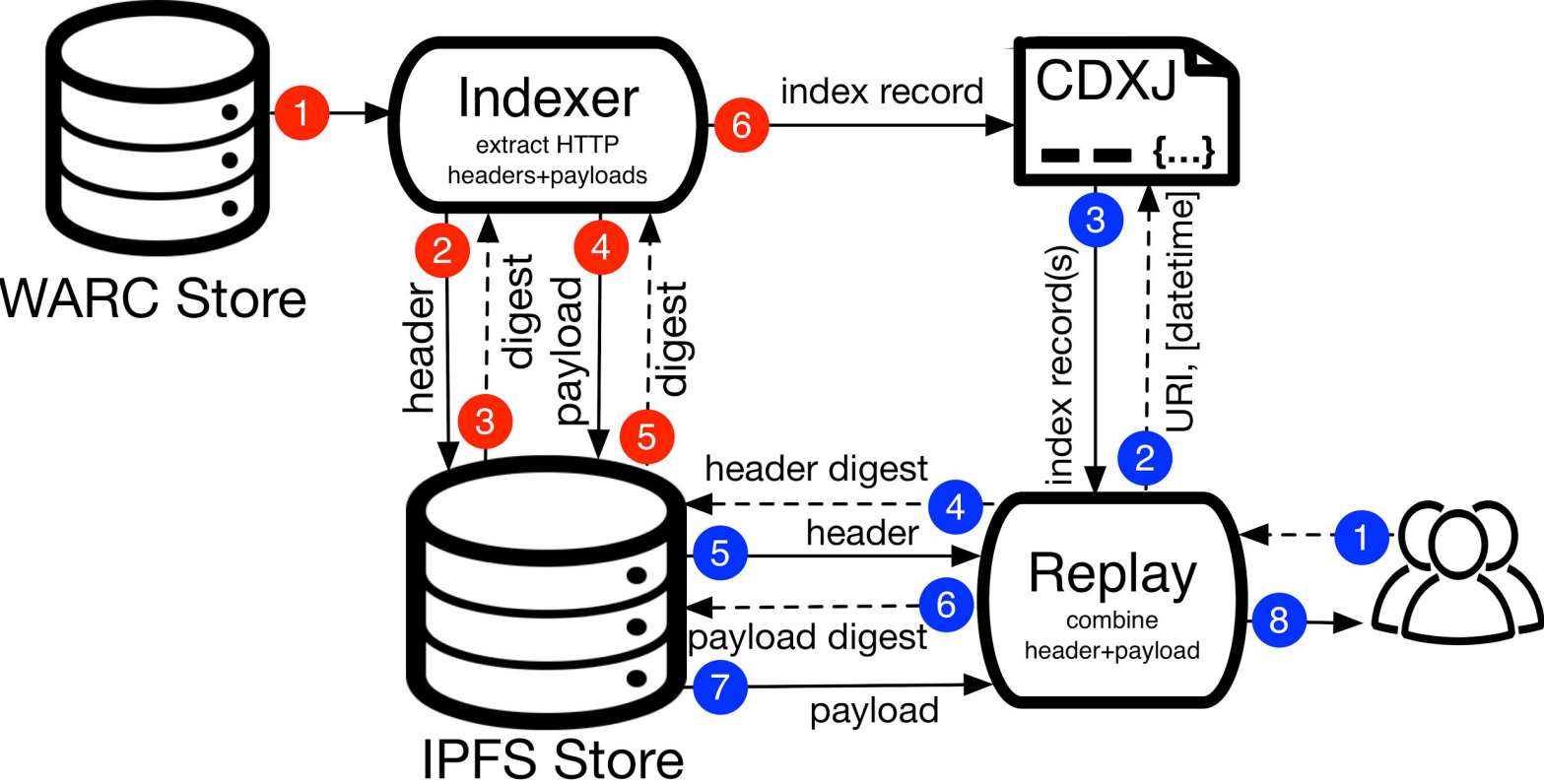
HTTP HEADER BLOCK

HTTP PAYLOAD BLOCK

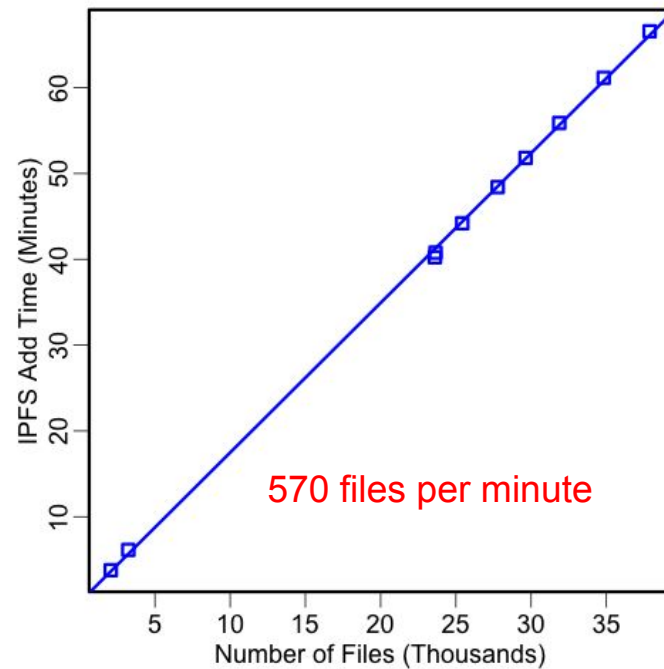
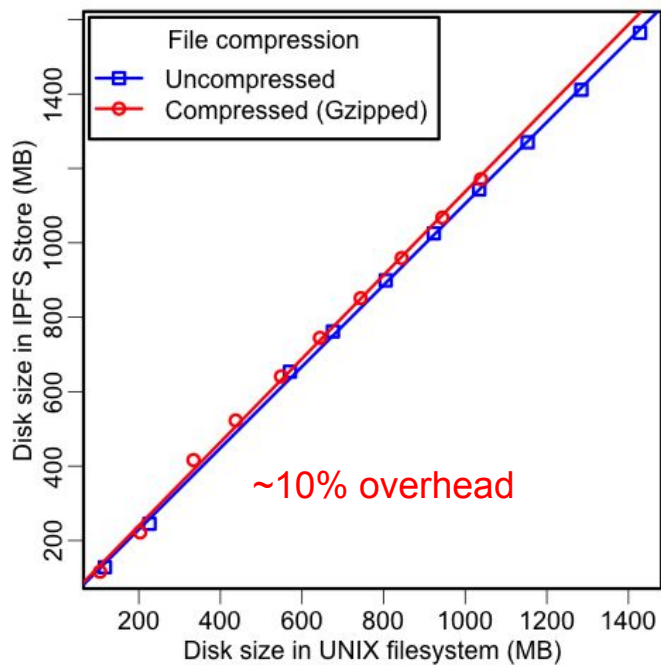
Reconstruct



Data Flow



Evaluation



- Reported IPFS slowness <https://github.com/ipfs/go-ipfs/issues/1216>
 - Has since been fixed, subsequent to IPWB-TPDL

Replay Time

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

Future Works

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

Conclusions

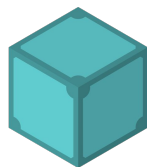
- A proof of concept system to leverage a novel approach to archiving and retrieval
- Evaluated storage and time costs and qualitative analysis
- It can only work for small archives in it's current state
- A path to answer “**who will archive the archives?**” question



InterPlanetary Wayback

Peer-to-Peer Permanence of Web Archives

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



IPFS + WayBack

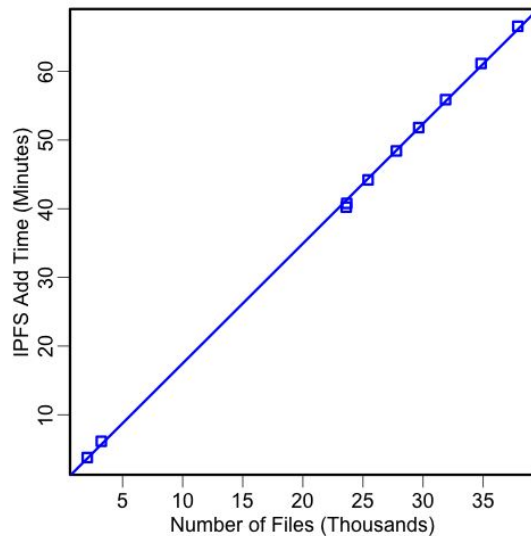
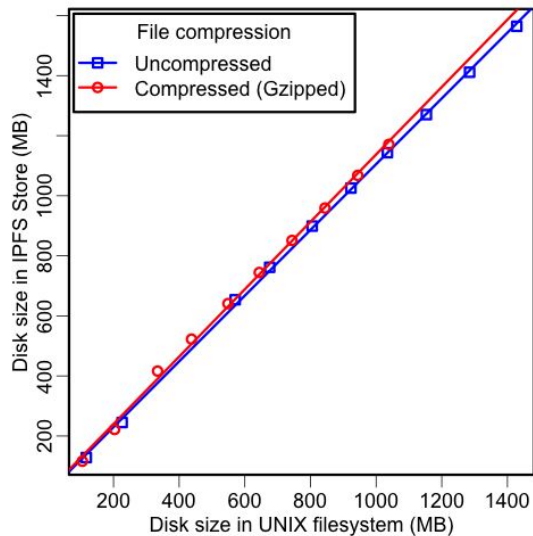
<http://github.com/oduwsdl/ipwb>

@WebSciDL

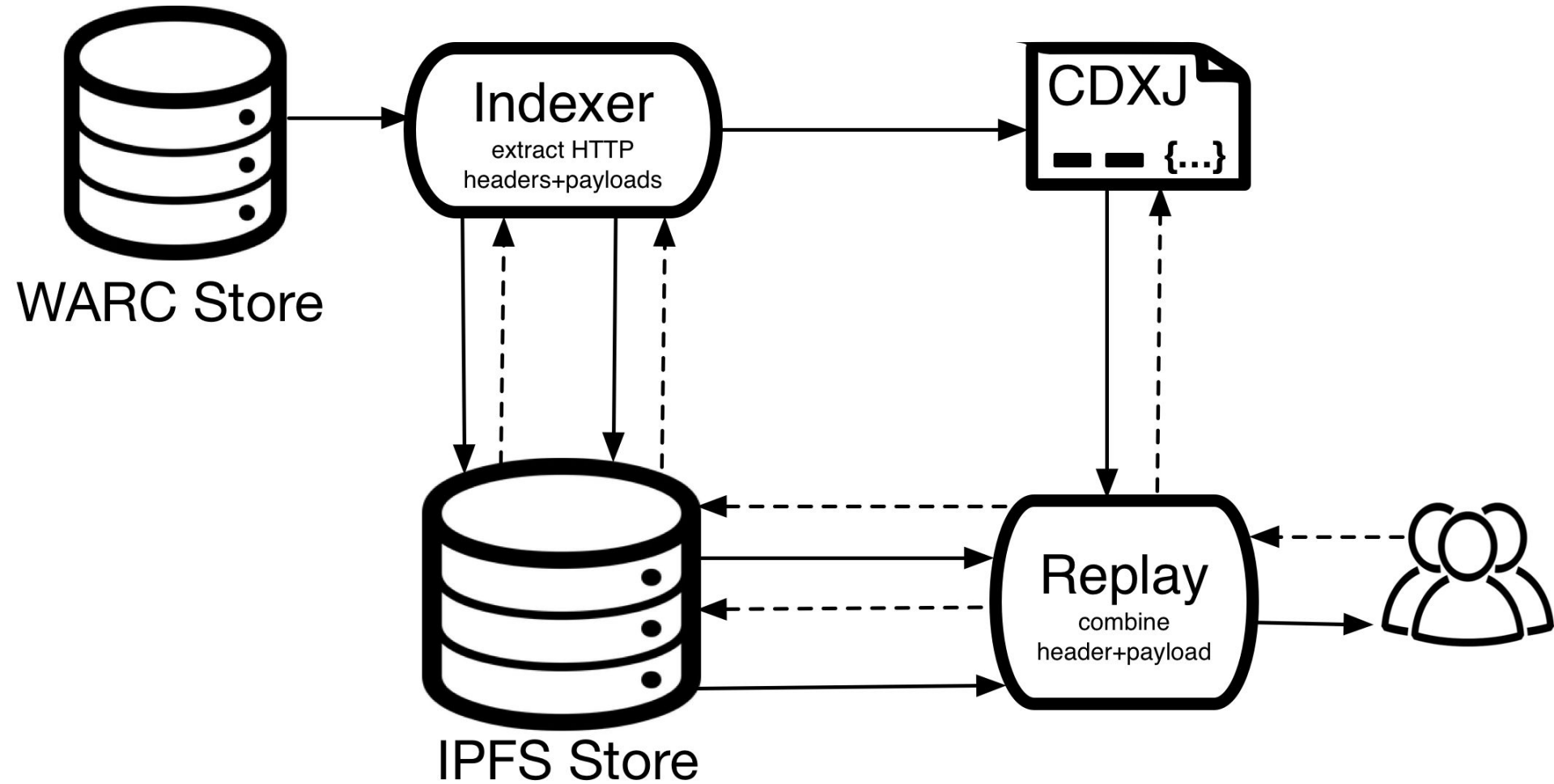
Support: NSF #1624067 via the Archives Unleashed Hackathon

Backup Slides

Evaluation



- Reported IPFS slowness <https://github.com/ipfs/go-ipfs/issues/1216>
 - Has since been fixed, subsequent to IPWB-TPDL





WARC Store



IPFS Store



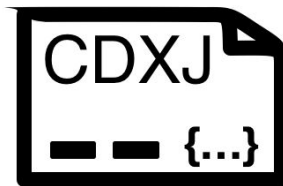


1



Indexer

extract HTTP
headers+payloads



WARC Store



IPFS Store





WARC Store

1



2

header

digest

4

payload

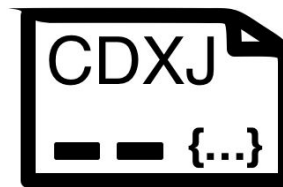
digest

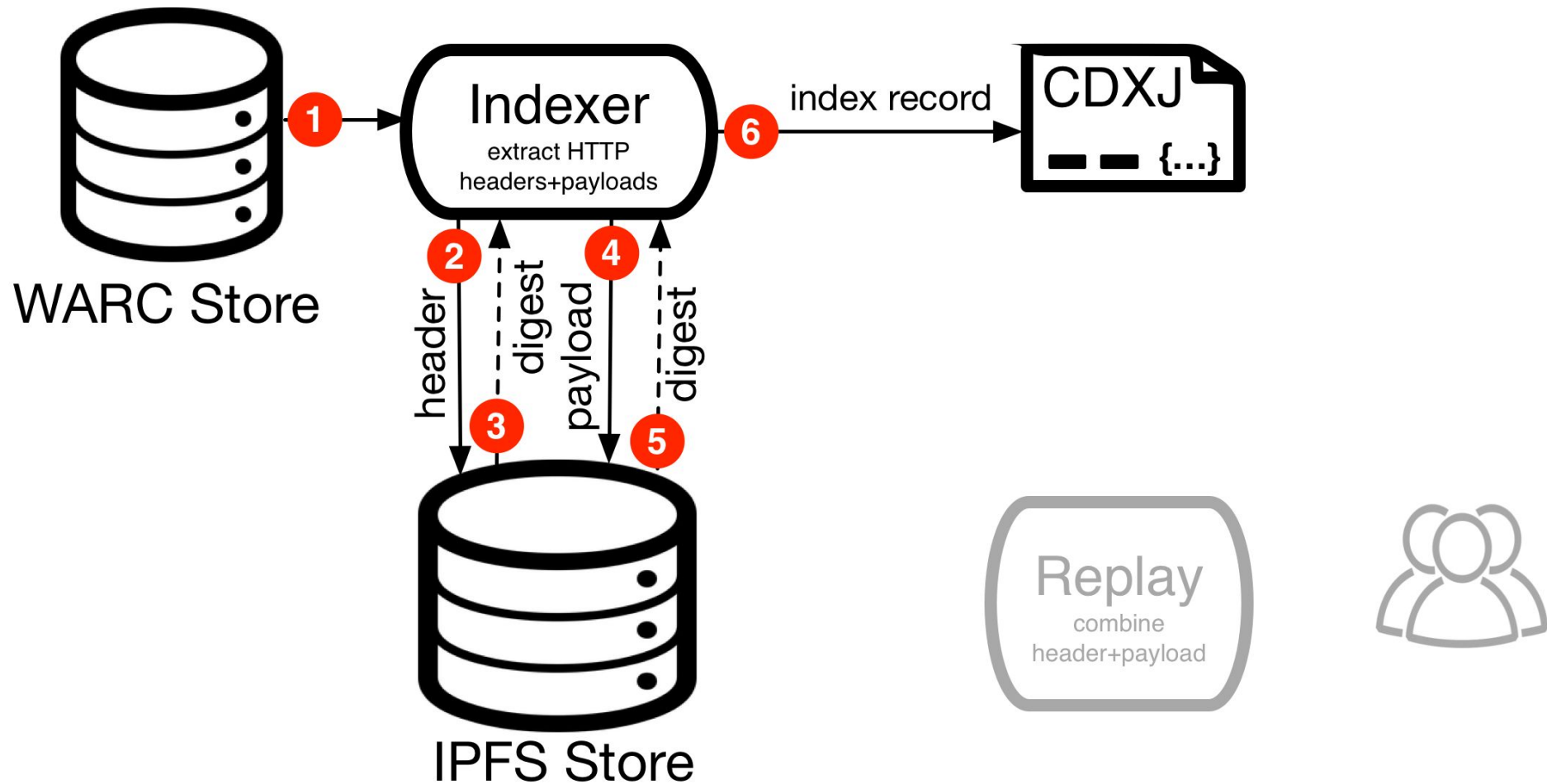
3

5



IPFS Store



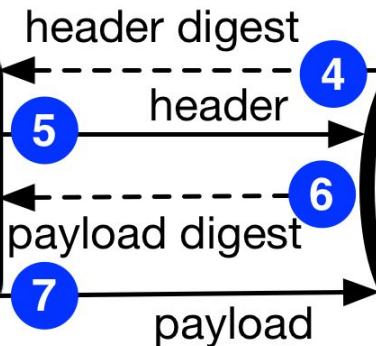
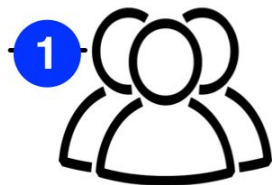
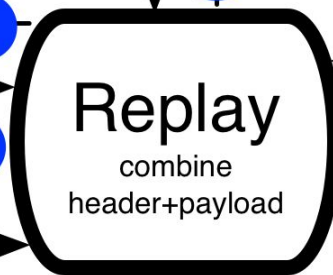




WARC Store



IPFS Store

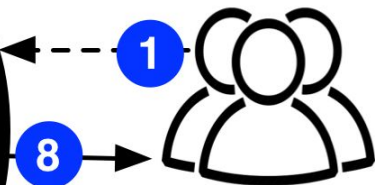
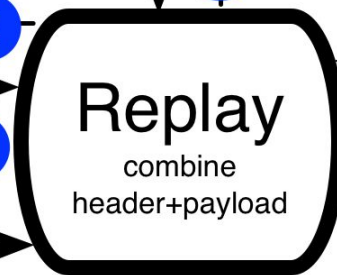




WARC Store



IPFS Store

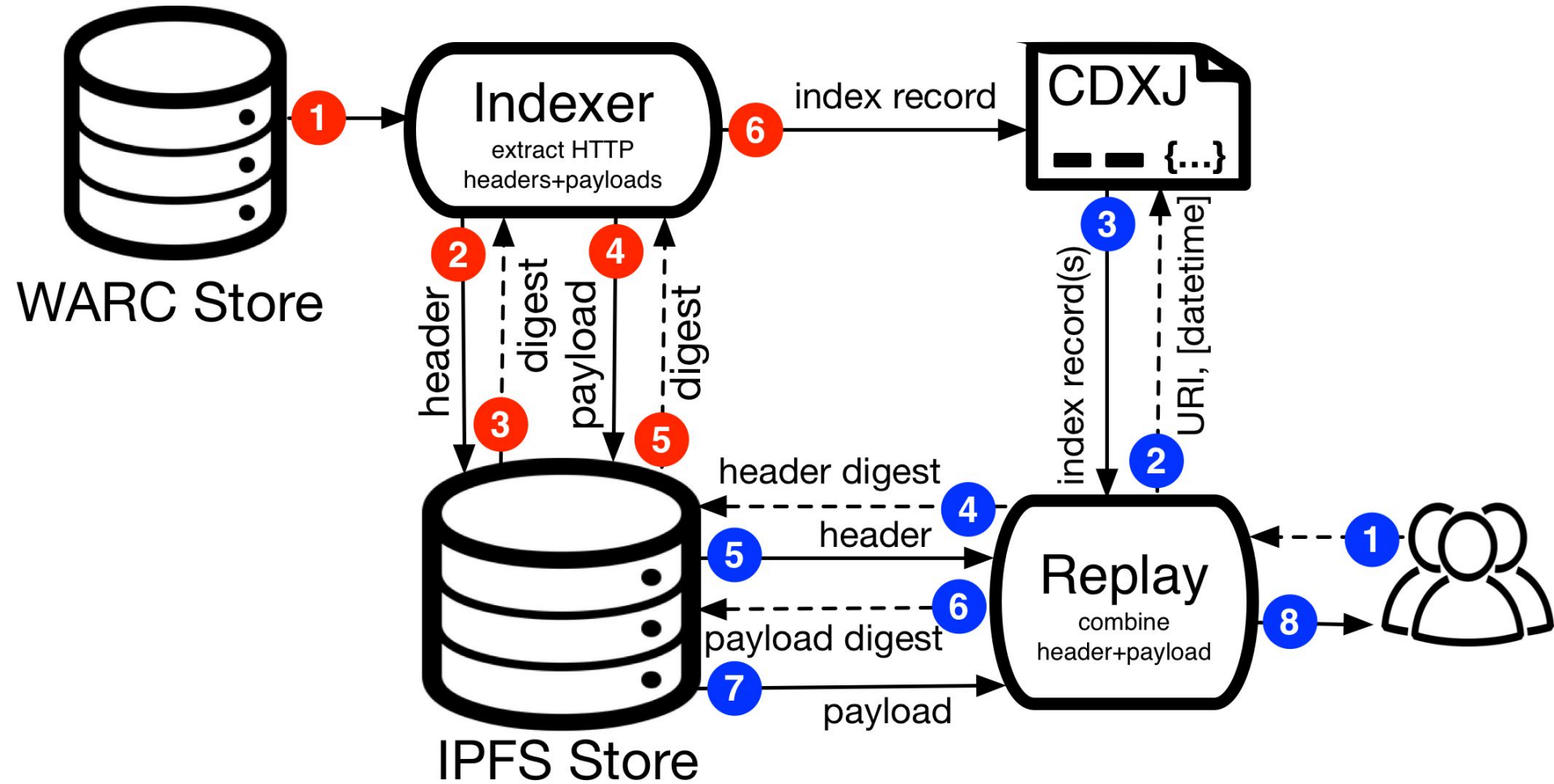


header digest

header

payload digest

payload



Methodology - IPWB WARC indexing

- `warc-response` record body extracted into temp files
 - HTTP header and entity body (payload) separated
 - Response metadata (e.g., datetime) retained
- temp files pushed into IPFS via locally running daemon
 - Two IPFS hashes (for header and payload) returned
- CDXJ record created representing `warc-response` contents
 - Contains URI-R, archived HTTP status, encoded IPFS hashes

Methodology - Replaying Archives

- Extension of pywb API to read CDXJ files
- On encountering IPFS URN, fetch `warc-response` temp files from IPFS using local daemon
 - This may occur on a separate machine using a separate daemon
- With WARC contents fetched, replay contents using pywb where the locator value in the CDXJ is used to dereference the temp files pulled from IPFS

CDXJ in IPWB

```
1  SURT_URI  DATETIME {  
2      "id": "WARC-Record-ID",  
3      "url": "ORIGINAL_URI",  
4      "status": "3-DIGIT_HTTP_STATUS",  
5      "mime": "Content-Type",  
6      "locator": "urn:ipfs/HEADER_DIGEST/PAYLOAD_DIGEST"  
7  }
```