

The Web ARChive (WARC) File Format

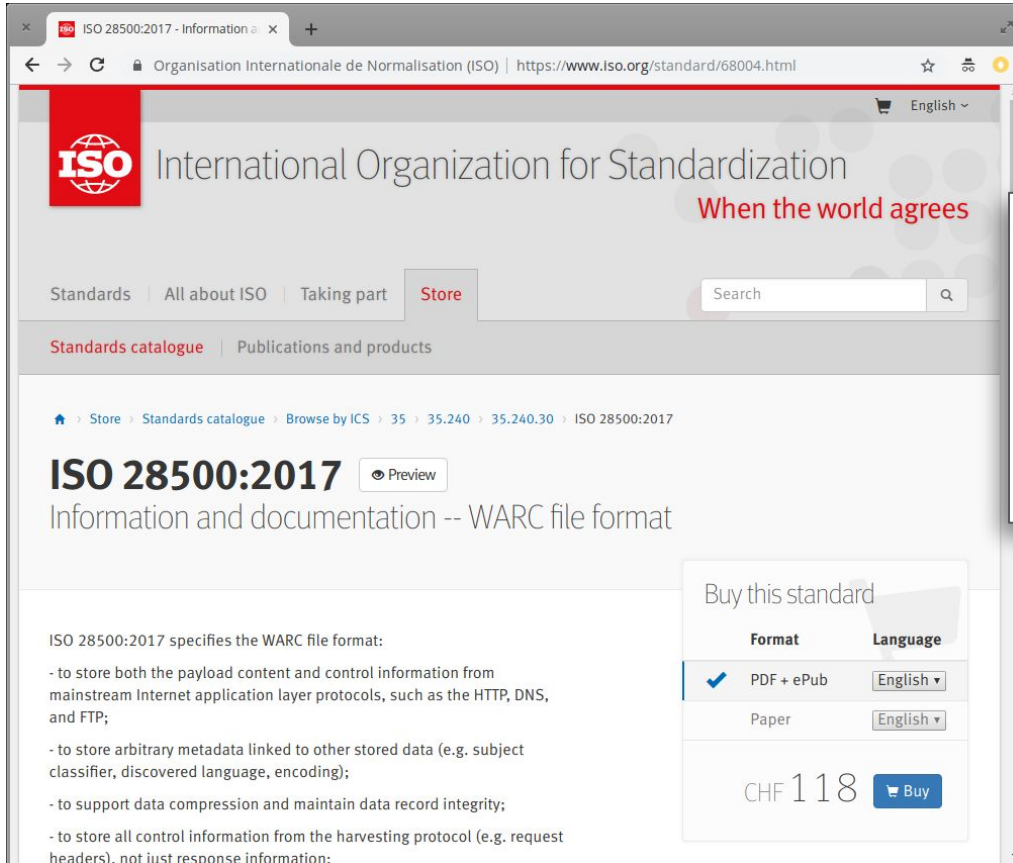
Sawood Alam

Web Science and Digital Libraries Research Group
Old Dominion University
Norfolk, Virginia, USA
@ibnesayeed

CS 531 Web Server Design
November 28, 2018



Web ARChive (WARC): ISO 28500 File Format



The screenshot shows the ISO 28500:2017 information page. The header includes the ISO logo and the text "International Organization for Standardization" and "When the world agrees". The main content area features the title "ISO 28500:2017" with a "Preview" button. Below the title, it says "Information and documentation -- WARC file format". A "Buy this standard" section is visible on the right, showing options for "Format" (PDF + ePub, Paper) and "Language" (English), with a price of CHF 118 and a "Buy" button. The left sidebar contains navigation links like "Standards", "All about ISO", "Taking part", and "Store".

ISO 28500:2017 specifies the WARC file format:

- to store both the payload content and control information from mainstream Internet application layer protocols, such as the HTTP, DNS, and FTP;
- to store arbitrary metadata linked to other stored data (e.g. subject classifier, discovered language, encoding);
- to support data compression and maintain data record integrity;
- to store all control information from the harvesting protocol (e.g. request headers). not just response information;

<https://twitter.com/AcademicsSay/status/1065299702881705985>



Shit Academics Say
@AcademicsSay



The best things in academic life are-

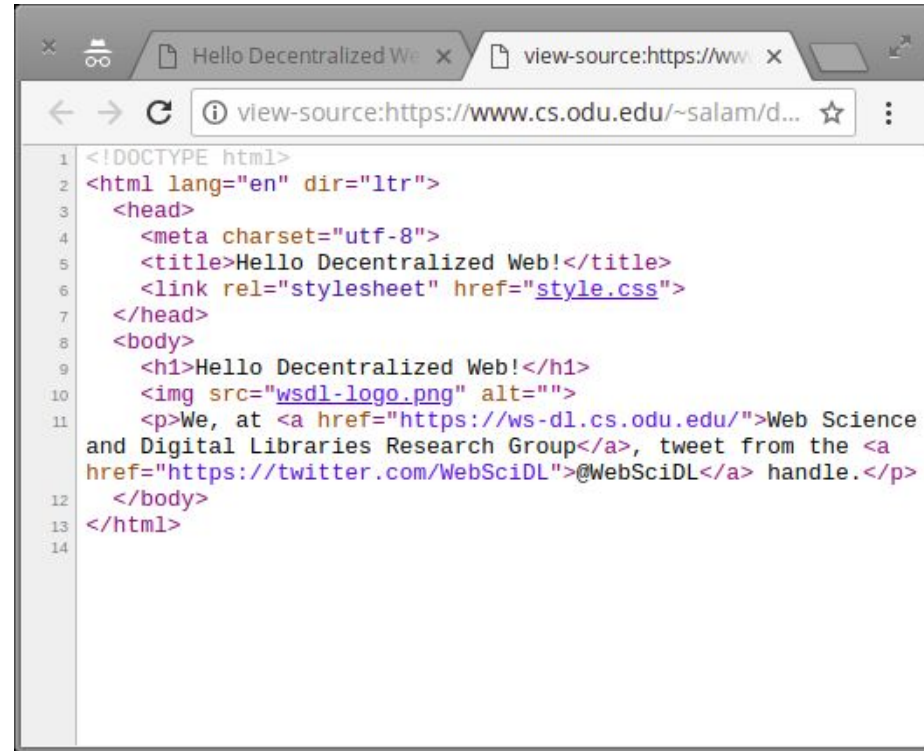
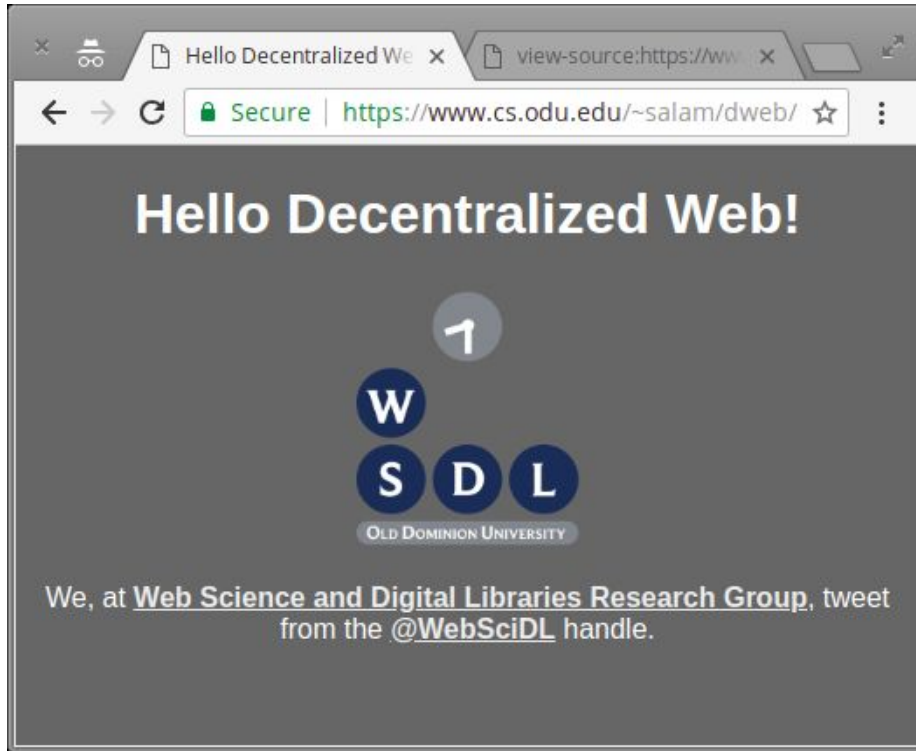
[To purchase full text click below]

♥ 3,543 12:43 PM - Nov 21, 2018



<https://github.com/iipc/warc-specifications>

Rendered HTML vs. Source Code



HTTP Response vs. WARC Record

The diagram compares an HTTP response (left) with a WARC record (right). Brackets indicate the following correspondences:

- WARC headers** (green bracket):
 - 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 headers** (orange bracket):
 - 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
- Payload** (blue bracket):
 - HTML content: <!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><p>We, at Web Science a nd Digital Libraries Research Group, tweet from the @WebSciDL handle.</p></body></html>

The WARC record is shown in a terminal window titled "less hello-dweb.warc 63x33". The HTTP response is shown in a terminal window titled "curl -i https://www.cs.odu.edu/~salam/dweb/ 63x25".

Why WARC and not Plain Filesystem?

- Number of inodes
- Name collision
- Deduplication
- Rich metadata
- Optimized for long-term Web preservation

WARC Record Types

- ★ warcinfo
- ★ response
- ★ resource
- ★ request
- ★ metadata
- ★ revisit
- ★ conversion
- ★ continuation

```
WARC-Type    = "WARC-Type" ":" record-type
record-type  = "warcinfo" | "response" | "resource"
               | "request" | "metadata" | "revisit"
               | "conversion" | "continuation"
```

<http://iipc.github.io/warc-specifications/specifications/warc-format/warc-1.1/>

WARC Indexing

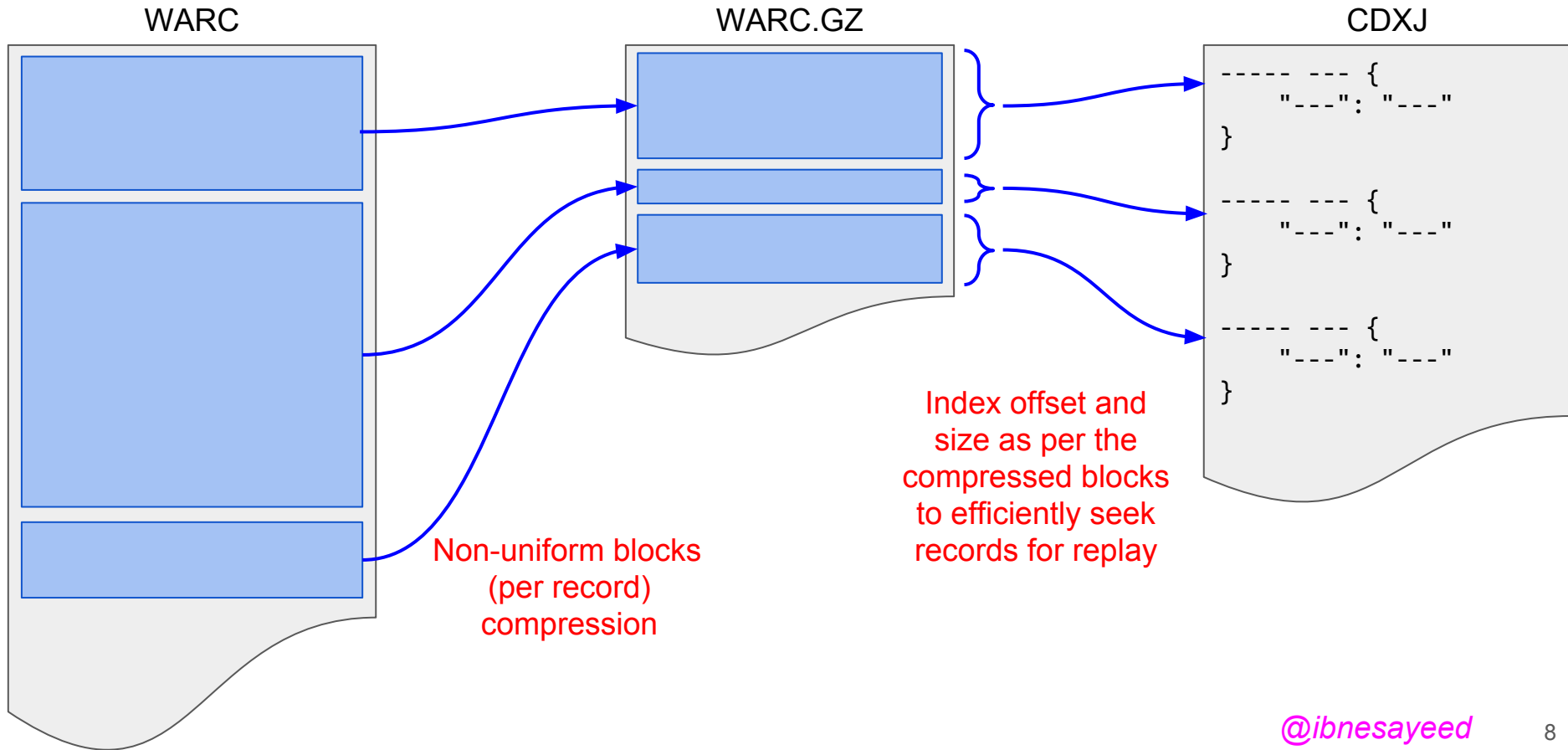
hello-dweb.warc

```
1 WARC/1.0
2 WARC-Type: response
3 WARC-Target-URI: https://www.cs.odu.edu/~salam/dweb/
4 WARC-Date: 2018-08-02T15:12:34Z
5 WARC-Record-ID: <urn:uuid:6e820124-b8a3-02ee-4012-f9be5578e09a>
6 Content-Type: application/http; msgtype=response
7 Content-Length: 654
8
9 HTTP/1.1 200 OK
10 Server: nginx
11 Date: Thu, 02 Aug 2018 15:06:54 GMT
12 Content-Type: text/html
13 Transfer-Encoding: chunked
14 Connection: keep-alive
15 Vary: Accept-Encoding
16 Front-End-Https: on
17
18 <!DOCTYPE html>
19 <html lang="en" dir="ltr">
20 <head>
21 <meta charset="utf-8">
22 <title>Hello Decentralized Web</title>
23 <link rel="stylesheet" href="style.css">
24 </head>
25 <body>
26 <h1>Hello Decentralized Web</h1>
27 
28 <p>Welcome, at cs href="https://www.cs.odu.edu/~Web Science and Digital
29 </body>
30 </html>
31
32 WARC/1.0
33 WARC-Type: response
34 WARC-Target-URI: https://www.cs.odu.edu/~salam/dweb/style.css
35 WARC-Date: 2018-08-02T15:12:34Z
36 WARC-Record-ID: <urn:uuid:dc4faeb2-6b8d-4200-a007-95c60b63f504>
37 Content-Type: application/http; msgtype=response
38 Content-Length: 421
39
40 HTTP/1.1 200 OK
41 Server: nginx
42 Date: Thu, 02 Aug 2018 15:12:34 GMT
43 Front-End-Https: on
44 Content-Type: text/css
45 Content-Length: 101
46 Last-Modified: Thu, 02 Aug 2018 15:06:10 GMT
47 ETag: "a1-37275269dc0"
48 Accept-Ranges: bytes
49 Vary: Accept-Encoding
50
51 html {
52   background: #6060;
53   text-align: center;
54   color: #fff;
55   font-family: sans-serif;
56 }
57
58 img {
59   width: 150px;
60 }
61
62 a {
63   font-weight: bold;
64   color: #00a0;
65 }
```

```
edu,odu,cs)/~salam/dweb/ 20180802012013 {
  "status_code": 200,
  "mime_type": "text/html",
  "offset": 0,
  "size": 998,
  "warc_file": "hello-dweb.warc"
}
```

```
edu,odu,cs)/~salam/dweb/style.css 20180802012013 {
  "status_code": 200,
  "mime_type": "text/css",
  "offset": 1001,
  "size": 771,
  "warc_file": "hello-dweb.warc"
}
```

WARC Compression



WARC Tools

- **Heritrix**: Web crawler
 - <https://github.com/internetarchive/heritrix3>
- **Wget**: Downloader CLI
 - <https://www.gnu.org/software/wget/>
- **Squidwarc**: Browser-based Web crawler
 - <https://github.com/N0taN3rd/Squidwarc>
- **WARCreate**: Chrome Extension to create WARC
 - <https://warcreate.com/>
- **Warcprox**: WARC writing MITM HTTP/S proxy
 - <https://github.com/internetarchive/warcprox>
- **warcio**: Python library to read/write WARC
 - <https://github.com/webrecorder/warcio>
- **Open Wayback**: Web archival replay system (Java)
 - <https://github.com/iipc/openwayback>
- **PyWB**: Web archival replay system (Python)
 - <https://github.com/webrecorder/pywb>
- **InterPlanetary Wayback (IPWB)**: Web archival replay system using IPFS
 - <https://github.com/oduwsdl/ipwb>
- **WAIL**: Web Archiving Integration Layer
 - <https://matkelly.com/wail>

WARC with Wget

Wget has built-in support for WARC creation, indexing, compression, and deduplication

```
$ man wget | grep "\-warc"  
--warc-file=file  
--warc-header=string  
--warc-max-size=size  
--warc-cdx  
--warc-dedup=file  
--no-warc-compression  
--no-warc-digests  
--no-warc-keep-log  
--warc-tempdir=dir
```

<https://www.gnu.org/software/wget/manual/wget.html>

WARC with WARCreate



WARC with warcio

Write a WARC file

```
from warcio.capture_http import capture_http
import requests

with capture_http('example.warc.gz'):
    requests.get('https://example.com/')
```

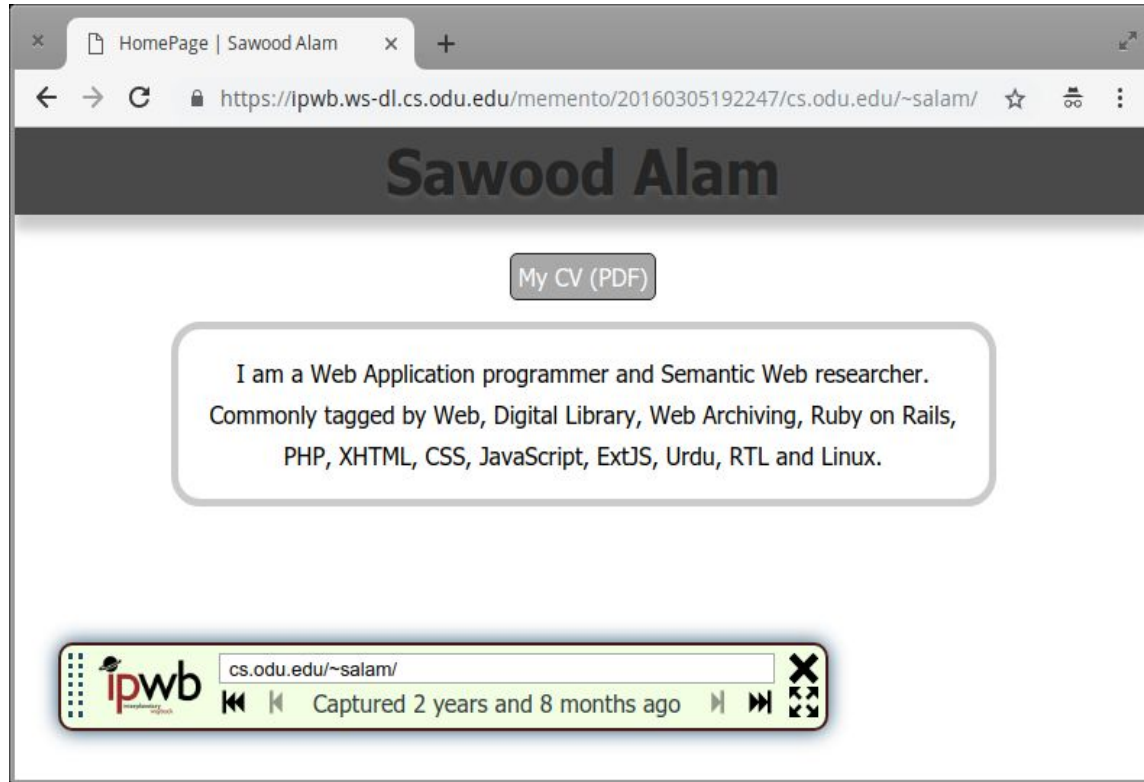
Read from a WARC file

```
from warcio.archiveiterator import ArchiveIterator

with open('example.warc.gz', 'rb') as stream:
    for record in ArchiveIterator(stream):
        if record.rec_type == 'response':
            print(record.rec_headers.get_header('WARC-Target-URI'))
```

WARC with IPWB

```
$ ipwb index salam.warc.gz | ipwb replay
```



WebPackage: Similar, but not the same!

- Package a group of related HTTP requests and responses to transmit and store together
- Optionally sign messages to allow third parties to store and deliver asynchronously
- Make browsers verify signed packages using origins' valid certificates
- Differences from WARC
 - Binary instead of textual
 - Not suitable for long-term preservation due to signing that would eventually expire

<https://github.com/WICG/webpackage>

Conclusions

- Web ARChive (WARC) is a well-supported and evolving ISO standard data format
- It is a text-based HTTP Message-like wrapper format
- It can store arbitrary number of HTTP request/response messages (and various other data types) along with a rich set of metadata
- Optimized for long-term Web preservation

<https://github.com/iipc/warc-specifications>