WARC Files

David Bui

WARC File Format

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/

@ibnesayeed 6

- Based on Internet Archive's ARC File Format. Hence the name Web ARChive.
- The WARC file consists of a concatenation of one or more WARC records.
- There are 8 types of WARC records seen as seen to the left.
- A WARC record consists of
  - The header
  - Then Record content block
- The header has mandatory named fields
  - Date
  - Type
  - Length of the record
  - Plus, other fields that assist in retrieval
- The content block contains resources in any format such as images or audio

Reference: https://www.loc.gov/preservation/digital/formats/fdd/fdd000236.shtml#:~:text=The%20WARC%20format%20is%20a,from%20the%20World%20Wide%20Web.&text=A%20WARC%20format%20file%20is,one%20or%20more%20WARC%20records.

```
WARC/1.0
WARC-Type: warcinfo
WARC-Date: 2011-02-25T18:32:19Z
WARC-Filename: WIDE-20110225183219005-04371-13730~crawl301.us.archive.org~9443.warc.gz
WARC-Record-ID: <urn:uuid:88fbcbee-f24e-47c1-b0c4-f7a9530ceb74>
Content-Type: application/warc-fields
Content-Length: 442

software: Heritrix/3.0.1-SNAPSHOT-20110127.213729 http://crawler.archive.org
ip: 207.241.232.79
hostname: crawl301.us.archive.org
format: WARC File Format 1.0
conformsTo: http://bibnum.bnf.fr/WARC/WARC_ISO_28500_version1_latestdraft.pdf
operator: kenji@archive.org
isPartOf: wide
description: seeds.txt
robots: obey
http-header-user-agent: Mozilla/5.0 (compatible; archive.org_bot +http://www.archive.org/details/archive.org_bot)
```

Warcinfo record

Response record with html content

CDX (Internet Archive)

- Consists of individual lines of text that each summarize a web document
- Starts with a CDX legend that describes how each line of data is formatted.
- Used to create index files of Warc filef

CDX A b e a m s c k r V v D d g M n
0-0-0checkmate.com/Bugs/Bug_Investigators.html 20010424210551 209.52.183.152 0-0-0checkmate.com:80/Bugs/Bug_Investigators.html text/html 200
58670fbe7432c5bed6f3dcd7ea32b221 a725a64ad6bb7112c55ed26c9e4cef63 - 17130110 59129865 1927657 6501523 DE_crawl6.20010424210458 - 5750

0-0-0checkmate.com/Bugs/Insect_Habitats.html 20010424210312 209.52.183.152 0-0-0checkmate.com:80/Bugs/Insect_Habitats.html text/html 200
d520038e97d7538855715ddcba613d41 30025030eeb72e9345cc2ddf8b5ff218 - 47392928 145482381 4426829 15345336 DE_crawl3.20010424210104 - 6356

0-0-0checkmate.com/Hot/index.html 20010424212403 209.52.183.152 0-0-0checkmate.com:80/Hot/index.html text/html 200
52242643710547ff4ce2605ed03ed9e2 b06d037c06e7ffd7afc6db270aca7645 - 21301376 62305547 1855363 6627262 DE_crawl6.20010424212307 - 6317

```
 CDX N b a m s k r M S V g
10,100,196,202)/musewebmain/dzyy/default.asp?ntmpkzh=236900 20110225232158 http://202.196.100.10/musewebmain/dzyy/default.asp?nTmpKzh=236900 text/html 200
4CVVD6FMLZCW73EVBVHVKBU4SPPOI3OE - - 587 824729713 testWARCfiles/WIDE-20110225221304846-04388-13730~crawl301.us.archive.org~9443.warc.gz
10,100,196,202)/musewebmain/dzyy/default.asp?ntmpkzh=265878 20110225190300 http://202.196.100.10/musewebmain/dzyy/default.asp?nTmpKzh=265878 text/html 200
4CVVD6FMLZCW73EVBVHVKBU4SPPOI3OE - - 587 231703167 testWARCfiles/WIDE-20110225184020081-04372-13730~crawl301.us.archive.org~9443.warc.gz
10,100,196,202)/musewebmain/dzyy/default.asp?ntmpkzh=265880 20110225190427 http://202.196.100.10/musewebmain/dzyy/default.asp?nTmpKzh=265880 text/html 200
4CVVD6FMLZCW73EVBVHVKBU4SPPOI3OE - - 587 243878729 testWARCfiles/WIDE-20110225184020081-04372-13730~crawl301.us.archive.org~9443.warc.gz
10,100,196,202)/musewebmain/dzyy/default.asp?ntmpkzh=266525 20110225232344 http://202.196.100.10/musewebmain/dzyy/default.asp?nTmpKzh=266525 text/html 200
4CVVD6FMLZCW73EVBVHVKBU4SPPOI3OE - - 586 987899236 testWARCfiles/WIDE-20110225220702321-04387-13730~crawl301.us.archive.org~9443.warc.gz
10,100,196,202)/musewebmain/dzyy/default.asp?ntmpkzh=266527 20110225232226 http://202.196.100.10/musewebmain/dzyy/default.asp?nTmpKzh=266527 text/html 200
4CVVD6FMLZCW73EVBVHVKBU4SPPOI3OE - - 586 828862927 testWARCfiles/WIDE-20110225221304846-04388-13730~crawl301.us.archive.org~9443.warc.gz
10,100,196,202)/musewebmain/dzyy/default.asp?ntmpkzh=600181887 20110225224003 http://202.196.100.10/musewebmain/dzyy/default.asp?nTmpKzh=600181887 text/
html 200 4CVVD6FMLZCW73EVBVHVKBU4SPPOI3OE - - 590 463925916 testWARCfiles/WIDE-20110225215415804-04385-13730~crawl301.us.archive.org~9443.warc.gz
```

Reference:
http://web.archive.org/web/20031226073353/http://www.archive.org/web/researcher/cdx_file_format.php

```
A canonized url
B news group
C rulespace category ***
D compressed dat file offset
F canonized frame
G multi-column language description (* soon)
H canonized host
I canonized image
J canonized jump point
K Some weird FBIS what's changed kinda thing
L canonized link
M meta tags (AIF) *
N massaged url
P canonized path
Q language string
R canonized redirect
S compressed record size
U uniqueness ***
V compressed arc file offset *
X canonized url in other href tags
Y canonized url in other src tags
Z canonized url found in script
a original url **
b date **
c old style checksum *
d uncompressed dat file offset
e IP **
f frame *
g file name
h original host
i image *
j original jump point
k new style checksum *
l link *
m mime type of original document *
n arc document length *
o port
p original path
r redirect *
s response code *
t title *
v uncompressed arc file offset *
x url in other href tages *
y url in other src tags *
z url found in script *
# comment
```

CDX legend (legend)

CDXJ

- CDX but with the data inside a JSON
- Internet archive specifies that CDXJ begins with a header specifying version: !OpenWayback-CDXJ 1.0
- Followed by the record indexes broken up into 4 parts
    1. Searchable URI
    2. Timestamp,
    3. Record Type,
    4. JSON block
- Below is a WebRecorder generated CDXJ file. Note that it is missing an initial header and a record type field.


```
1  de,kaze-online,shop)/images/products/small/pb0281.jpg 20110225183218 {"url": "http://shop.kaze-online.de/images/products/small/PB0281.jpg", "mime": "image/
   jpeg", "status": "200", "digest": "QG6N6SOXUHFK2BGT6EEGMNQMALUW5YAE", "length": "3418", "offset": "492", "filename": "example.warc.gz"}
2  de,kaze-online,shop)/images/products/small/pb0281.jpg 20110225183218 {"url": "http://shop.kaze-online.de/images/products/small/PB0281.jpg", "mime":
   "application/warc-fields", "digest": "63ISG6NPUL62BROJ33LNYF4IIT2FXURK", "length": "409", "offset": "4359", "filename": "example.warc.gz"}
```