Starting shortly 🕒
Archive-It Advanced Training

Introduction to the WARC

References:
bit.ly/warc-intro
INTRODUCTION TO THE WARC

Learning objectives

- Understand the key characteristics of the Web ARChive (WARC) format for storage and preservation
- Identify aspects that are specific to Archive-It
- Describe WARC contents and where/how to find them

Prerequisite:

- Some (beginners’ OK!) knowledge of web transfer protocols and conventions

bit.ly/warc-intro
WHAT IS A WARC?

A **WARC** is a file.
WHAT IS A WARC?

A **WARC** is a **container file**.

*It concatenates several files into one object.*

WHAT IS A WARC?

A **WARC** is a container file for storing web content.

WARC files are produced by crawlers, proxies, and other manual utilities that retrieve files from a web server.

WHAT IS A WARC?

A **WARC** is a container file for storing web content *in its original context*.

WARC files contain technical and provenance metadata about the collection and arrangement of media.
WHAT IS A WARC?

A **WARC** is a container file **standard** for storing web content in its original context.

*The WARC file format standard was published by the International Organization for Standardization (ISO) committee on technical interoperability as ISO 28500.*

WHAT IS A WARC?

A **WARC** is a container file standard for storing web content in its original context, *maintained by the International Internet Preservation Consortium (IIPC)*.

An international organization of web archivists is responsible for developing our shared standard.

(VERY BRIEF!) HISTORY OF THE WARC

1996: Internet Archive creates ARC files of web crawl data

2005: Archive-It pilot partners run their first web crawls

2009: WARC file format 1.0 published as ISO 28500

2017: IIPC revises the WARC specification to version 1.1

[bit.ly/warc-intro]
ANATOMY OF A WARC

A **WARC** is a container file standard for storing web content in its original context.

*It includes metadata about its creation and contents, records of server requests and responses, and each response’s full payload.*
# ANATOMY OF A WARC

<table>
<thead>
<tr>
<th>.warc</th>
<th>warcinfo</th>
<th>request</th>
<th>response</th>
<th>revisit</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>resource</td>
<td>conversion</td>
<td>continuation</td>
<td>metadata</td>
</tr>
</tbody>
</table>
ANATOMY OF A WARC

WARC record types

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

Identifies the file and describes when and how it was created:

WARC-Type: warcinfo
WARC-Record-ID: <urn:uuid:cfb2fbba-7b99-4e95-83b4-9313884c83b9>
WARC-Filename: ARCHIVEIT-8232-TEST_CRAWL-JOB1111215-SEED2166618-20200320173416774-00000-xqtcu3m8.warc.gz
WARC-Date: 2020-03-20T17:34:16Z
Content-Type: application/warc-fields
Content-Length: 116
software: warcprox 2.4.26
hostname: wbgrp-svc408.us.archive.org
ip: 207.241.232.59
format: WARC File Format 1.0
ANATOMY OF A WARC

WARC record types

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

Records the client’s original request for content from the server:

WARC-Type: request
WARC-Record-ID: <urn:uuid:997673b5-154e-4173-a597-8c4dbaa412c6>
WARC-Target-URI: https://www.netpreserve.org/
WARC-Date: 2020-03-20T17:34:14Z
WARC-Concurrent-To: <urn:uuid:845ac1f6-9d82-4e24-9693-fc5d9bcceedfe>
WARC-Block-Digest: sha1:YQQEFRPXTLNBEYX5VJBK6M27KRP4UY
Content-Type: application/http,msgtype=request
Content-Length: 420

bit.ly/warc-intro
ANATOMY OF A WARC

WARC record types
- warcinfo
- request
- response
- revisit
- resource
- conversion
- continuation
- metadata

Records the client’s original request for content from the server:

GET /blog/HTTP/1.1
User-Agent: Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/73.0.3683.83 Safari/537.36
Accept: text/html,application/xhtml+xml,application/xml;q=0.9,*/*;q=0.8
Accept-Charset: ISO-8859-1,utf-8;q=0.7,*;q=0.7
Accept-Encoding: gzip, deflate
Accept-Language: en-us,en;q=0.5
X-Forwarded-For: 6.214.43.172
Host: www.netpreserve.org
Via: 1.1 warcprox
ANATOMY OF A WARC

WARC record types
- warcinfo
- request
- response
- revisit
- resource
- conversion
- continuation
- metadata

Records the server’s subsequent response to the client:

WARC-Type: response
WARC-Record-ID: <urn:uuid:a3415adf-1cd3-447c-80c7-e68a18752f05>
WARC-Target-URI: https://widgets.wp.com/likes/index.html
WARC-Date: 2020-03-20T18:57:07Z
WARC-Payload-Digest: sha1:CORSFN5DKI2BBKIK7XKYQMYTTT3MKIC
Content-Type: application/http; msgtype=response
Content-Length: 386
ANATOMY OF A WARC

WARC record types

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

Records the server’s subsequent response to the client:

HTTP/1.1 200 OK
Server: nginx
Date: Thu, 19 Mar 2020 18:57:07 GMT
Content-Type: text/html
Content-Length: 126
Accept-Ranges: bytes

<!DOCTYPE html>
<html>
<head><link rel="stylesheet" href="style.css"/></head>
<body>
<div id="target"></div>
</body>
</html>
ANATOMY OF A WARC

WARC record types
- warcinfo
- request
- response
- revisit
- resource
- conversion
- continuation
- metadata

Documents a de-duplicated server response:

WARC-Type: revisit
WARC-Record-ID: <urn:uuid:46195e2f-463d-45ff-a3c9-8d107ea8ad12>
WARC-Target-URI: https://s0.wp.com/i/favicon.ico
WARC-Date: 2020-03-19T18:57:29Z
WARC-Profile: http://netpreserve.org/warc/1.0/revisit/identical-payload-digest
WARC-Payload-Digest: sha1:AWHWSLHL7LFI7DOK4QXUVH323OGGHPS3
WARC-Refers-To: <urn:uuid:dab1d9c6-e81d-4366-bb52-886923b7d3e6>
WARC-Refers-To-Target-URI: https://s1.wp.com/i/favicon.ico
WARC-Refers-To-Date: 2020-03-19T18:56:21Z
Content-Type: application/http; msgtype=response
Content-Length: 362

bit.ly/warc-intro
ANATOMY OF A WARC

WARC record types

Documents a de-duplicated server response:

HTTP/1.1 200 OK
Server: nginx
Date: Thu, 19 Mar 2020 18:57:29 GMT
Content-Type: image/x-icon
Content-Length: 5430
Connection: close
Last-Modified: Fri, 13 Nov 2015 04:17:50 GMT
Vary: Accept-Encoding
ETag: "5645646e-1536"
Expires: Fri, 28 Aug 2020 04:10:03 GMT
Cache-Control: max-age=31536000
Accept-Ranges: bytes

bit.ly/warc-intro
ANATOMY OF A WARC

**WARC record types**

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

Houses content that was created or retrieved without any URI-scheme:

**WARC-Type:** resource
**WARC-Record-ID:** <urn:uuid:49a5a6bc-a0eb-4d58-9ca8-75ffe2e56434>
**WARC-Target-URI:** screenshot:http://www.netpreserve.org/blog/
**WARC-Date:** 2019-12-19T17:53:11Z
**WARC-Block-Digest:** sha1:GCDC2JZRN52NG2SNE6V52HC5BWUAFNNC
**WARC-Payload-Digest:** sha1:GCDC2JZRN52NG2SNE6V52HC5BWUAFNNC
**Content-Type:** image/jpeg
**Content-Length:** 182270

?&??]s???m4????i?&??k?O??H?[?&??]?s???m4????i?&??k?O??H?[?&??]?s????m4????i?&??k?O??H?[?&??]?s????m4????i?&??k?O??H?[?&??]?s????m4????i?&??k?O??H[?&??]h

# ANATOMY OF A WARC

<table>
<thead>
<tr>
<th>WARC record types</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>warcinfo</td>
<td>request</td>
</tr>
<tr>
<td>response</td>
<td>revisit</td>
</tr>
<tr>
<td>resource</td>
<td>conversion</td>
</tr>
<tr>
<td>continuation</td>
<td>metadata</td>
</tr>
</tbody>
</table>

A WARC record for content that was migrated into a new format:

```
WARC-Type: conversion
WARC-Record-ID: <urn:uuid:27ef1r9j-defr-25f9-328r-16859347dddd>
WARC-Target-URI: http://www.archive.org/images/logoc.jpg
WARC-Date: 2026-09-19T19:00:40Z
WARC-Block-Digest: sha1:XQMRY75YY4ZWC6JAT6KNXKD37F7MOEK
WARC-Refers-To: <urn:uuid:92283950-ef2f-4d72-b224-f54c6ec90bb0>
Content-Type: image/neoimg
Content-Length: 934
```

*Not supported yet generally by known Wayback rendering software*
A WARC record for content that was truncated by storing another WARC file:

WARC-Type: continuation
WARC-Record-ID: <urn:uuid:70653950-a77f-b212-e434-7a7c6ec909ef>
WARC-Target-URI: http://www.archive.org/images/logoc.jpg
WARC-Date: 2016-09-19T17:20:24Z
WARC-Block-Digest: sha1:T7HXETFVA92MSS7ZENMFZY6ND6WF7KB7
WARC-Payload-Digest: sha1:CCHXETFVJD2MUZY6ND6SS7ZENMWF7KQ2
WARC-Segment-Origin-ID: <urn:uuid:39509228-ae2f-11b2-763a-aa4c6ec90bb0>
WARC-Segment-Number: 2
WARC-Segment-Total-Length: 1902
WARC-Identified-Payload-Type: image/jpeg
Content-Length: 302

....
ANATOMY OF A WARC

WARC record types
- warcinfo
- request
- response
- revisit
- resource
- conversion
- continuation
- metadata

Additional capture- or environment-specific information about a WARC record:

WARC-Type: metadata
WARC-Record-ID: <urn:uuid:742428bc-0b4e-4971-b048-5de3a1e9e560>
WARC-Target-URI: https://netpreserveblog.files.wordpress.com/iipc_logo_fullcolor.png
WARC-Date: 2020-03-19T18:56:17Z
Content-Type: application/warc-fields
Content-Length: 476

force-fetch:
hopsFromSeed: 1
fetchTimeMs: 61
charsetForLinkExtraction: ISO-8859-1
INTRODUCTION TO THE WARC

Key advantages:

- Store all the things!
- Extensible and conformant to many collecting and rendering tools
- Standardized and maintained openly by web archivists

Special challenges:

- Volume, count, and “chunking”
- Interdependence and portability
THE WEB ARCHIVING SYSTEMS API

Archive-It web application

warcs.archive-it.org

Wayback
WARCs
Metadata
Partner

bit.ly/warc-intro
**KEY WARC REFERENCES**

⭐️ **WARC specification repository** - IIPC

⭐️ **WARC file format specification 28500:2017** - ISO

⭐️ **WARC File Format (ISO 28500) Pre-print drafts** - BnF

⭐️ **WARC format description** - Library of Congress

⭐️ **Details for WARC 1.0** - PRONOM

⚠️ **Storage and preservation policy** - Archive-It

⚠️ **Find and download your WARC files** - Archive-It

Archive-It Advanced Training

Introduction to the WARC

References:

bit.ly/warc-intro