

WARC Format and Beyond

John Kunze, California Digital Library

Mark Middleton, Hanzo Ltd

Clément Oury, French national library



international internet preservation consortium

The WARC standard

John Kunze, California Digital Library

WARC history

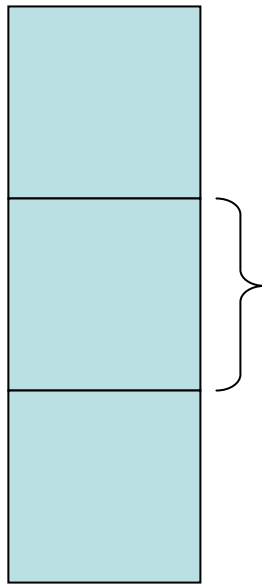
- WARC = Web ARChive file format
 - Created by IIPC
- WARC is next generation of ARC file format
 - ARC format created by the Internet Archive
 - Most legacy web archives in ARC
- Original discussion: Sept 2004
 - First Internet Draft: May 2005
 - First ISO Working Draft: Feb 2006
 - Final ISO Draft: June 2008
 - Final Publication: May 2009

WARC introduction

- A (W)ARC file is a sequence of content blocks, each preceded by a small text header
 - Both allow easy recording of content blocks
 - Only WARC supports *related* content blocks

(W)ARC File Anatomy

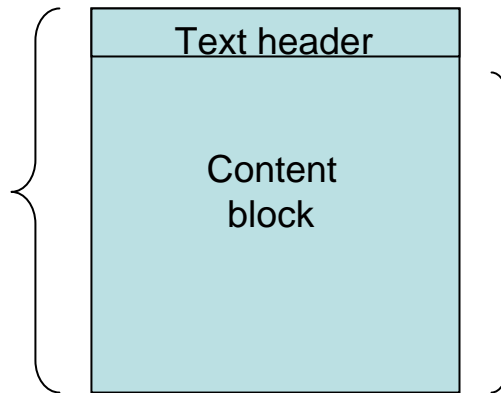
(W)ARC File



·
·
·

Append at will

(W)ARC Record



Length, source URI, date, type, ...

E.g., HTTP response headers and *length* bytes of HTML, GIF, PDF, ...

ARC Header and Content

`http://www.oac.cdlib.org/ 128.48.120.68 20050727235250 text/html 11182`

`HTTP/1.1 200 OK`

`Date: Wed, 27 Jul 2005 23:52:49 GMT`

`Server: Apache/1.3.27 (Unix) mod_perl/1.27`

`Last-Modified: Thu, 02 Jun 2005 00:04:46 GMT`

`ETag: "3cb67-2aa6-429e4d1e"`

`Accept-Ranges: bytes`

`Content-Length: 10918`

`Connection: close`

`Content-Type: text/html`

`<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.0 Transitional//EN">`

`<html>`

`...`

`</html>`

(W)ARC in Context

- Content blocks are not files
- Content blocks are not web pages
 - ... but separate blocks making up a page
- Not all blocks come from web sites
 - In ARC: DNS and first *filedesc* record
 - In WARC, also metadata, conversions, etc.
- Records are sort of peers of files
 - Many “files” in one file for speed and ease

(W)ARCs and Crawling

- One crawl often in multiple (W)ARCs
- Standard tools index each record start
- (W)ARC: records can be order-independent
 - File can be exploded and recombined easily
 - File can be used as a container for anything

In the beginning

... of a (W)ARC file, it may take a few records before you see interesting content, e.g.,

1. file-descriptive record
2. dns:foo.bar
3. <http://foo.bar/robots.txt>
4. maybe finally a record you wanted to harvest

WARC Goals, part 1

- Ability to store arbitrary metadata linked to other stored data (e.g., subject classifier, discovered language, encoding)
- Support for data compression and maintenance of data record integrity
- Ability to store all control information from the harvesting protocol (e.g., request headers), not just response information.

WARC Goals, part 2

- Ability to store the results of data migrations linked to other stored data
- Ability to store a duplicate detection event
- Sufficiently different from the legacy ARC
- Ability to store globally unique record identifiers
- Support for deterministic handling of long records (e.g., truncation, segmentation).

WARC fields, part 1 of 3

WARC-Target-URI

WARC-IP-Address

WARC-Date

Content-Type

Content-Length

WARC-Record-ID

WARC-Refers-To

WARC-Type

WARC-Type values

Warcinfo

Response

Resource

Request

Metadata

Revisit

Conversion

Continuation

... future types ...

WARC fields, part 2 of 3

WARC-Block-Digest

WARC-Payload-Digest

WARC-Warcinfo-ID

WARC-Concurrent-To

WARC-Filename

WARC-Profile

WARC-Identified-Payload-Type

WARC fields, part 3 of 3

WARC-Truncated

WARC-Segment-Origin-ID

WARC-Segment-Number

WARC-Segment-Total-Length

WARC Metadata Example

```
WARC/1.0
WARC-Type: metadata
WARC-Target-URI: http://www.archive.org/images/logoc.jpg
WARC-Date: 2006-09-19T17:20:24Z
WARC-Record-ID: <urn:uuid:16da6da0-bcdc-49c3-927e-57494593b943>
WARC-Concurrent-To: <urn:uuid:92283950-ef2f-4d72-b224-f54c6ec90bb0>
Content-Type: application/warc-fields
WARC-Block-Digest: sha1:UZY6ND6CCHXETFVJD2MSS7ZENMWF7KQ2
Content-Length: 59

via: http://www.archive.org/
hopsFromSeed: E
fetchTimeMs: 565
```

WARC conclusion

- WARC extends ARC's web archiving ability
- WARC remains simple, open, fast, general
 - E.g., LANL journal archiving
- ISO 28500 publication May 2009



international internet preservation consortium

WARC Tools

Mark Middleton, Hanzo Ltd

WARC Usage



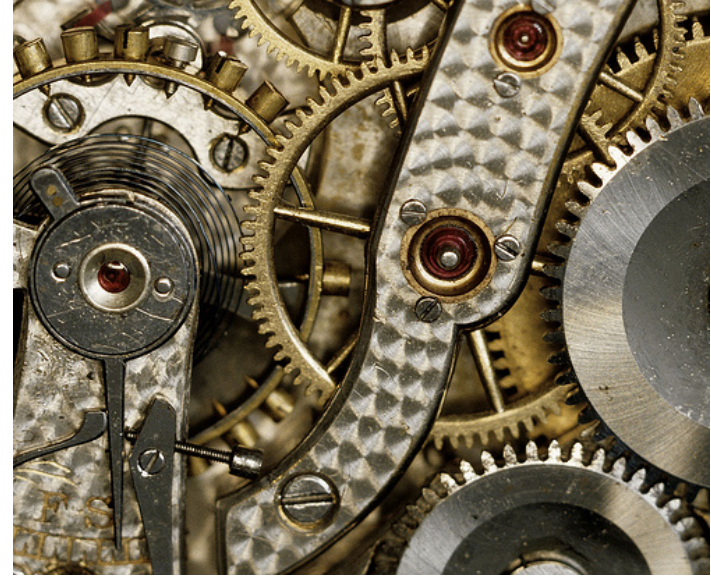
Clément Oury, French national library

Introduction

- Everybody finds WARC attractive
 - Because it has extended features
 - Because it's a standard
- People can now use WARC
 - Many tools already available
- Why haven't all heritage institutions moved to WARC yet?

WARC format: challenges

- ARC format was straightforward...
- Specifications: 4 pages
- 2 record types
- 9 header fields



- WARC is a bit more complex...
- Specifications: 28 pages
- 8 record types
- 17 header fields

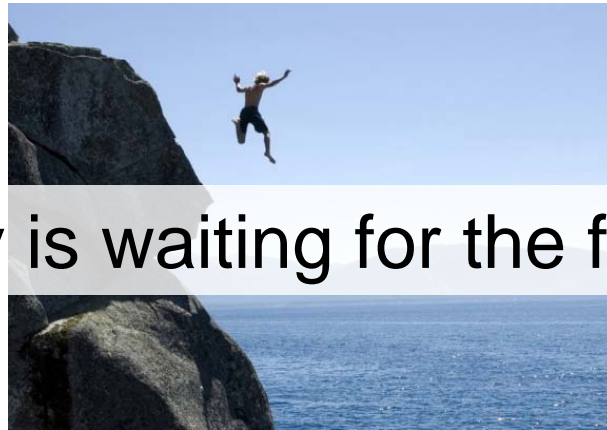
A leap into the unknown?

How much will the transition cost?

When to begin?

How long will the conversion take?

What to start with?



Everybody is waiting for the first to begin

How to manage two generations of formats at the same time?

Can I throw my ARC files away?

Anticipate issues and imagine solutions together



- Task force: grouping web archivists and tools developers together
- First task: “implementation guidelines”
 - WARC standard to provide generic rules only
 - Implementation guidelines to set recommendations on how to write and design WARC files according to functional use cases
- Two examples
 - recording provenance information
 - ensuring interoperability

Provenance information: at the record level

WARC/1.0
WARC-Type: warcinfo
WARC-Record-ID:<D1D2D3D4D5>
[Other header fields]

software: Heritrix 1.12.0...
hostname: crawling017.archive.org
ip: 207.241.227.234
[...]

WARC/1.0
WARC-Type: response
WARC-Warcinfo-ID:<D1D2D3D4D5>
WARC-Record-ID:<A1A2A3A4A5>
[Other header fields]

[http response here]
[image/jpeg binary data here]

WARC/1.0
WARC-Type: request
WARC-Record-ID:<B1B2B3B4B5>
WARC-Concurrent-To:<A1A2A3A4A5>
[Other header fields]

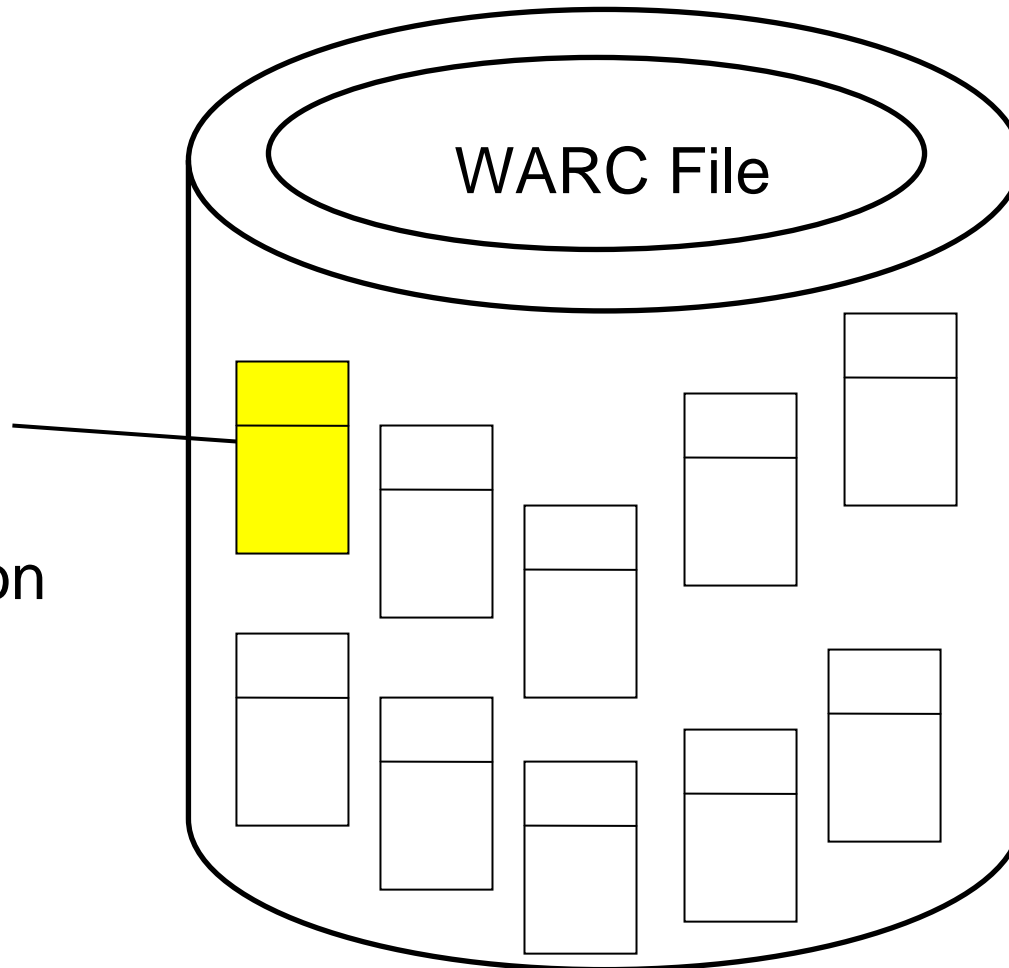
[http request here]

WARC/1.0
WARC-Type: metadata
WARC-Record-ID:<C1C2C3C4C5>
WARC-Concurrent-To:<A1A2A3A4A5>
[Other header fields]

via: http://www.archive.org/
hopsFromSeed: E
fetchTimeMs: 565
[...]

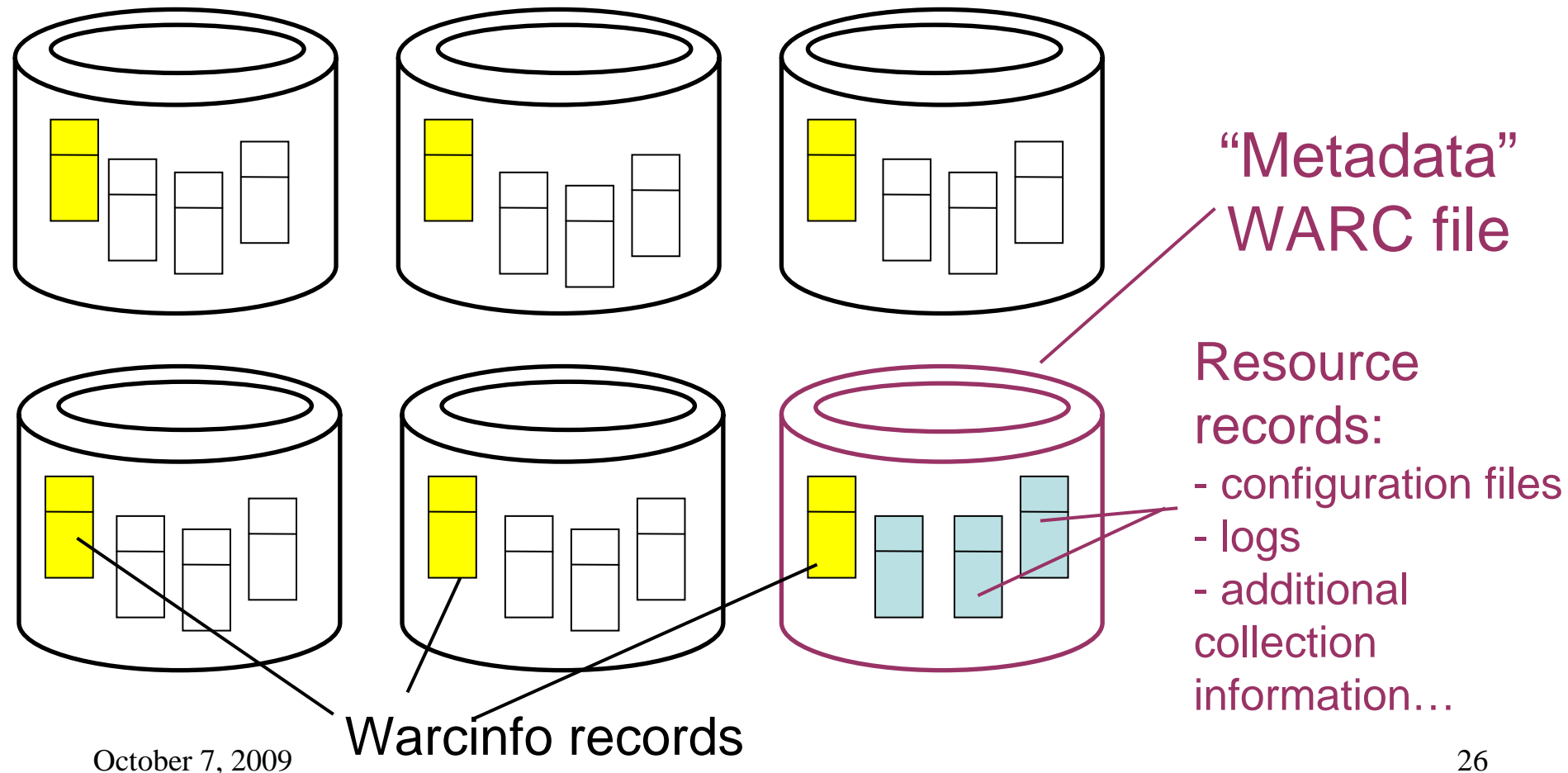
Provenance information: at the WARC file level

**Warcinfo
record:**
High level
configuration
information



Provenance information: at the crawl instance level

Set of WARC files from the same crawl instance



All information useful for...

- Quality Assurance
- Collection management
- Prioritizing preservation actions
- Keeping track of the way web archives were built up

Interoperability issues

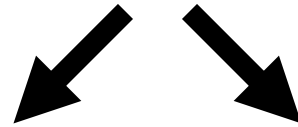
- There are many ways to design a WARC file issued from a web crawl...
- ... and many more ways to create WARC files issued from
 - container format conversion
 - repackaging
 - or other data management operations

One single example: Date of a converted WARC record

http://www.dryswamp.edu:80/index.html 127.10.100.2 19961104142103 text/html 202
[Block]

Original ARC record

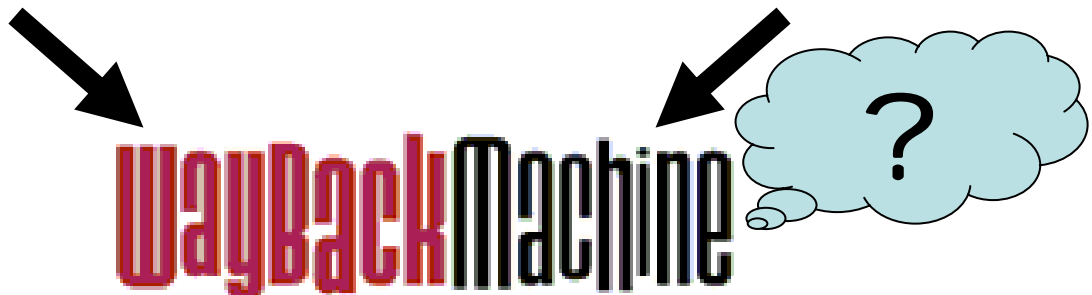
Converted WARC record, first way



Converted WARC record, second way

WARC/1.0 WARC-Type: response WARC-Target-URI: http://www.archive.org/images/logoc.jpg WARC-Date: 1996-11-04T14:21:03Z WARC-Warcinfo-ID: <urn:uuid:d7ae5c10- e6b3-4d27-967d-34780c58ba39> [Other headers fields]
[Block]

WARC/1.0 WARC-Type: response WARC-Target-URI: http://www.archive.org/images/logoc.jpg WARC-Date: 2009-09-20T05:15:59Z WARC-Concurrent-To: <urn:uuid:92283950- ef2f-4d72-b224-f54c6ec90bb0> [Other headers fields]
[Block]



Other topics...

- Choosing and using unique identifiers
- Recording output of payload identification or characterization
- Managing information on viruses

To conclude: What do we need now?

- Share the implementation guidelines
- Maintain task force. More questions left and to come
- Design metrics
 - Time for processing and transition?
 - Cost (machine, labor)?
 - Conversion validation processes?
- Design and test transition plans
- Document and share transition experiences



Questions.. or coffee?



Image credits:

- http://www.preparationmariage.com/IMG/jpg/Fotolia_120123_S.jpg
- http://www.hpceurope.com/img/prdts/presse2009/VIGN_engrenage_rectifie.jpg
- <http://www.flickr.com/photos/villeneuve53/1808995620/>
- http://www.euphoriasmoothies.com/confidential/euphoria_leap_of_faith.jpg
- <http://www.russiablog.org/RedBullCanWeightlifting.png>
- <http://www.flickr.com/photos/careytilden/115435168/>
- <http://www.flickr.com/photos/papalars/2197212826/>
- <http://www.flickr.com/photos/herzogbr/2274372747/>