WARC: an Archiving Format for the Web

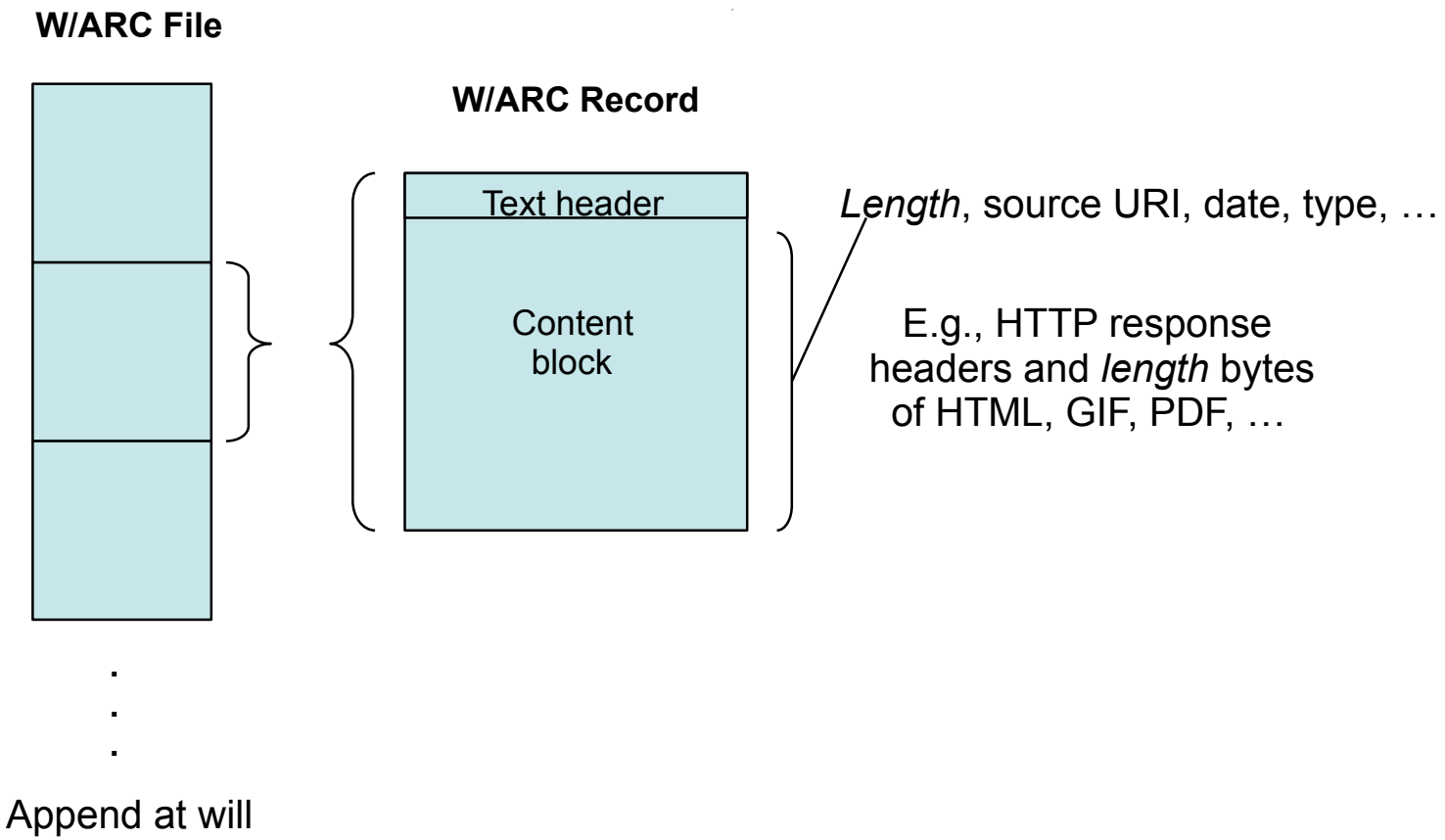
22 September 2005

John Kunze, California Digital Library

WARC Overview

- WARC = Web ARChive file format
- Next generation of ARC, called for by IIPC
  - ARC format created by the Internet Archive
  - Over 600TB of ARCs gathered since 1996
- An ARC or WARC file is a simple sequence of content blocks, each introduced by a small text header
  - ARCs for crawlers to write captured content easily
  - WARCs for captured and *related* content blocks
- Support in Heritrix; later (?) Alexa, HTTrack

W/ARC File Anatomy



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 are order-independent
  - File be exploded and recombined easily
  - File be used as a container for anything
- Typical order: file-descriptive record, dns:foo.bar, http://foo.bar/robots.txt, and then first interesting content

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 Header Overview

- Text line of positional parameter tokens
- Plus optional *named parameters*, e.g.,

```
warc/0.8 7587 response http://www.archive.org/images/logo.jpg
    20050708010101 message/http
    uuid:a4b26b6b-f918-4136-af04-f859d75aebe5
IP-Address: 207.241.224.241
Related-Record-ID: uuid:f569983a-ef8c-4e62-b347-295b227c3e51
Check-Method: sha1:2ZWC6JAT6KNXKD37F7MOEKXQMRY75YY4
blank_line
```

- Compare to ARC header example

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

WARC Record Types

- Warcinfo
- Response
- Resource
- Request
- Metadata
- Revisit
- Conversion
- Continuation

WARC Record Header Id

- Globally unique, persistent enough:

```
warc/0.8 7587 response http://www.archive.org/images/logo.jpg
    20050708010101 message/http
    uuid:a4b26b6b-f918-4136-af04-f859d75aebe5
IP-Address: 207.241.224.241
Related-Record-ID: uuid:f569983a-ef8c-4e62-b347-295b227c3e51
Check-Method: sha1:2ZWC6JAT6KNXKD37F7MOEKXQMRY75YY4
blank_line
```

- Any globally unique URI is ok record id
  - UUID/GUID: two registries (OUI, vendor)
  - ARK, URN, HDL, DOI, URL: one registry

WARC Named Parameters

- IP-Address: IP-address
- Check-Method: algorithm:value [will move]
- Related-Record-ID: record-id
  - required of 'revisit' and 'conversion' records
  - primary record and id conventions optional
- Segment-Origin-ID: record-id
- Segment-Number: integer
- Truncated: reason-token
- E.g., 'length' or 'time' for exceeding a limit
- Warcinfo-ID: record-id

WARC Metadata Example

```
warc/0.8 395 metadata http://www.archive.org/images/logo.jpg
    20050708010101 text/xml
    http://ark.cdlib.org/ark:/13030/xt12rk835gm/_s
Related-Record-ID: http://ark.cdlib.org/ark:/13030/
   xt12rk835gm


<?xml version="1.0"?>
<harvestmetadata
xmlns="http://archive.org/harvest/0.8/">
<discovered-via>http://www.archive.org<discovered-via>
<download-time-ms>565</download-time-ms>
</harvestmetadata>
```

- Similar examples for 'revisit' record type

Conclusion

- WARC extends ARC's web archiving ability
- WARC remains simple, open, fast, general
  - E.g., LANL journal archiving
- Work in progress:

  http://cvs.sourceforge.net/viewcvs.py/archive-access/archive-access/src/docs/warc/
- Co-authors: Allan Arvidson, John Kunze, Gordon Mohr, Michael Stack; comments to
  - jak@ucop.edu