

# Don't WARC Away: Preservation Metadata & Web Archives

Jefferson Bailey & Maria LaCalle, Internet Archive

ALA 2015 | ALCTS PARS | June 27, 2015  
@jefferson\_bail | maria@archive.org



# Don't WARC Away: Preservation Metadata & Web Archives

Jefferson Bailey & Maria LaCalle, Internet Archive

ALA 2015 | ALCTS PARS | June 27, 2015  
@jefferson\_bail | maria@archive.org



- We are a non-profit Digital Library & Archive founded in 1996
- 20+PB unique data: 10PB web, ~8m text, 2m vid, 2m aud, 100K soft, etc
- We work in a former church and it's awesome
- Developed: Heritrix, Wayback, warccprox, Umbra, NutchWax, ARC format
- Engineers, librarians/archivists, program staff



# INTERNET ARCHIVE **WayBack**Machine

- <https://archive.org/web>
- Largest and oldest publicly available web archive in existence
- 485,000,000,000+ URLs (that's billions)
- Like a billion websites, domain agnostic
- Content in 40+ Languages
- Periodic snapshot; 1b+ URLs per week





- <https://archive-it.org/>
- Web archiving service used by 370+ institutions
- 3500+ collection, 10 billion+ URLs
- 49 states and 19 countries
- Libraries, archives, museums, governments, non-profits, etc.
- User groups, Annual Meeting, collaborative and educational projects



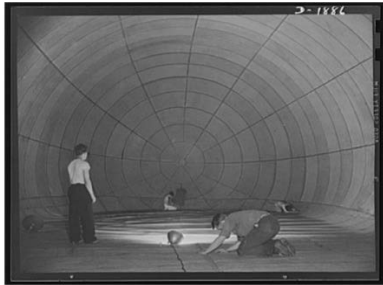
# What is a web archive?

- **Web archiving** is the process of collecting portions of web content, preserving the collections, and then providing access to the archives - for use and re use.
- A **web archive** is a collection of archived URLs grouped by theme, event, subject area, or web address.
- A **web archive** contains as much as possible from the original resources and documents the change over time. It recreates the experience a user would have had if they had visited the live site on the day it was archived.



# Web archive community

WEB ARCHIVING IN THE UNITED  
STATES: A 2013 SURVEY  
AN NDSA REPORT



## NDSA 2013 Survey

- 70% of respondents using Archive-It
- 17% were using California Digital Library's Web Archiving Service
- 81% of organizations devoting one half FTE or less to web archiving

## IIPC 2013 Survey

*Is your web archiving collection  
integrated in your preservation system?*

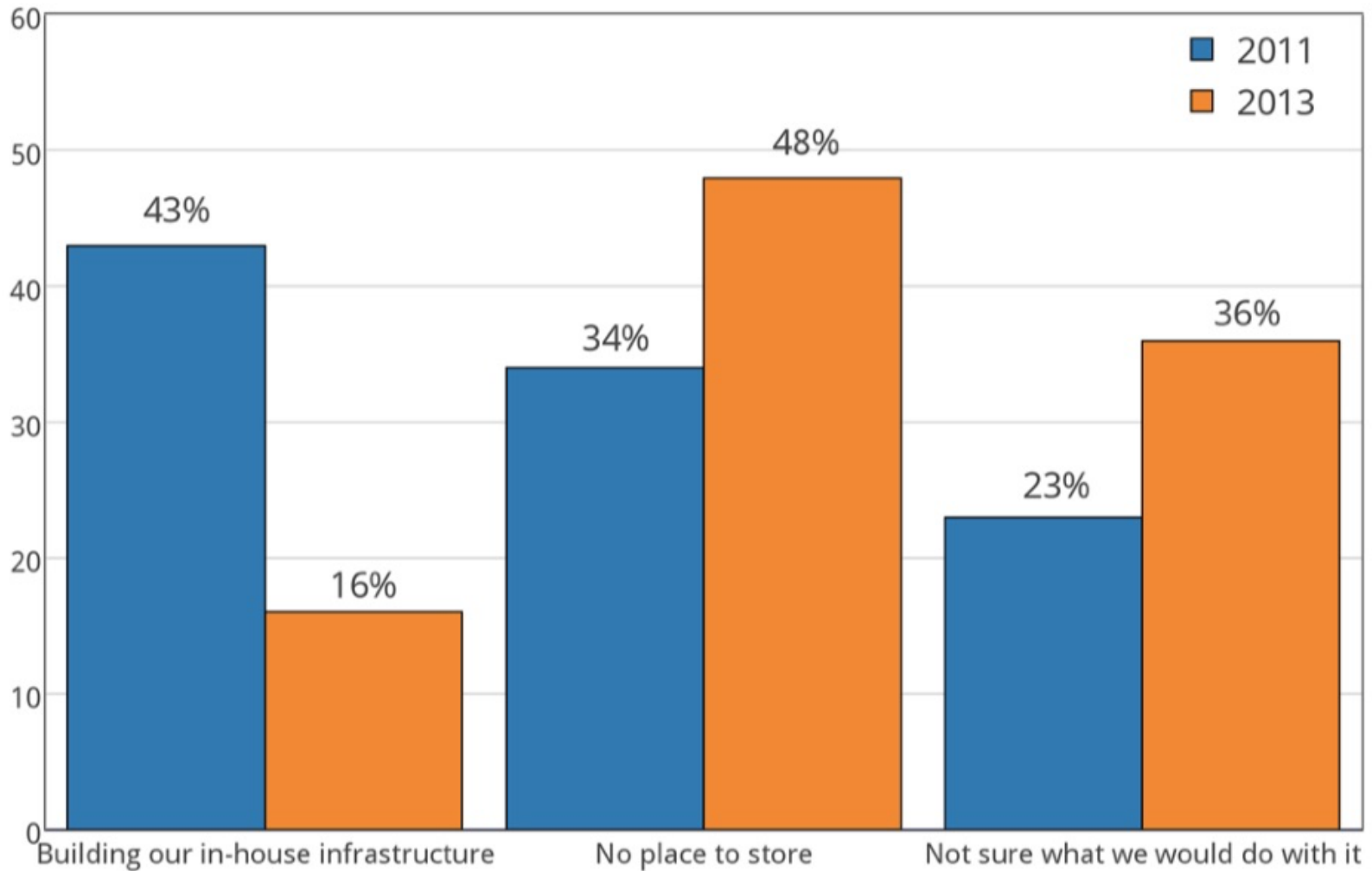
37% Yes

26% Planning to

37% Have not integrated their web  
collection



# REASONS FOR NOT TRANSFERRING DATA FROM AN EXTERNAL SERVICE



# Format Obsolescence: the David Rosenthal perspective

**The vast majority of information generated today will not survive 100 years for reasons that have nothing to do with the interpretability of the bits.**



# WARC (Web ARChive) Format



- ISO 28500:2009
- Combines multiple digital resources into an aggregate archival file together with related information
- Container file
- Written by crawlers
- Concatenated raw content
- For long-term storage and preservation



# WARC: the What and What Not

## The What

- Four required fields:
  - Record Identifier (URI)
  - Content Length/ Record Body Size
  - Timestamp
  - WARC Record Type: 8 different types but most common is the archived response/resource (HTML, pdf, JavaScript...)
- WARC contains extensive technical metadata

## The What Not

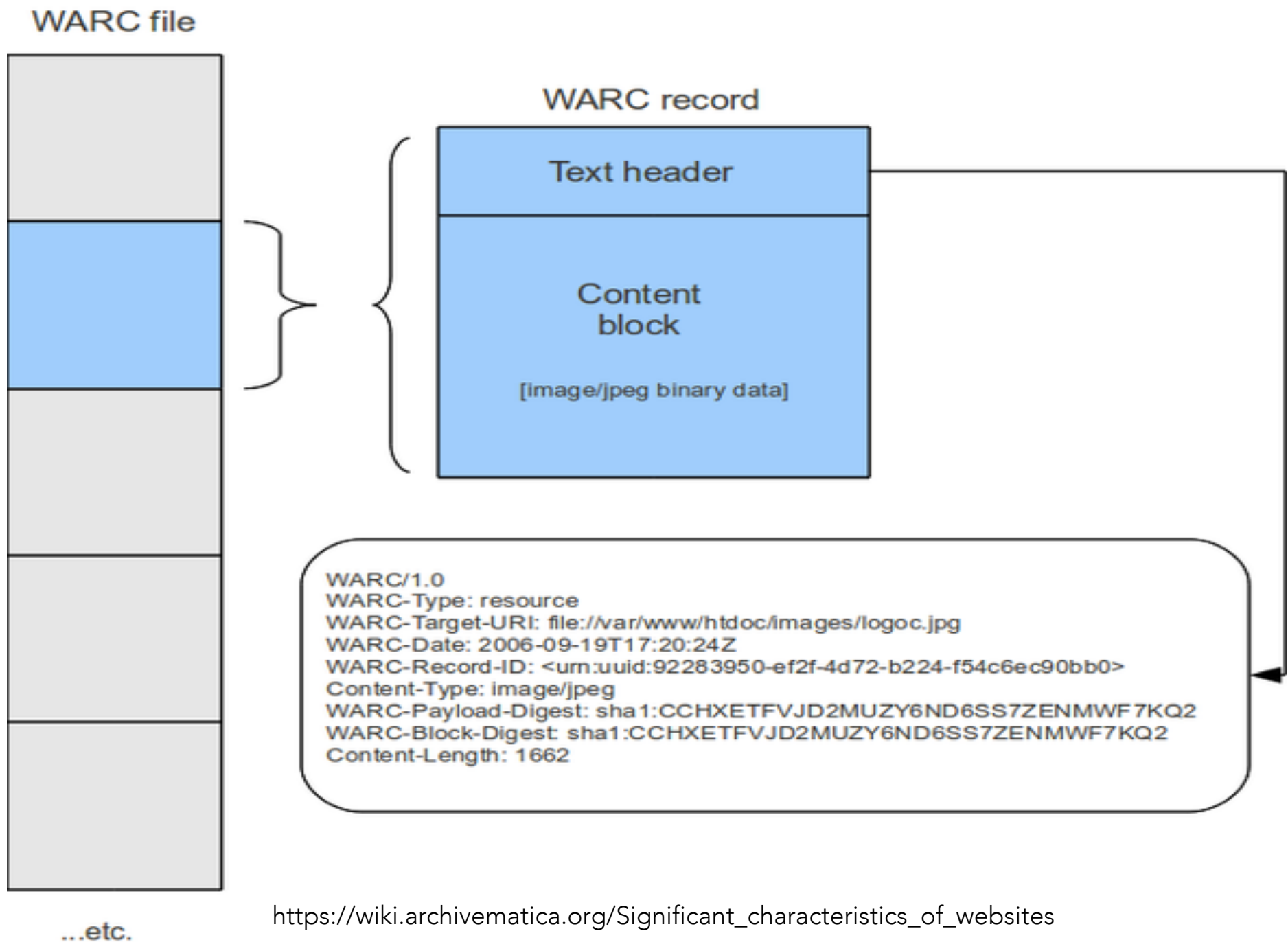
- Rights and permissions
- Descriptions
- Agents & Events
- File format identification
- Validation
- Characterization

# WARC: The Guts

The eight types of WARC records:

- warcinfo – defines records that follow
- response – scheme-specific response (full http response)
- resource – direct retrieval w/o protocol
- request – full http request w/ headers
- metadata – further describe/explain harvested resource (hopsFromSeed, fetchTime)
- revisit – revisitation of previously archived content (dedupe)
- conversion – transformations
- continuation – completion across segmentation





[https://wiki.archivematica.org/Significant\\_characteristics\\_of\\_websites](https://wiki.archivematica.org/Significant_characteristics_of_websites)

ARCHIVEIT-3336-DAILY-26569-20141107000046608-00000-wbgrp-crawl051.us.archive.org-6442.warc

OPEN FILES

ARCHIVEIT-3336-DAILY-26569-20141107000046608-00000-wbgrp-crawl051.us.archive.org-6442.warc

```
4107 WARC/1.0
4108 WARC-Type: response
4109 WARC-Target-URI: http://sfbay.craigslist.org/sfc/roo/4727260590.html
4110 WARC-Date: 2014-11-07T00:00:56Z
4111 WARC-Payload-Digest: sha1:XLKPTY4QQ0F24LEBJYDMUIINKENXU5A0
4112 WARC-IP-Address: 208.82.238.146
4113 WARC-Record-ID: <urn:uuid:6ef7262f-515b-4f3d-8c28-2e00b28b9161>
4114 Content-Type: application/http; msgtype=response
4115 Content-Length: 13253
4116
4117 HTTP/1.1 200 OK
4118 Connection: close
4119 Cache-Control: max-age=300, public
4120 Last-Modified: Fri, 07 Nov 2014 00:00:56 GMT
4121 Date: Fri, 07 Nov 2014 00:00:56 GMT
4122 Vary: Accept-Encoding
4123 Content-Type: text/html; charset=UTF-8
4124 X-MCP-Cache-Control: max-age=2592000, public
4125 X-Frame-Options: SAMEORIGIN
4126 Server: Apache
4127 Expires: Fri, 07 Nov 2014 00:05:56 GMT
4128
4129 <!DOCTYPE html>
4130 <html class="no-js">
4131 <head>
4132   <title>master bedroom with a pirate bathroom in a 3b/2b apartment. $1520/m</title>
4133   <meta name="robots" content="NOARCHIVE,NOFOLLOW">
4134   <meta name="description" content="Hi there, We are looking for one new roommate to fill our 3b/2b apartment at 320 Capp St. The
4135   room is available on Nov. 9th-ish. If you're interested, please reply with a bit about yourself and your...">
4136   <meta name="twitter:card" content="preview">
4137   <meta property="og:description" content="Hi there, We are looking for one new roommate to fill our 3b/2b apartment at 320 Capp
4138   St. The room is available on Nov. 9th-ish. If you're interested, please reply with a bit about yourself and your...">
4139   <meta property="og:image" content="http://images.craigslist.org/00p0p_l3IbaeevjXx_600x450.jpg">
4140   <meta property="og:site_name" content="craigslist">
4141   <meta property="og:title" content="master bedroom with a pirate bathroom in a 3b/2b apartment. $1520/m">
4142   <meta property="og:type" content="article">
4143   <meta property="og:url" content="http://sfbay.craigslist.org/sfc/roo/4727260590.html">
4144   <meta name="viewport" content="initial-scale=1.0, user-scalable=1">
4145   <link type="text/css" rel="stylesheet" media="all" href="//www.craigslist.org/styles/cl.css?v=d7d4e51d9bac1e78a31bb15f752864cf">
4146   <link type="text/css" rel="stylesheet" media="all" href="//www.craigslist.org/styles/leaflet-stock.
4147   css?v=ea2cbe352bcad5eae1c267a9dd15a5c1">
4148   <!--[if lt IE 9]>
4149   <script src="//www.craigslist.org/js/html5shiv.min.js?v=42e8031bc7ca9d67a48f4a5feff7bf29" type="text/javascript" ></script>
4150   <![endif]>
4151   <!--[if lte IE 7]>
4152   <script src="//www.craigslist.org/js/json2.min.js?v=178d4ad319e0e0b4a451b15e49b71bec" type="text/javascript" ></script>
4153   <![endif]>
```

resource

42 matches

Tab Size: 4 Plain Text

# Challenges to Preservation Metadata



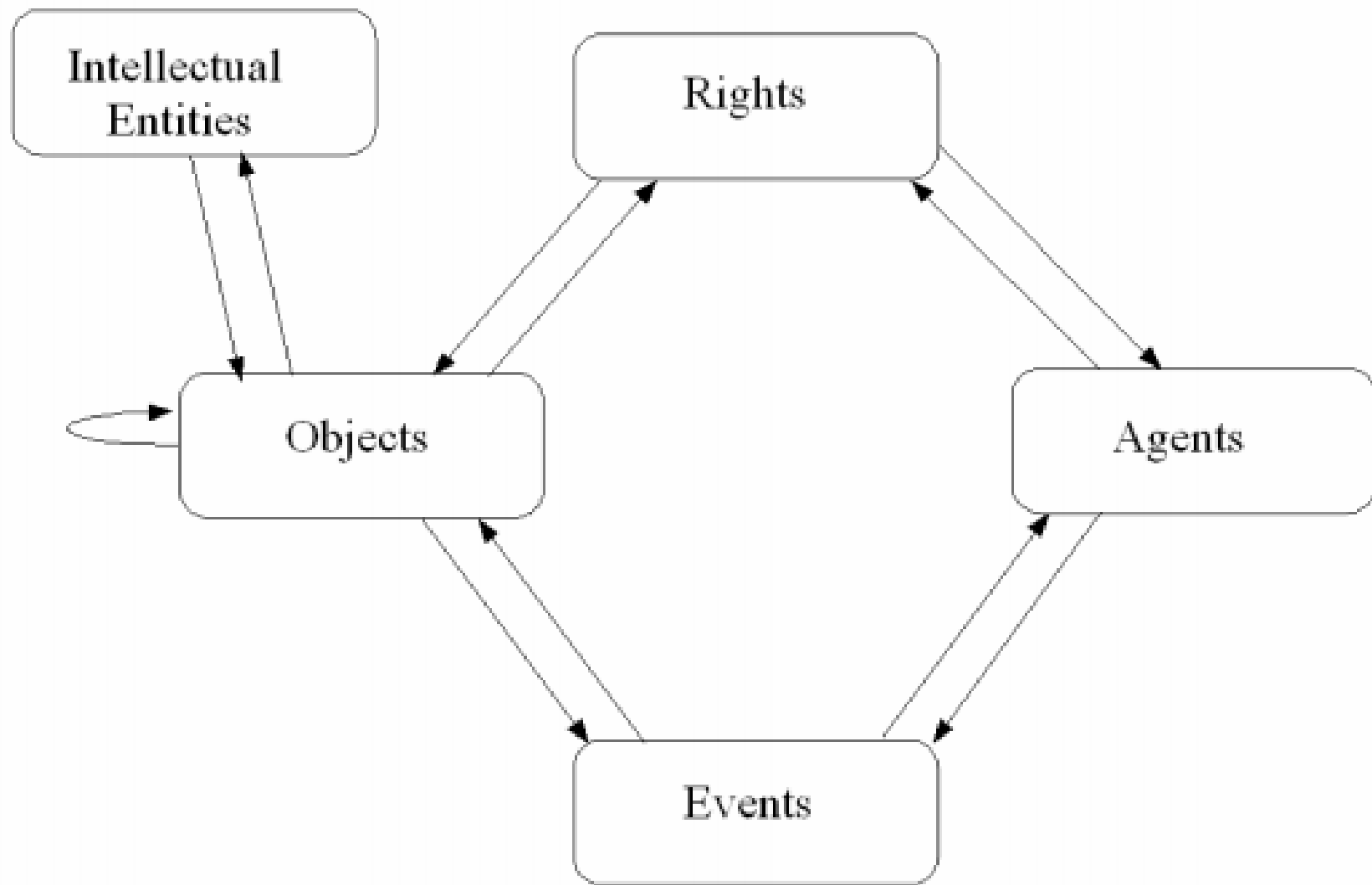
- Concatenated nature
  - Unpack every resource?
- Dispersed for storage
  - Arbitrary placement of resources in WARC files
  - Duplication / revisit
- Unreliable mimes + format verification/obsolescence
- Differentiated preservation actions
- Volume of data

# AIT 2015 Partner Survey

- 80% of respondents do not currently store local copies of their WARC files
  - 53% plan on doing so in the future
  - 41% are considering this for the future
- 20% ingest their WARC files into a digital preservation system or long-term repository
- 14% create metadata for WARC files

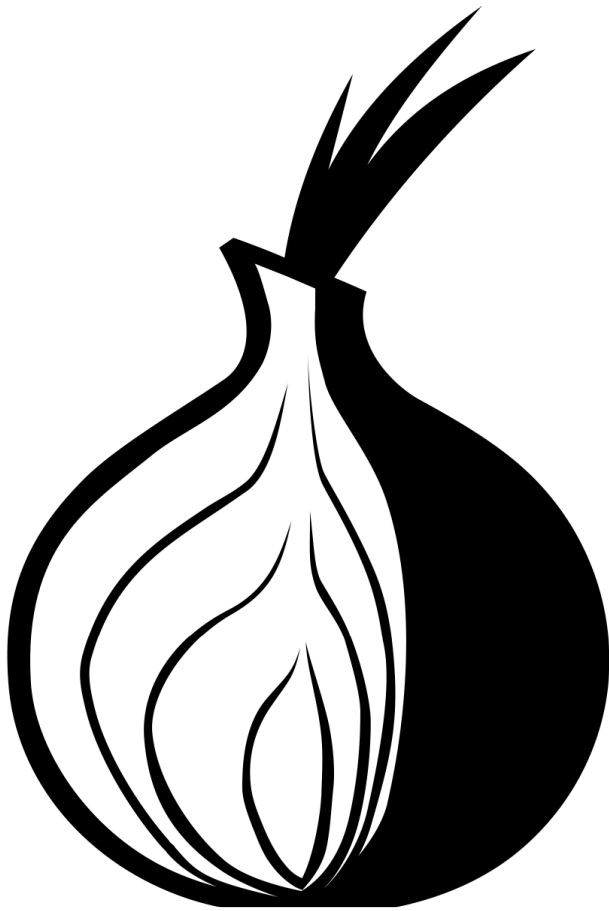




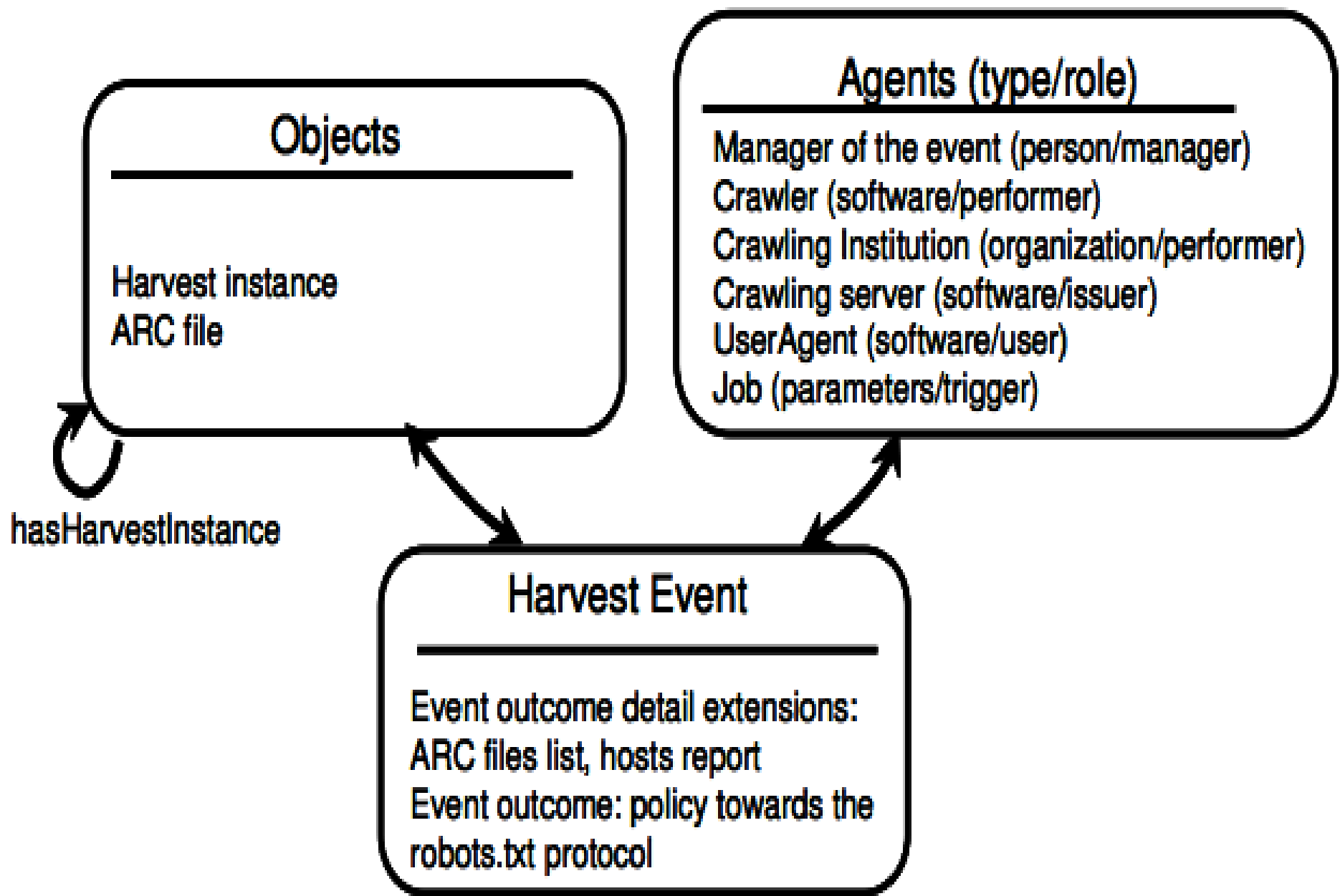


**Figure 1: The PREMIS Data Model**

# The “Onion” Model



- objectCharacteristics
- compositionLevel
- Logical distinctions?
- *“The individual filestream objects are not composition levels of the package file object. They should be considered separate objects, each with their own composition levels.”*
- WARC
  - Record
    - Record type
      - Header
        - Content block
          - Payload
            - Bitstream
              - On and on



# <structMap> & <admSec>

```
<mets:structMap TYPE="logical">
  <!-- the website containing webpages -->
  <mets:div TYPE="WEBSITE">

    <!-- the first webpage -->
    <mets:div TYPE="WEBPAGE"/>

    <!-- definitios of image -->
    <mets:div TYPE="ASSOCIATEDOBJECT" />

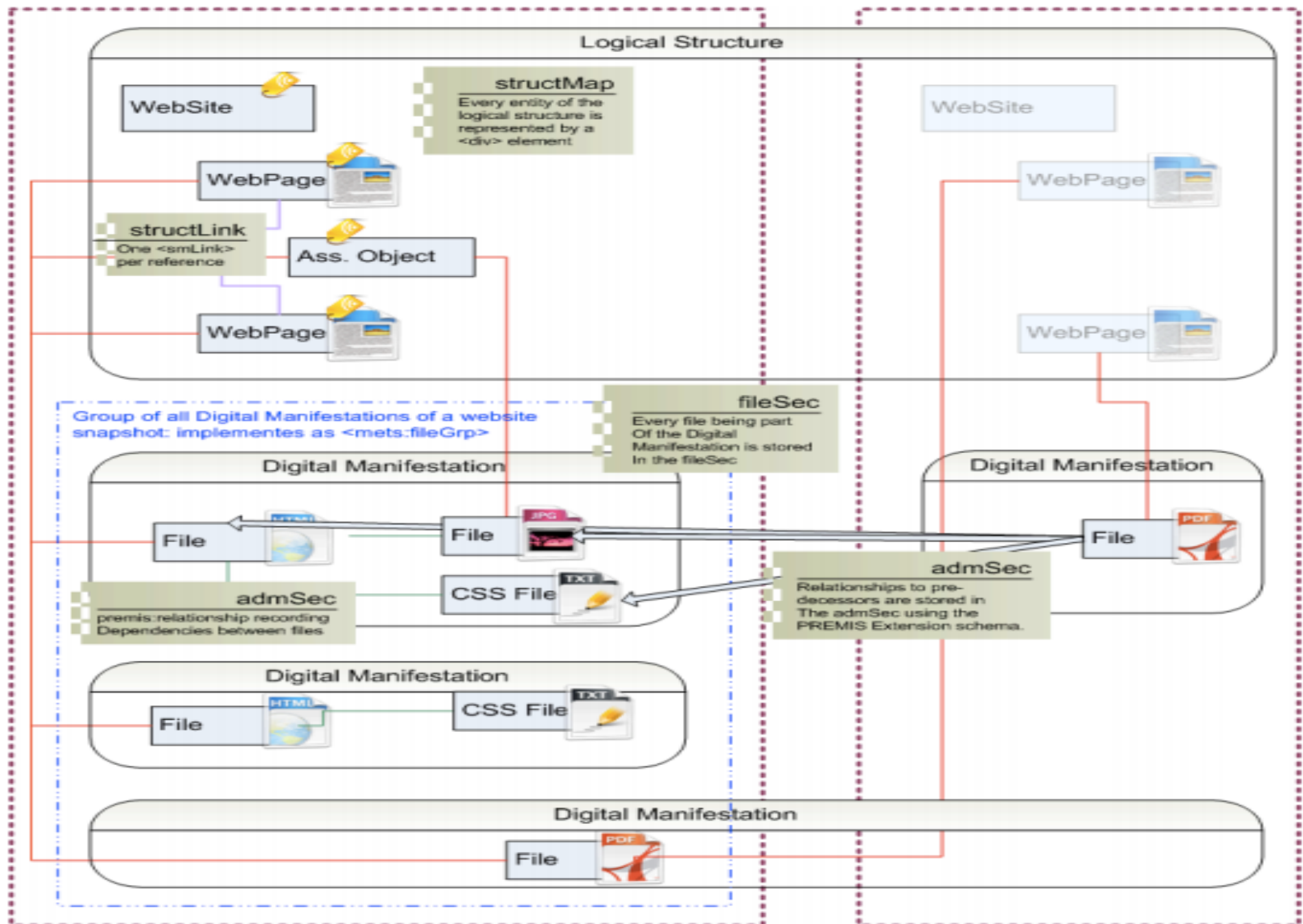
    <!-- the second webpage -->
    <mets:div TYPE="WEBPAGE" />

  </mets:div>
</mets:structMap>
```

**Figure 2.** <structMap> in METS representing the logical structure of a harvested website

```
<premis:event>
  <premis:eventIdentifier>
    <premis:eventIdentifierType>local
    </premis:eventIdentifierType>
    <premis:eventIdentifierValue>event01
    </premis:eventIdentifierValue>
  </premis:eventIdentifier>
  <premis:eventType>migration
  </premis:eventType>
  <premis:eventDateTime>2006-07-16T19:20:30
  </premis:eventDateTime>
  <premis:linkingAgentIdentifier>
    <premis:linkingAgentIdentifierType>local
    </premis:linkingAgentIdentifierType>
    <premis:linkingAgentIdentifierValue>
      agent001
    </premis:linkingAgentIdentifierValue>
  </premis:linkingAgentIdentifier>
</premis:event>
```

**Figure 5.** Representation of an event in PREMIS



Marcus Enders. A METS based information package for long term accessibility of Web Archives. (2010)

**THE GOGGLES**



**THEY DO NOTHING**

# Premises that complicate PREMIS

- Little local acquisition
- Format opacity
- Concatenation / compression
- Crawler variance
- Policies / Agents
- Scale, scale, scale

# "Practical" Approaches

- Data redundancy over metadata granularity
- Utilize Crawl/Crawler-specific resources
  - `mimetype-report.txt`
  - `crawl-report.txt`
  - CDX index
- Utilize additional crawl reporting
  - Host reports, etc
- Decomposition levels?
- Simplify events/agents/objects

# Closing/Discussion Thoughts

- Forecasting Obsolescence
- Collection vs. Control
- Institutional vs. Technological
- Lightweight Tonnage of Data



# THANKS!

Jefferson Bailey, Internet Archive  
[jefferson@archive.org](mailto:jefferson@archive.org) | [@jefferson\\_bail](https://twitter.com/jefferson_bail)

Maria LaCalle, Internet Archive  
[maria@archive.org](mailto:maria@archive.org)

Internet Archive  
<https://archive.org>

Archive-It  
<https://archive-it.org>



- We are a non-profit Digital Library & Archive founded in 1996
- 20+PB unique data: 10PB web, ~8m text, 2m vid, 2m aud, 100K soft, etc
- We work in a former church and it's awesome
- Developed: Heritrix, Wayback, warccprox, Umbra, NutchWax, ARC format
- Engineers, librarians/archivists, program staff



# INTERNET ARCHIVE **WayBack**Machine

- <https://archive.org/web>
- Largest and oldest publicly available web archive in existence
- 485,000,000,000+ URLs (that's billions)
- Like a billion websites, domain agnostic
- Content in 40+ Languages
- Periodic snapshot; 1b+ URLs per week





- <https://archive-it.org/>
- Web archiving service used by 370+ institutions
- 3500+ collection, 10 billion+ URLs
- 49 states and 19 countries
- Libraries, archives, museums, governments, non-profits, etc.
- User groups, Annual Meeting, collaborative and educational projects



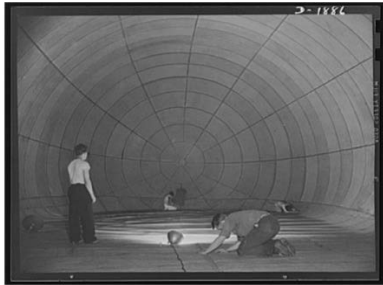
# What is a web archive?

- **Web archiving** is the process of collecting portions of web content, preserving the collections, and then providing access to the archives - for use and re use.
- A **web archive** is a collection of archived URLs grouped by theme, event, subject area, or web address.
- A **web archive** contains as much as possible from the original resources and documents the change over time. It recreates the experience a user would have had if they had visited the live site on the day it was archived.



# Web archive community

WEB ARCHIVING IN THE UNITED  
STATES: A 2013 SURVEY  
AN NDSA REPORT



## NDSA 2013 Survey

- 70% of respondents using Archive-It
- 17% were using California Digital Library's Web Archiving Service
- 81% of organizations devoting one half FTE or less to web archiving

## IIPC 2013 Survey

*Is your web archiving collection  
integrated in your preservation system?*

37% Yes

26% Planning to

37% Have not integrated their web  
collection



# Format Obsolescence: the David Rosenthal perspective

**The vast majority of information generated today will not survive 100 years for reasons that have nothing to do with the interpretability of the bits.**



# WARC (Web ARChive) Format



- ISO 28500:2009
- Combines multiple digital resources into an aggregate archival file together with related information
- Container file
- Written by crawlers
- Concatenated raw content
- For long-term storage and preservation



# WARC: the What and What Not

## The What

- Four required fields:
  - Record Identifier (URI)
  - Content Length/ Record Body Size
  - Timestamp
  - WARC Record Type: 8 different types but most common is the archived response/resource (HTML, pdf, JavaScript...)
- WARC contains extensive technical metadata

## The What Not

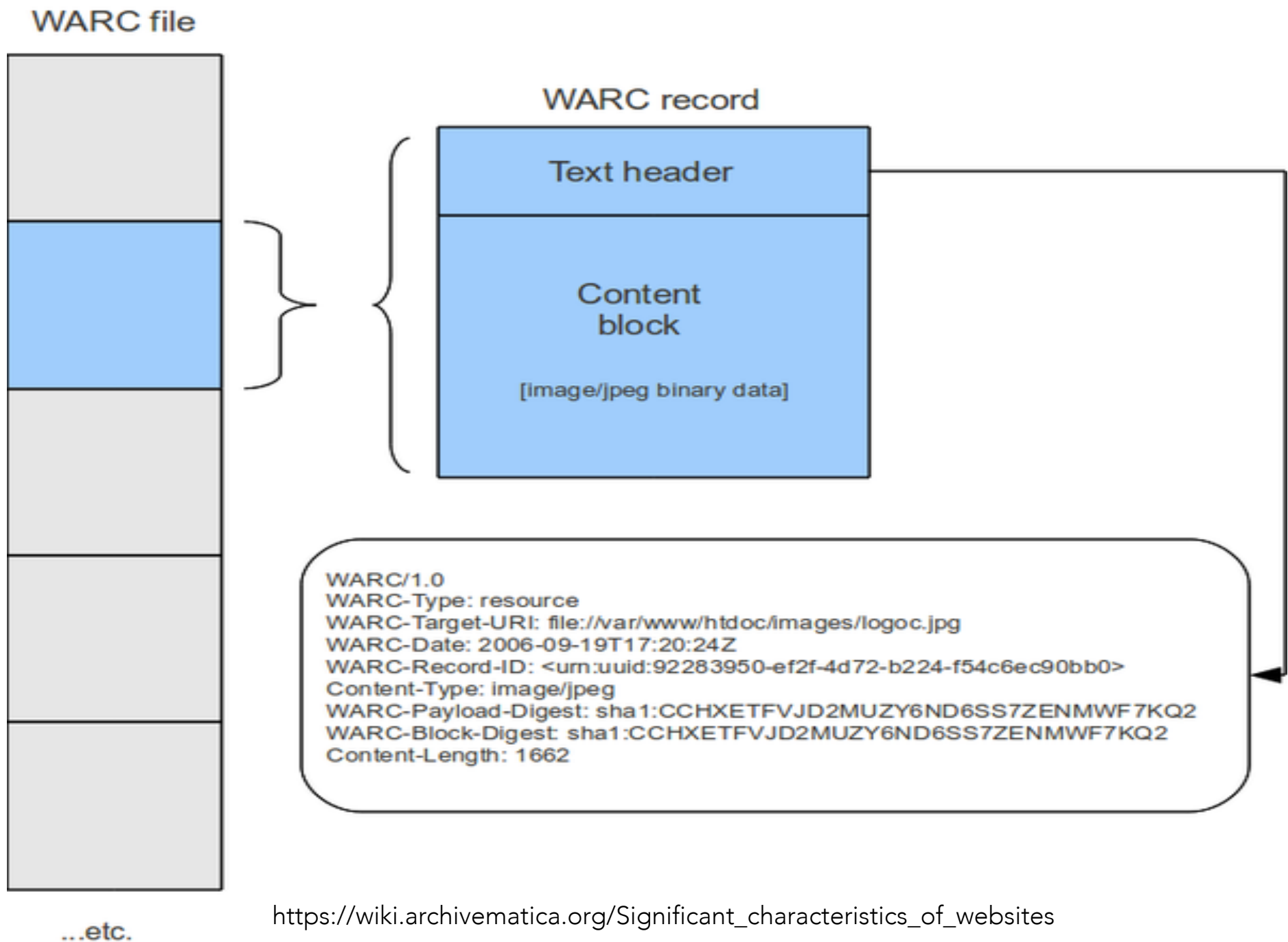
- Rights and permissions
- Descriptions
- Agents & Events
- File format identification
- Validation
- Characterization

# WARC: The Guts

The eight types of WARC records:

- warcinfo – defines records that follow
- response – scheme-specific response (full http response)
- resource – direct retrieval w/o protocol
- request – full http request w/ headers
- metadata – further describe/explain harvested resource (hopsFromSeed, fetchTime)
- revisit – revisitation of previously archived content (dedupe)
- conversion – transformations
- continuation – completion across segmentation





[https://wiki.archivematica.org/Significant\\_characteristics\\_of\\_websites](https://wiki.archivematica.org/Significant_characteristics_of_websites)

ARCHIVEIT-3336-DAILY-26569-20141107000046608-00000-wbgrp-crawl051.us.archive.org-6442.warc

OPEN FILES

ARCHIVEIT-3336-DAILY-26569-20141107000046608-00000-wbgrp-crawl051.us.archive.org-6442.warc

```
4107 WARC/1.0
4108 WARC-Type: response
4109 WARC-Target-URI: http://sfbay.craigslist.org/sfc/roo/4727260590.html
4110 WARC-Date: 2014-11-07T00:00:56Z
4111 WARC-Payload-Digest: sha1:XLKPTY4QQ0F24LEBJYDMUIINKENXU5A0
4112 WARC-IP-Address: 208.82.238.146
4113 WARC-Record-ID: <urn:uuid:6ef7262f-515b-4f3d-8c28-2e00b28b9161>
4114 Content-Type: application/http; msgtype=response
4115 Content-Length: 13253
4116
4117 HTTP/1.1 200 OK
4118 Connection: close
4119 Cache-Control: max-age=300, public
4120 Last-Modified: Fri, 07 Nov 2014 00:00:56 GMT
4121 Date: Fri, 07 Nov 2014 00:00:56 GMT
4122 Vary: Accept-Encoding
4123 Content-Type: text/html; charset=UTF-8
4124 X-MCP-Cache-Control: max-age=2592000, public
4125 X-Frame-Options: SAMEORIGIN
4126 Server: Apache
4127 Expires: Fri, 07 Nov 2014 00:05:56 GMT
4128
4129 <!DOCTYPE html>
4130 <html class="no-js">
4131 <head>
4132   <title>master bedroom with a pirate bathroom in a 3b/2b apartment. $1520/m</title>
4133   <meta name="robots" content="NOARCHIVE,NOFOLLOW">
4134   <meta name="description" content="Hi there, We are looking for one new roommate to fill our 3b/2b apartment at 320 Capp St. The
4135   room is available on Nov. 9th-ish. If you're interested, please reply with a bit about yourself and your...">
4136   <meta name="twitter:card" content="preview">
4137   <meta property="og:description" content="Hi there, We are looking for one new roommate to fill our 3b/2b apartment at 320 Capp
4138   St. The room is available on Nov. 9th-ish. If you're interested, please reply with a bit about yourself and your...">
4139   <meta property="og:image" content="http://images.craigslist.org/00p0p_l3IbaeevjXx_600x450.jpg">
4140   <meta property="og:site_name" content="craigslist">
4141   <meta property="og:title" content="master bedroom with a pirate bathroom in a 3b/2b apartment. $1520/m">
4142   <meta property="og:type" content="article">
4143   <meta property="og:url" content="http://sfbay.craigslist.org/sfc/roo/4727260590.html">
4144   <meta name="viewport" content="initial-scale=1.0, user-scalable=1">
4145   <link type="text/css" rel="stylesheet" media="all" href="//www.craigslist.org/styles/cl.css?v=d7d4e51d9bac1e78a31bb15f752864cf">
4146   <link type="text/css" rel="stylesheet" media="all" href="//www.craigslist.org/styles/leaflet-stock.
4147   css?v=ea2cbe352bcad5eae1c267a9dd15a5c1">
4148   <!--[if lt IE 9]>
4149   <script src="//www.craigslist.org/js/html5shiv.min.js?v=42e8031bc7ca9d67a48f4a5feff7bf29" type="text/javascript" ></script>
4150   <![endif]>
4151   <!--[if lte IE 7]>
4152   <script src="//www.craigslist.org/js/json2.min.js?v=178d4ad319e0e0b4a451b15e49b71bec" type="text/javascript" ></script>
4153   <![endif]>
```

resource

42 matches

Tab Size: 4 Plain Text

# Challenges to Preservation Metadata



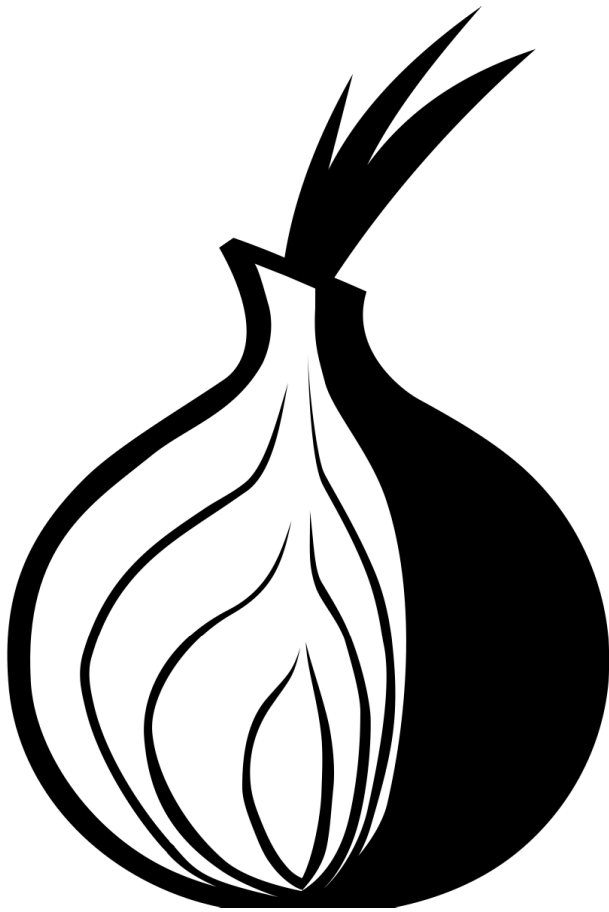
- Concatenated nature
  - Unpack every resource?
- Dispersed for storage
  - Arbitrary placement of resources in WARC files
  - Duplication / revisit
- Unreliable mimes + format verification/obsolescence
- Differentiated preservation actions
- Volume of data

# AIT 2015 Partner Survey

- 80% of respondents do not currently store local copies of their WARC files
  - 53% plan on doing so in the future
  - 41% are considering this for the future
- 20% ingest their WARCs into a digital preservation system or long-term repository
- 14% create metadata for WARC files



# The “Onion” Model



- objectCharacteristics
- compositionLevel
- Logical distinctions?
- *“The individual filestream objects are not composition levels of the package file object. They should be considered separate objects, each with their own composition levels.”*
- WARC
  - Record
    - Record type
      - Header
        - Content block
          - Payload
            - Bitstream
              - On and on

**THE GOGGLES**

**THEY DO NOTHING**

# Premises that complicate PREMIS

- Little local acquisition
- Format opacity
- Concatenation / compression
- Crawler variance
- Policies / Agents
- Scale, scale, scale

# "Practical" Approaches

- Data redundancy over metadata granularity
- Utilize Crawl/Crawler-specific resources
  - `mimetype-report.txt`
  - `crawl-report.txt`
  - CDX index
- Utilize additional crawl reporting
  - Host reports, etc
- Decomposition levels?
- Simplify events/agents/objects

# Closing/Discussion Thoughts

- Forecasting Obsolescence
- Collection vs. Control
- Institutional vs. Technological
- Lightweight Tonnage of Data



# THANKS!

Jefferson Bailey, Internet Archive  
[jefferson@archive.org](mailto:jefferson@archive.org) | [@jefferson\\_bail](https://twitter.com/jefferson_bail)

Maria LaCalle, Internet Archive  
[maria@archive.org](mailto:maria@archive.org)

Internet Archive  
<https://archive.org>

Archive-It  
<https://archive-it.org>

