

**ARCHIVE-IT**

**Partner Meeting 2017 | Portland, OR**



**This Archive-It Goes to 11 (2006-2017)**

# Staff Updates & Shout Outs

## NEW AIT TEAM MEMBERS

- Mary Haberle ([mhaberle@archive.org](mailto:mhaberle@archive.org)), Web Archivist
- Hiring an **Engineering Manager, Archive-It** now! (Check IA jobs page)

## OTHER NOTABLES

- New Web Archiving Team hires:
  - Gemma Batson, European Program Manager (AIT & Web Group)
  - Jaimie Murdock, Web Crawl Engineer
  - Bryan Newbold, Web & Open Data Engineer
- In the house: Barbara, Karl, Kyrie, Lori, Maria, Mouse, Neil, Sylvie
- Wish you were here: Adam, James, Jillian, Mark, Noah, Vinay

#aitpdx @archiveitorg



# Community Updates New Partners (73)

Harvard Kennedy School of Government  
Harvard Divinity School  
University of Wisconsin, Madison SILS  
Sisters of Charity Federation  
Lewis and Clark College  
University of North Texas  
U.S. Food and Drug Administration  
Bruderhof  
U.S. Department of Labor  
Rooftop Alternative PreK-8 School  
School District of Philadelphia  
St. John Newmann  
Cass Junior High School  
Norfolk Collegiate School  
Williams Middle Magnet School  
The University of Oklahoma - SLIS  
Haugton Library (Harvard)  
Museum of Comparative Zoology (Harvard)  
Lamont Library (Harvard)

Baruch College  
Skidmore College  
Augustana College  
University of Texas - Arlington  
Auraria Library  
Senator John McCain  
Ivy Plus Libraries  
Phillip Exeter Academy  
Hamilton College  
Department of Education  
Missouri State Archives  
Lynn University  
Getty Research Institute  
Federal Reserve Bank of St Louis  
Bowdoin College  
Park School of Baltimore  
American Legion  
Georgia College  
Union of Concerned Scientists  
California Pirate Party  
National Library of Norway  
University of Arizona

Bibliotheksservice-Zentrum Baden-Württemberg  
Grand Valley State University  
Colorado State University  
City of Vancouver Archives  
University of Nebraska - Lincoln  
University of Michigan Special Collections  
British Columbia Institute of Technology  
Space Telescope Science Institute (MAST)  
Nationwide Mutual Insurance Company  
District of Columbia Public Library  
California Polytechnic State University  
Local Government Association of the UK  
California Earthquake Authority  
Nelson-Atkins Museum of Art  
Federal Reserve Bank of Minneapolis  
Royal Ontario Museum  
NAACP Legal Defense Fund  
Oregon Health & Science University  
Brock University  
Lakehead University  
University of Ontario Institute of Technology

University of Western Ontario  
Scholars Portal  
University of Ottawa  
Vancouver Island University  
Duke University Medical Center  
Archive  
University of Guelph  
Rhode Island Office of Library &  
Information Services  
Memorial University of Newfoundland  
Mt. San Antonio College  
Russell Sage Foundation  
OSI Systems Inc  
Laurentian University  
National Library of Norway  
University of Arizona

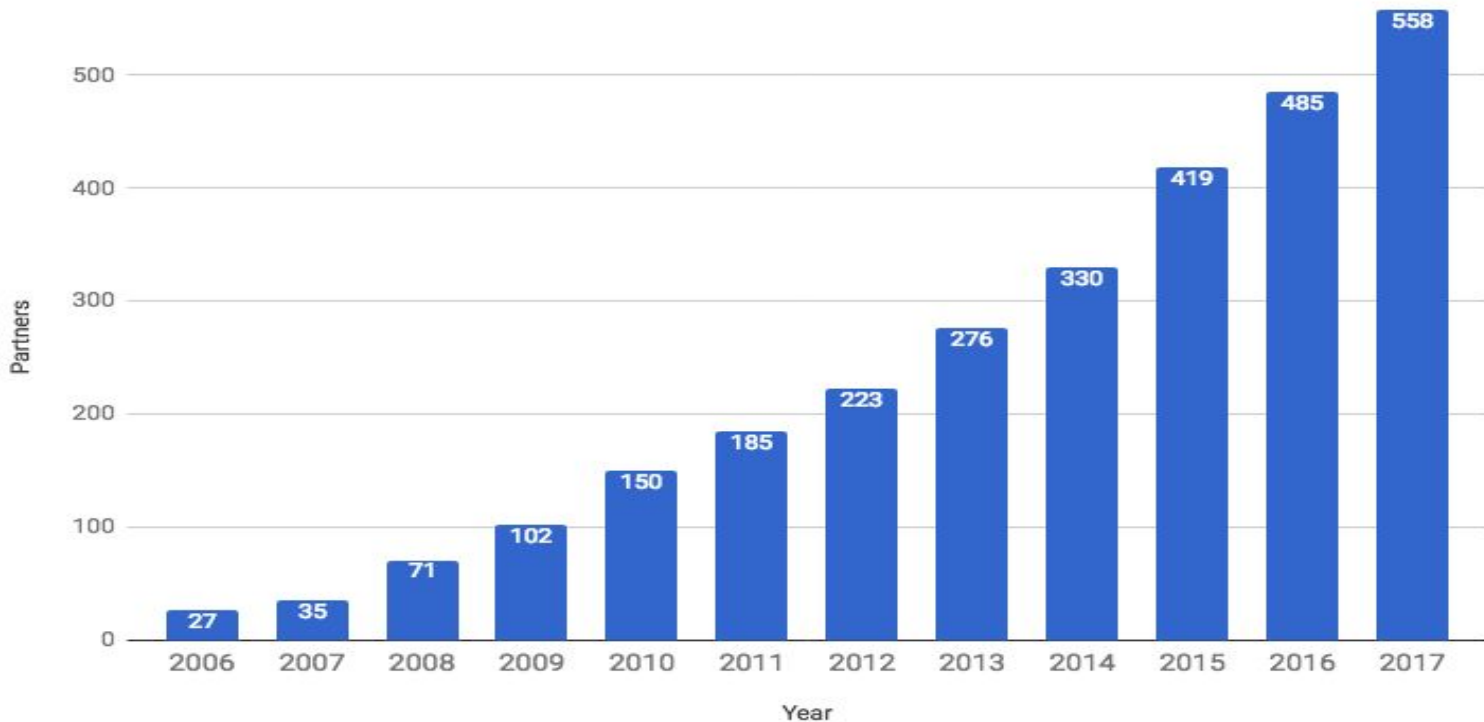
**2006-today: 523 Archive-It partners | 36 in 2017 so far**

#aitpdx @archiveitorg



# AIT Community Updates

Partners vs. Year



#aitpdx @archiveitorg



# AIT Community Updates



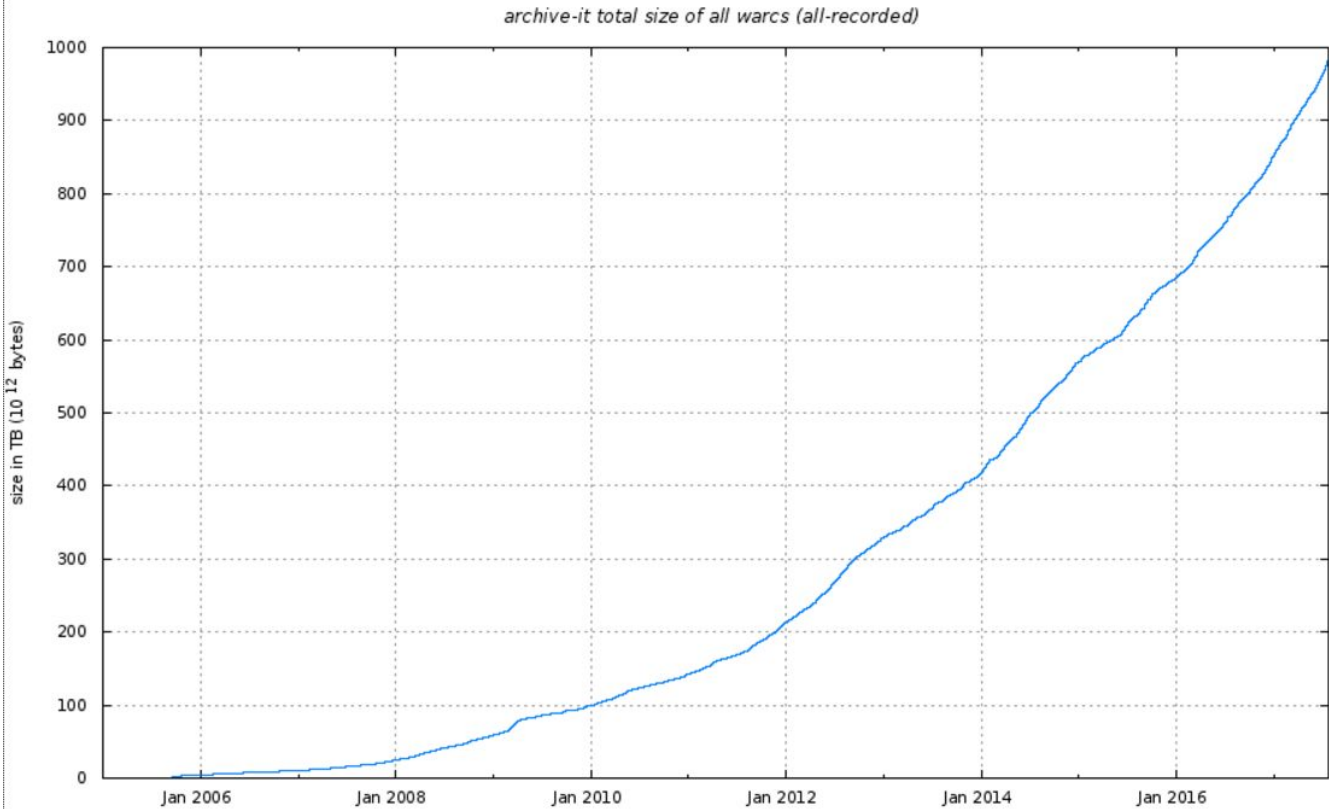
## Consortial & Other Partnerships

- Canada-wide web archiving group (COPPUL/OCUL) - 24 members
- BSZ - 15 members
- Harvard University Libraries - 9 members
- WorldCat
- TechSoup
- Other consortia: LIPA, CDL, KAIC, Yale, Stanford, Five Colleges

#aitpdx @archiveitorg



# AIT Community Updates



#aitpdx @archiveitorg



# AIT stats-a-thon

- Total URLs: 23 Billion+ URLs preserved
  - (+37% from 16B 1Y ago)
- Total Data: ~985TB
  - (was ~800TB a year ago)
- Total Collections: ~8700 public+private
  - (~7200 a year ago)
- Total Seeds: ~1,150,000+ (woo!)
  - (1,000,000 a year ago)
- Total Crawls: ~295,926 (new stat)

#aitpdx @archiveitorg





# Grants & Strategic Activities



“Combining Social Media Storytelling With Web Archives”  
- Connecting AIT collections to Storify



“Global Events & Trends Archive” (GETAR)  
- Events archiving, data mining, K-12 education



“New Measures Research Project”  
- Snapshot crawls of 600+ community news sites



L3S Research Center & Archives Unleashed  
- ArchiveSpark, Archive-It Datathons, AU Hackathons



Society of California Archivists

#aitpdx @archiveitorg



# Grants & Strategic Activities

The Andrew W. Mellon Foundation



NYU



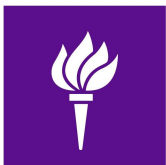
The screenshot shows a web browser window with the Archive-It interface. The address bar displays the URL `https://partner.qa-archive-it.org/567/collections/4049/seeds/1372363/metadata`. The page title is `http://www.campjulie.com`. The metadata section shows the URL `http://composers.dlts.org:8089/plugins/composers/archiveit?resource_id=MSS.479` under the heading 'Related Archival Materials'. There are buttons for 'Import from Worldcat' and 'Edit'. The page is part of the 'New York University Archive of Contemporary Composers' Websites' collection.

#aitpdx | @archiveitorg



# Grants & Strategic Activities

The Andrew W. Mellon Foundation



NYU



## Narrow Your Results

**Group** Sort By: **Count** | (A-Z)

1 Semiannual (9)  
10 Semiannual (8)  
11 Semiannual (9)  
12 Semiannual (9)  
13 Semiannual (9)

**More** ▾

**Creator** Sort By: **Count** | (A-Z)

Chambers, Wendy Mae. (2)  
Mumma, Gordon, 1935- (2)  
Young, Katherine, 1980- (2)  
Adamo, Mark. (1)  
Adler, Christopher. (1)

**More** ▾

**Publisher** Sort By: **Count** | (A-Z)

Vimeo (1)

**Language** Sort By: **Count** | (A-Z)

English (8)  
French (1)

Sites for this collection are listed below. Narrow your results at left, or enter a search query below to find a site, specific URL or to search the text of archived webpages.

Search

Clear

Sites

Search Page Text

◀ Prev Page

Page 2 of 3 (204 Total Results)

Next Page ▶

Sort By: **Title (A-Z)** | Title (Z-A) | URL (A-Z) | URL (Z-A)

URL: <http://www.campjulie.com>

Captured 5 times between May 5, 2017 and May 5, 2017

### Related Archival Materials

**Julie Covello Papers**

Extent: 36 Digital Objects

URL: <http://www.campjulie.com/wp-content/plugins/download-monitor/download.php?id=16>

No Captures were found for this URL.

Title: Carl Schimmel

URL: <http://www.carlschimmel.com/>

Captured 15 times between Feb 12, 2014 and May 4, 2016

Creator: Schimmel, Carl, 1975-

Title: Carman Moore

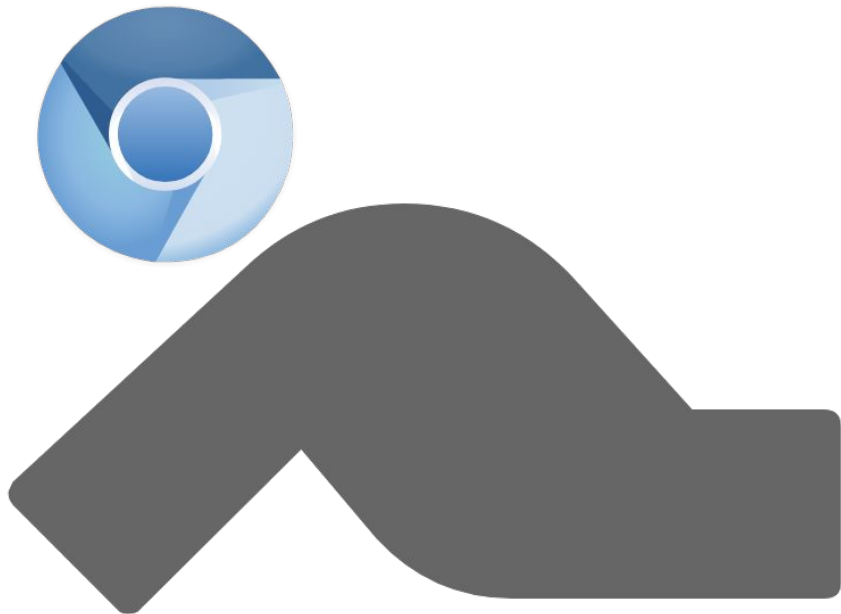
URL: <http://www.carmanmoore.com/>

Captured 8 times between Feb 12, 2014 and Dec 1, 2016

#aitpdx | @archiveitorg



# Grants & Strategic Activities



- Brozzler in production in Archive-It; 3 dozen orgs using it
- <https://github.com/internetarchive/brozzler> -- Continued improvements to A/V and SM
- NL contributors (NLA), 22 forks
- Multiple NLS testing locally
- Enhancements to warcprow
- Dozens of behaviors written for Umbra, IA browser automation tool (behaviors used by Brozzler)

# Grants & Strategic Activities

You are viewing a temporarily archived web page, collected at the request of [Karl-Franz Blumenthal](#), using [Archive-It](#). This page was captured in a test crawl on 21:37:27 Apr 24, 2017, run in the [Capture Technology Development Collection](#). If you would like this crawl to become part of your collection permanently, you will need to save it from the [crawl report](#). The information on this web page may be out of date. See [All versions](#) of this archived page.

Up next Autoplay

**ACE-CHUBB** ACE buys Chubb; Bulls Rally Around Promising Profitability  
AmigoBulls 619 views

Rockwell Returns  
Chubb 86 views

ACE AND CHUBB ARE NOW ONE  
Chubb In Asia Pacific 6,167 views

Chubb Advantage: The Value of Chubb Automobile Insurance  
Chubb 147 views

Healthcare Practice  
Chubb 302 views

MALJU JADI AGEN ASURANSI ??  
MCMY DEBY 4,363 views

**Commercial Insurance**  
Chubb  
Subscribe 1,974  
1,944 views

Add to Share More 7 5

Published on Sep 12, 2016  
Category People & Blogs  
License Standard YouTube License

You are viewing an archived web page, collected at the request of [Wisconsin Historical Society](#), using [Archive-It](#). This page was captured on 16:28:48 Nov 30, 2016, and is part of the [Wisconsin State Government](#) collection. The information on this web page may be out of date. See [All versions](#) of this archived page. [Video Metadata](#)

facebook

See more of Governor Scott Walker by logging into Facebook  
Message this Page, learn about upcoming events and more. If you don't have a Facebook account, you can create one to see more of this Page.

Sign Up Log In

Not Now

45th Governor of Wisconsin http://walker.wis.gov  
29K,710 people like this and 14,995 people follow this

You are viewing an archived web page, collected at the request of [ACE INA Holdings, Inc.](#), using [Archive-It](#). This page was captured on 19:24:16 Jan 31, 2017, and is part of the [YouTube](#) collection. The information on this web page may be out of date. See [All versions](#) of this archived page. [Video Metadata](#)

Up next Autoplay

**ACE-CHUBB** ACE buys Chubb; Bulls Rally Around Promising Profitability  
AmigoBulls 422 views

Healthcare Practice  
Chubb 218 views

Chubb Environmental Liability  
Chubb 24 views

ACE AND CHUBB ARE NOW ONE  
Chubb In Asia Pacific 5,484 views

ESIS Integrated Disability Management  
Chubb 65 views

Chubb Advantage: The Real Deal about Vacation Home Insurance  
Chubb 246 views

**Multiline Capabilities**

**Commercial Insurance**  
Chubb  
Subscribe 1,609  
1,405 views

Add to Share More 8 5

Published on Sep 12, 2016  
Category People & Blogs  
License Standard YouTube License

<https://wayback.archive-it.org/8884/20170301085056/https://111--en-0972v7f.googlevideo.com/video...aq=2&ajsignature=43AC258F91AC7E740568F4E71F95979486ACF5EA378A6A7B7D05687EED079734215A10E>

You are viewing a temporarily archived web page, collected at the request of [Karl-Franz Blumenthal](#), using [Archive-It](#). This page was captured in a test crawl on 23:18:45 Mar 22, 2017, run in the [Capture Technology Development Collection](#). If you would like this crawl to become part of your collection permanently, you will need to save it from the [crawl report](#). The information on this web page may be out of date. See [All versions](#) of this archived page.

WALKER: WE GO #WIWORKING

Like Follow Share

Photos

Government Official

45th Governor of Wisconsin http://walker.wis.gov

297,017 people like this and 265,498 people follow this

Phone: (608) 265-1212  
Walker at gov/wisworking

#aitpdx | @archiveitorg



# Grants & Strategic Activities



- **WASAPI project (Web Archiving Systems APIs)**
  - National Symposium at IA, Feb 2017
  - White paper out in Q3
  - Export API spec in production & use
  - Use cases for development purposes
  - Additional integrations from partners
  - Community models for collaborative dev
  - 2018 work?
  - **MORE TESTERS WANTED!!**

#ait16 | @archiveitorg



# Grants & Strategic Activities



Archive-It API v1.0

downloader\_eal

Api Root · Webdata Query List

## Webdata Query List

OPTIONS

GET

API endpoint that allows webdata files to be queried for and listed.

GET /wasapi/v1/webdata?crawl=297306

HTTP 200 OK

Vary: Accept

Allow: GET, POST, HEAD, OPTIONS

Content-Type: application/json

```
{
  "request-url": "https://partner.archive-it.org/wasapi/v1/webdata?crawl=297306",
  "previous": null,
  "includes-extra": false,
  "count": 5,
  "files": [
    {
      "size": 343102,
      "locations": [
        "https://warcs.archive-it.org/webdatafile/ARCHIVEIT-5425-CRAWL_SELECTED_SEEDS-JOB297306-SEED1358138-20170426175506870-00000-qfeyzrpo.warc.gz",
        "https://archive.org/download/ARCHIVEIT-5425-CRAWL_SELECTED_SEEDS-JOB297306-SEED1358138-20170426-00000/ARCHIVEIT-5425-CRAWL_SELECTED_SEEDS-JOB297306-SEED1358138-20170426175506870-00000-qfeyzrpo.warc.gz"
      ],
      "filename": "ARCHIVEIT-5425-CRAWL_SELECTED_SEEDS-JOB297306-SEED1358138-20170426175506870-00000-qfeyzrpo.warc.gz",
      "crawl": 297306,
      "checksums": {
        "md5": "326f3241484456956636f75401110f"
      }
    }
  ]
}
```

<https://partner.archive-it.org/wasapi/v1/webdata>

<https://github.com/WASAPI-Community/data-transfer-apis/tree/master/ait-specification>



#aitpdx | @archiveitorg



# Grants & Strategic Activities



README.md

build passing coverage 89% version dev

## wasapi-downloader

Java command line application to download crawls from WASAPI.

### Local Setup

You'll need the following prerequisites installed on your local computer:

- Java (7)

<https://github.com/sul-dlss/wasapi-downloader>



#aitpdx | @archiveitorg





# Grants & Strategic Activities



Empowering Public Libraries  
to Create Community History  
Web Archives

## PROGRAM PARTNERS



CLEVELAND PUBLIC LIBRARY



Queens Library



San Francisco Public Library

- IMLS-funded project to train public librarians in web archiving to preserve online local history & community memory
- Archive-It with partners Queens PL, San Francisco PL, Cleveland PL, OCLC WebJunction
- **Applications Open -- due August 25**
- 15 public libraries will participate
- 2 year project, estimated 35+ TB of local history web archives; cohort model to seed community
- Outputs: OER training materials, videos, guidebooks, cohort, etc.

<https://archive-it.org/blog/projects/community-webs/>

#aitpdx | @archiveitorg



# Research & Development

## CDX “14”

- urlkey
- timestamp
- original
- mimetype
- statuscode
- digest
- redirect
- robotflags
- length
- offset
- filename

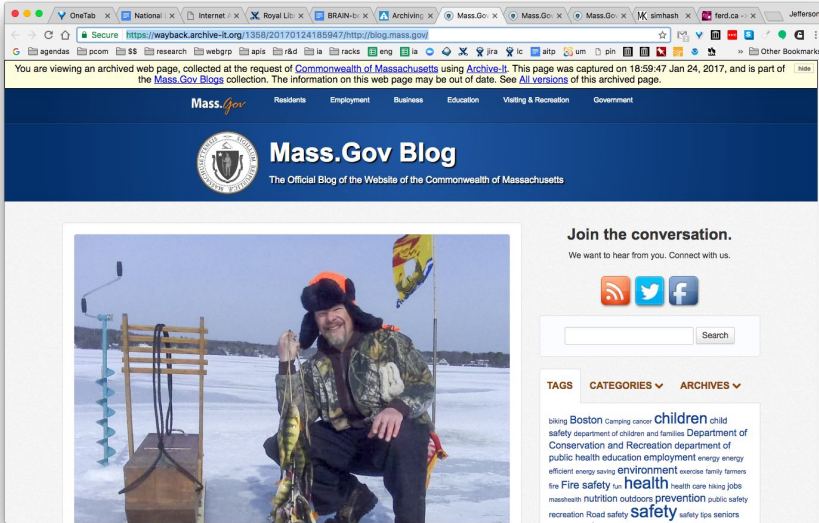
- language
- simhash
- sha256

3 added columns in new “expanded” CDX -- currently in production in Archive-It

These fields are not available for public access.

- Uses CLD2 interpreter
- Running on 1B+/month crawls, AIT and domains
- Writing to lang-dirs or CDX

# Research & Development



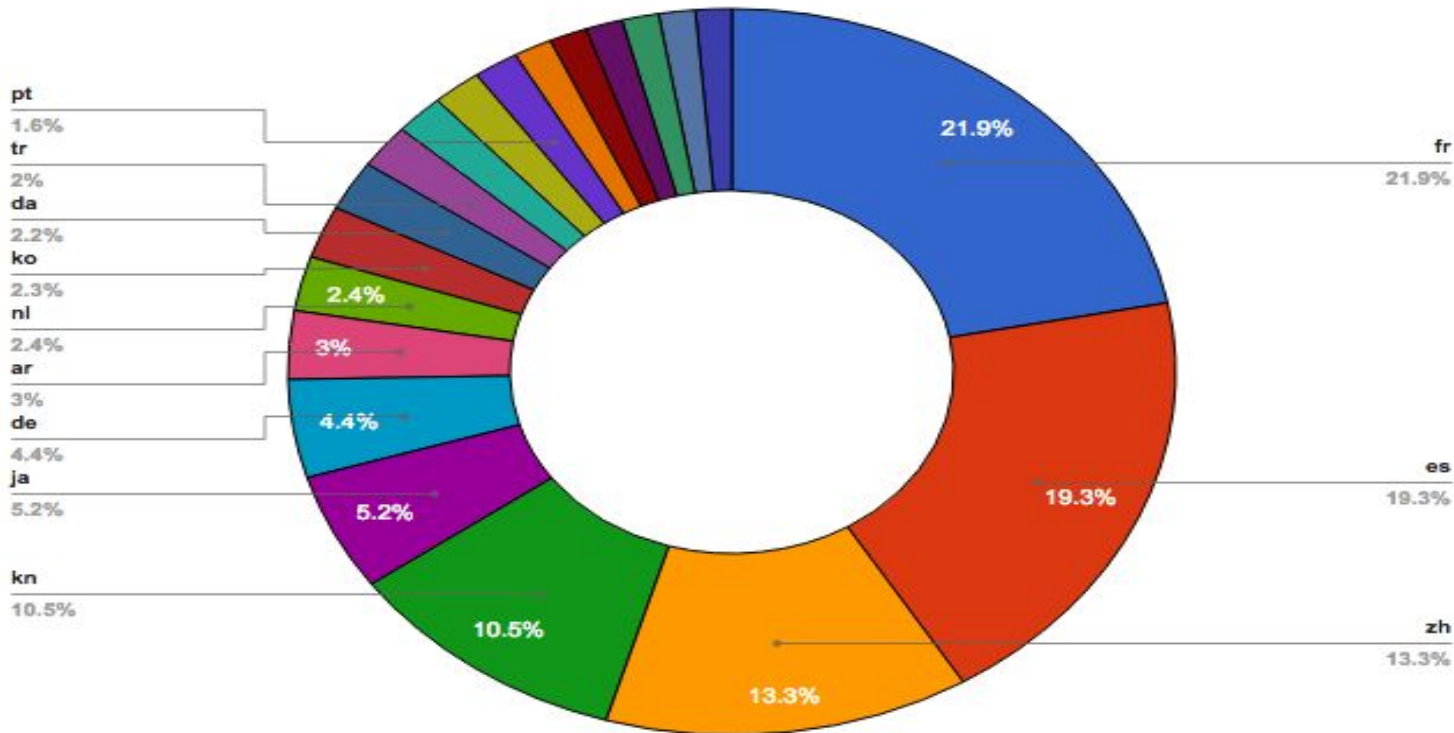
<https://wayback.archive-it.org/1358/20170124185947/http://blog.mass.gov/> (original)

<https://wayback.archive-it.org/1358/20170224190056/http://blog.mass.gov/> (89% similarity - 1 month)

<https://wayback.archive-it.org/1358/20170524190004/http://blog.mass.gov/> (78% similarity - 4 months)

# Research & Development

Top 20 non-en languages in Archive-It (Jan-May 2017)

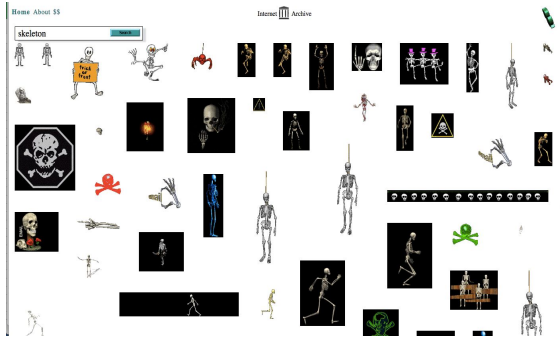
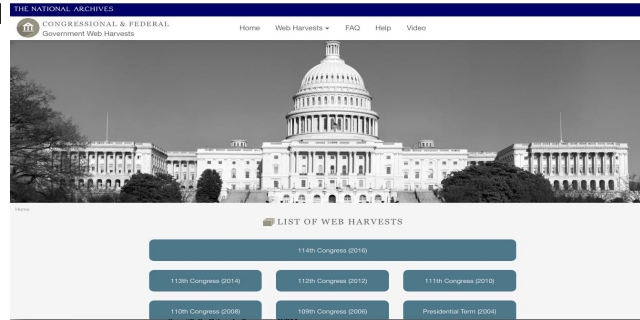
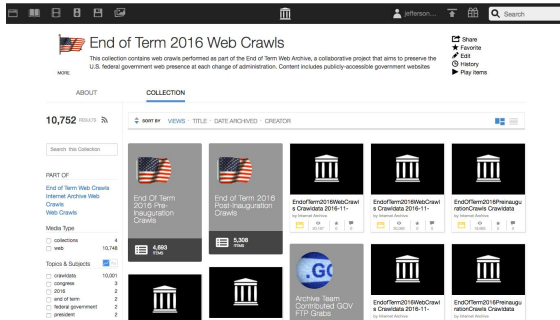


#aitpdx | @archiveitorg



# Research & Development Web Group Activities

- 250+ .gov/.mil archived
- Worked with White House to preserve Obama social media
- <https://gificities.org/> (animated GIFs from Geocities)
- <https://www.webharvest.gov/> (new web archive portal for NARA)
- Nation Wide Web portal for country-specific archive
- WAT API -- API build on extracted derivative files
- New Researcher workshops and notebooks



# AIT Special Projects Updates

- Other Events: White House Hackathon, WARCshop, SAA WA Section Webinar on APIs
- K-12 & Edu: 10 K-12 partners, 7 MLIS program partners
- Collaborative Spontaneous Events collecting; (Cuba; US election)
- Training & Support:
  - Video Curriculum
  - Special Topics Webinar Trainings
- Communications:
  - Blog posts, Collections of the Week & Quarterly Update Calls
  - More conferences and events than we can list

#aitpdx @archiveitorg



# Archive-It New Features

## Full Text Search in Elasticsearch

- Total rebuild of full-text search in Elasticsearch
- 6.5 billion docs & 52TB full-text index
- Improved relevance ranking, metadata search, performance
- Average query time (including queries across all collections) < 2 seconds!
- Seamless support of existing public site and web app FTS
- Ongoing support of OpenSearch



# Archive-It New Features

## CDX/C API

The image shows a browser window displaying a CDX/C API response. The response is a text-based format with fields separated by spaces. Colored lines and labels identify specific fields: red lines for 'urlkey', 'original', and 'login'; blue lines for 'timestamp', 'mimetype', 'digest', 'offset', 'file name', and 'length'. The first line of the response is: `com, twitter)/internetarchive 20161206224935 https://twitter.com/internetarchive/ text/html 200 L5DWB6VD575XT05QPCCKE7KEQXG4GQ56 --`

<https://support.archive-it.org/hc/en-us/articles/115001790023-Access-Archive-It-s-Wayback-index-with-the-CDX-C-API>



#aitpdx | @archiveitorg





# Archive-It New Features

## Scope Rules Refactor

- Mostly a backend change
- Scope rules are stored and accessed in a more straightforward way = faster loading
- Collection level “Host Rules” and “Expand Scope” are combined in the UI

Home / Collections / Climate Change / Scope

### Climate Change

Created: Oct 27, 2010 by system Updated: Sep 26, 2016 by system

Overview Seeds Crawls **Collection Scope** Metadata Wayback QA

You can also add scope rules at the Seed level.

#### Add Collection Scope Rule

Select Collection Scope Rule Type...

- Select Collection Scope Rule Type...
- Block Hosts
- Add Data Limit
- Add Document Limit
- Ignore Robots.txt
- Block URL If...
- Expand Scope to include URL If...

Type to Search...

Applies To	Rule	Uri Match	Value	Controls
For host akamaihd.net	ignore robots.txt			<input type="checkbox"/>
For host dotearth.blogs.nytimes.com	block entire host			<input type="checkbox"/>

Help

# Archive-It New Features

## Private Collection Pages

The screenshot displays the 'Digital Preservation' dashboard for a collection. At the top, it shows 'Created: Dec 6, 2016 by system' and 'Updated: May 11, 2017 by Alt\_Demo'. The navigation tabs include Overview, Seeds, Crawls, Collection Scope, Metadata, and Wayback QA. The 'Collection Data' section features a donut chart for 'Collection Data vs Account' showing 170.9 GB Archived, and a 'Total Data Archived' box with 170.9 GB of Data and 3,506,552 Documents. The 'Collection Settings' section has dropdown menus for 'Private' and 'Active', and a 'Save' button. A 'Private Collection Link' is provided as <https://archive-it.org/collections/74dce26a-fd52-4701-9abb-e1b05e5fd0fd>. The 'Scheduled Crawls' table is shown at the bottom.

Frequency	Active Seeds	Next Crawl	Last Crawl	Time Limit	Data Limit	Doc. Limit	Crawl Frequency
Monthly	2	Aug 7, 2017	Jul 7, 2017	3 Days	No Limit	No Limit	<a href="#">Edit Schedule &gt;</a>

<https://support.archive-it.org/hc/en-us/articles/208334003-Controlling-access-to-your-web-archives-#privatecollection>

#aitpdx | @archiveitorg



# Archive-It: in Development



Find Archive-It Media files

(Audio, Video, PDF, PPT, ...)

Support for collection (c:), host/domain (site:), mime (m:), year (y:), and sha-1 digest (k:) filters  
Example MIME filters - m:ppt, m:pdf, m:audio, m:video, m:flv, m:x-flv, m:mp3

[Find In Wayback](#)



Tools



Subscription Service

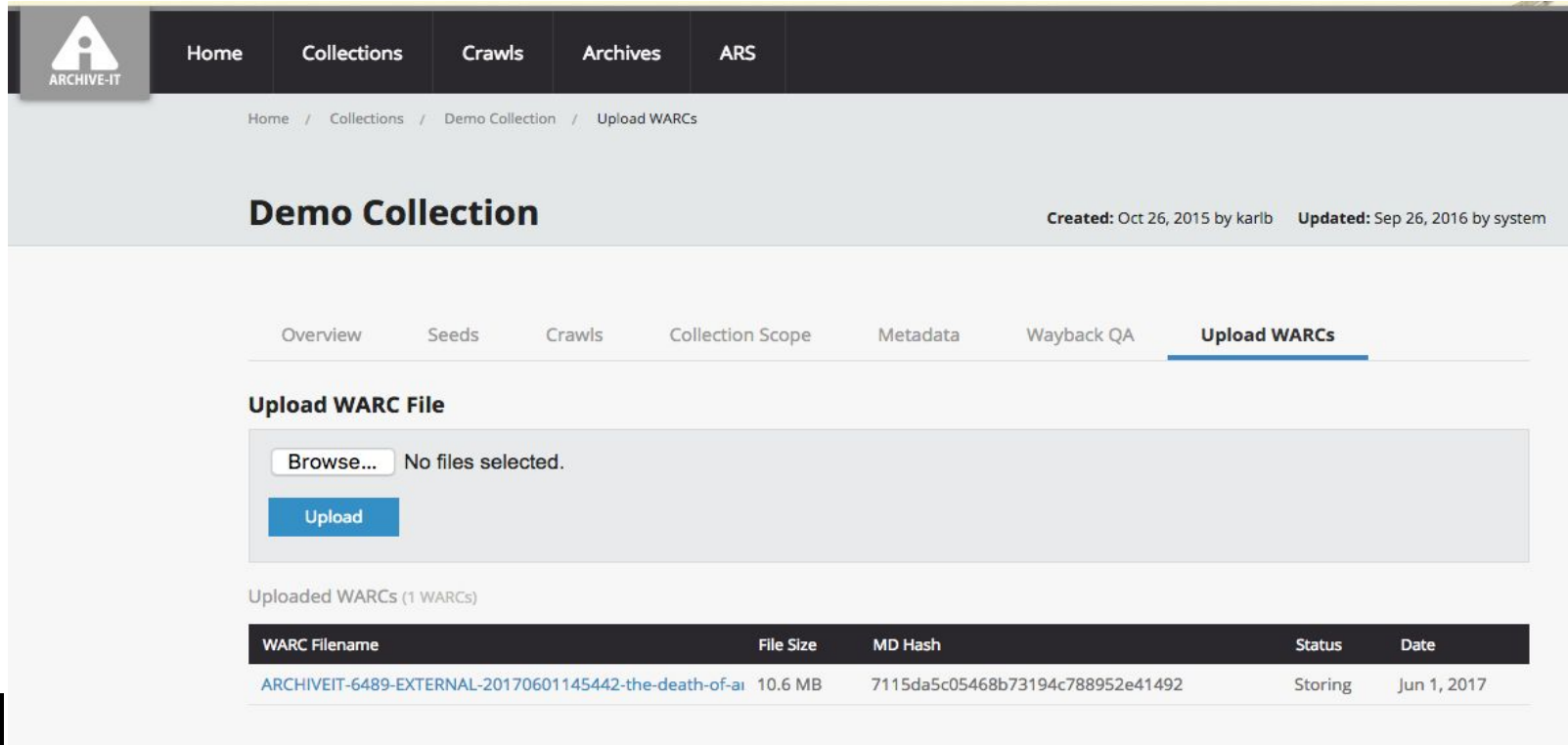


Save Page Now

#aitpdx | @archiveitorg



# Archive-It: in Development WARC Uploader



The screenshot shows the Archive-It WARC Uploader interface. At the top is a dark navigation bar with the Archive-It logo and menu items: Home, Collections, Crawls, Archives, and ARS. Below the navigation bar is a breadcrumb trail: Home / Collections / Demo Collection / Upload WARC. The main content area is titled "Demo Collection" and includes metadata: "Created: Oct 26, 2015 by karlb" and "Updated: Sep 26, 2016 by system". A horizontal menu below the title contains tabs for Overview, Seeds, Crawls, Collection Scope, Metadata, Wayback QA, and Upload WARC. The "Upload WARC" tab is selected. Under this tab, there is a section titled "Upload WARC File" containing a "Browse..." button, the text "No files selected.", and an "Upload" button. Below this is a section titled "Uploaded WARC (1 WARC)" which contains a table with the following data:

WARC Filename	File Size	MD Hash	Status	Date
<a href="#">ARCHIVEIT-6489-EXTERNAL-20170601145442-the-death-of-ai</a>	10.6 MB	7115da5c05468b73194c788952e41492	Storing	Jun 1, 2017

#aitpdx | @archiveitorg



# Archive-It: in Development



- UI Updates/Improvements (ongoing!)
- New and improved reports backend
- Operational Resiliency
- WorldCat integration



# AIT Looking Ahead

- Brozzler (for all!)
- Partner Data API
- Moving Seeds from one collection to another
- Wayback Updates: Python Wayback/Partner WB controls
- ARS

