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), 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
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
Haughton 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**

<table>
<thead>
<tr>
<th>Year</th>
<th>Partners</th>
</tr>
</thead>
<tbody>
<tr>
<td>2006</td>
<td>27</td>
</tr>
<tr>
<td>2007</td>
<td>35</td>
</tr>
<tr>
<td>2008</td>
<td>71</td>
</tr>
<tr>
<td>2009</td>
<td>102</td>
</tr>
<tr>
<td>2010</td>
<td>150</td>
</tr>
<tr>
<td>2011</td>
<td>185</td>
</tr>
<tr>
<td>2012</td>
<td>223</td>
</tr>
<tr>
<td>2013</td>
<td>276</td>
</tr>
<tr>
<td>2014</td>
<td>330</td>
</tr>
<tr>
<td>2015</td>
<td>419</td>
</tr>
<tr>
<td>2016</td>
<td>485</td>
</tr>
<tr>
<td>2017</td>
<td>558</td>
</tr>
</tbody>
</table>

#aitpdx @archiveitorg
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
AIT Community Updates

archive-it total size of all warcs (all-recorded)

#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)
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
Grants & Strategic Activities
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 warcprox
- Dozens of behaviors written for Umbra, IA browser automation tool (behaviors used by Brozzler)
Grants & Strategic Activities
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!!
Grants & Strategic Activities

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

#aitpdx | @archiveitorg
Grants & Strategic Activities

https://github.com/sul-dlss/wasapi-downloader
Grants & Strategic Activities

--

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/
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

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

These fields are not available for public access.
Research & Development

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

- fr: 21.9%
- zh: 13.3%
- es: 19.3%
- ja: 5.2%
- kn: 10.5%
- de: 4.4%
- en: 3%
- ar: 2.4%
- ni: 2.3%
- ko: 2.2%
- da: 2%
- tr: 1.6%
- pt: 1.6%
- Others: 24.6%
Research & Development
Web Group Activities

- 250+ .gov/.mil archived
- Worked with White House to preserve Obama social media
- [https://gifcities.org/](https://gifcities.org/) (animated GIFs from Geocities)
- [https://www.webharvest.gov/](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
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

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
Archive-It New Features
Private Collection Pages

https://support.archive-it.org/hc/en-us/articles/208334003-Controlling-access-to-your-web-archives-
#privatecollection
Archive-It: in Development
Archive-It: in Development
WARC Uploader

Demo Collection

Upload WARC File

Uploaded WARC File (1 WARC)

<table>
<thead>
<tr>
<th>WARC Filename</th>
<th>File Size</th>
<th>MD Hash</th>
<th>Status</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARCHIVE-IT-6489-EXTERNAL-20170601145442-the-death-of-ai</td>
<td>10.6 MB</td>
<td>7115da5c05468b73194c788952e41492</td>
<td>Storing</td>
<td>Jun 1, 2017</td>
</tr>
</tbody>
</table>
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