Preserving born digital community records

The Association for Rural & Small Libraries

Fargo, North Dakota-Moorhead, Minnesota
October 27, 2016

Maria Praetzellis, Program Manager, Web Archiving | maria@archive.org
Internet Archive & Archive-It | @internetarchive & @archiveitorg
Agenda

- The what and why of IA
- How to use IA
- Building libraries together
• We are a non-profit Digital Library & Archive founded in 1996
• We work in a former church and it’s awesome (and you are welcome to visit!)
• Extensive open-source technology development for archiving/preservation
• Team is a mix of engineers, librarians/archivists, program staff
28,000,000,000,000,000 Bytes Archived

(28 PetaBytes)
2.4 million books online
Open Library is yours to borrow, read & search.

Books to Read  The World's classic literature at your fingertips. Over 1,000,000 free ebook titles available.

Books to Borrow  Here's a sample of recently returned books from the eBook lending library.
TV Collection

1.2 million shows
900,000 hours
TV News Archive

Quotes

- Russia, this is a topic that is -- I will give you an opportunity to -- >> take them one at a time. >> get nuclear weapons. >> he never

- >>> according to a study released today by the southern poverty law center, which surveyed 2,000 teachers. Andrea

- Punched women who made heartbreaking choice to end a pregnancy. >>> why did Donald Trump say that? >> look, it's --

- Night of music and debauchery: we say welcome to the emcees of cabaret, Randy Harrison joining us on "the 9".
Political ads broadcast 283,824 times over 26 markets

The Political TV Ad Archive collects political ads in the 2016 election. In addition to tracking airings across key primary states, the collection includes ads that may air elsewhere or exclusively on social media.
Software Preservation and Emulation

25,000 software titles
20 Years of Archiving the Web

510 billion Web URLs
THE WAYBACK MACHINE

https://archive.org/web/

The largest publicly available web archive in existence.

> 510 Billion URLs
> 100 million websites
> 40 languages
> ~1 billion URLs added per week
> ~14 PB (petabytes!) of data
What is a web archive?

A web archive contains as much as possible from the original resources and documents their change over time. It is a priority to recreate the same experience a user would have had if they had visited the live site on the day it was archived.
92 days

Average life-span of a website
Free hosting, storage and access!

- Hosting is free
- Content is preserved in perpetuity
- Files automatically derived into multiple formats
- Metadata capabilities
- Institution pages and collections
Distributed Preservation
An Ethnologic Dictionary
of the
Navaho Language.

The Franciscan Fathers,
Saint Michaels, Arizona.
<table>
<thead>
<tr>
<th>File Name</th>
<th>Date</th>
<th>Time</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>ethologicnavaho00editrich.djvu</td>
<td>26-Jul-2006</td>
<td>13:28</td>
<td>17.7M</td>
</tr>
<tr>
<td>ethologicnavaho00editrich.gif</td>
<td>26-Jul-2006</td>
<td>09:56</td>
<td>338.2K</td>
</tr>
<tr>
<td>ethologicnavaho00editrich.pdf</td>
<td>07-Nov-2011</td>
<td>00:19</td>
<td>39.4M</td>
</tr>
<tr>
<td>ethologicnavaho00editrich_abbyy.gz</td>
<td>26-Jul-2006</td>
<td>11:50</td>
<td>14.2M</td>
</tr>
<tr>
<td>ethologicnavaho00editrich_archive.torrent</td>
<td>16-May-2016</td>
<td>11:16</td>
<td>18.4K</td>
</tr>
<tr>
<td>ethologicnavaho00editrich_bw.pdf</td>
<td>06-Nov-2011</td>
<td>21:58</td>
<td>29.4M</td>
</tr>
<tr>
<td>ethologicnavaho00editrich_dc.xml</td>
<td>25-Jul-2006</td>
<td>05:08</td>
<td>480.0B</td>
</tr>
<tr>
<td>ethologicnavaho00editrich_djvu.txt</td>
<td>26-Jul-2006</td>
<td>15:36</td>
<td>959.5K</td>
</tr>
<tr>
<td>ethologicnavaho00editrich_djvu.xml</td>
<td>26-Jul-2006</td>
<td>12:00</td>
<td>8.0M</td>
</tr>
<tr>
<td>ethologicnavaho00editrich_files.xml</td>
<td>16-May-2016</td>
<td>11:16</td>
<td>6.1K</td>
</tr>
<tr>
<td>ethologicnavaho00editrich_jp2.zip</td>
<td>26-Jul-2006</td>
<td>09:32</td>
<td>822.2M</td>
</tr>
<tr>
<td>ethologicnavaho00editrich_jpg.zip</td>
<td>26-Jul-2006</td>
<td>09:34</td>
<td>366.1M</td>
</tr>
<tr>
<td>ethologicnavaho00editrich_marc.xml</td>
<td>25-Jul-2006</td>
<td>05:08</td>
<td>3.6K</td>
</tr>
<tr>
<td>ethologicnavaho00editrich_meta.mrc</td>
<td>25-Jul-2006</td>
<td>05:08</td>
<td>986.0B</td>
</tr>
<tr>
<td>ethologicnavaho00editrich_meta.sqlite</td>
<td>16-May-2016</td>
<td>11:16</td>
<td>6.0K</td>
</tr>
<tr>
<td>ethologicnavaho00editrich_meta.xml</td>
<td>30-Apr-2016</td>
<td>21:11</td>
<td>2.0K</td>
</tr>
<tr>
<td>ethologicnavaho00editrich_metadata.xml</td>
<td>25-Jul-2006</td>
<td>05:08</td>
<td>501.0B</td>
</tr>
<tr>
<td>ethologicnavaho00editrich_raw_jp2.zip</td>
<td>26-Jul-2006</td>
<td>09:29</td>
<td>1.1G</td>
</tr>
<tr>
<td>scandata.zip</td>
<td>25-Jul-2006</td>
<td>21:15</td>
<td>46.0M</td>
</tr>
<tr>
<td>Archive File</td>
<td>Size (bytes)</td>
<td>MD5</td>
<td></td>
</tr>
<tr>
<td>--------------</td>
<td>--------------</td>
<td>-----------------</td>
<td></td>
</tr>
<tr>
<td>ARCHIVED-7520-CRAWL SELECTED SEEDS-J02238492-20160927085956479-00001.warc.gz</td>
<td>769565642</td>
<td>8251ae452ab850e005da5e32632e122e</td>
<td></td>
</tr>
<tr>
<td>ARCHIVED-7520-CRAWL SELECTED SEEDS-J02238506-20160927083831229-00019.warc.gz</td>
<td>1134009000</td>
<td>307df79e339746e612e6bc30300</td>
<td></td>
</tr>
<tr>
<td>ARCHIVED-7520-CRAWL SELECTED SEEDS-J0238562-20160927082431724-00017.warc.gz</td>
<td>1176211094</td>
<td>037907e0139e93113306d921580a</td>
<td></td>
</tr>
<tr>
<td>ARCHIVED-7520-CRAWL SELECTED SEEDS-J0238562-20160927083636434-00014.warc.gz</td>
<td>1260260840</td>
<td>3666e75b6535b350f6878091e25f5</td>
<td></td>
</tr>
<tr>
<td>ARCHIVED-7520-CRAWL SELECTED SEEDS-J0238562-2016092708086004-00015.warc.gz</td>
<td>1003168868</td>
<td>a8f791918c9b579974e83646a064ab</td>
<td></td>
</tr>
<tr>
<td>ARCHIVED-7520-CRAWL SELECTED SEEDS-J0238562-20160927080940867-00014.warc.gz</td>
<td>1046817082</td>
<td>05775276a12e2e2d9007c6e616e98ac</td>
<td></td>
</tr>
<tr>
<td>ARCHIVED-7520-CRAWL SELECTED SEEDS-J03823856-20160927071512619-00011.warc.gz</td>
<td>1005393023</td>
<td>4125165e921e18217b968098474d1</td>
<td></td>
</tr>
<tr>
<td>ARCHIVED-7520-CRAWL SELECTED SEEDS-J03823856-20160927074299510-00012.warc.gz</td>
<td>1099533355</td>
<td>f792a9c2e49e3a184d9e6e07c0fd</td>
<td></td>
</tr>
<tr>
<td>ARCHIVED-7520-CRAWL SELECTED SEEDS-J038238493-20160927071701638-00014.warc.gz</td>
<td>1916945133</td>
<td>83323b0893d5b939b39e7e8614f426c</td>
<td></td>
</tr>
<tr>
<td>ARCHIVED-7520-CRAWL SELECTED SEEDS-J0923856-20160927054249931-00011.warc.gz</td>
<td>1135477144</td>
<td>61480b79d0ce4d1d9a989413e75affed</td>
<td></td>
</tr>
<tr>
<td>ARCHIVED-7520-CRAWL SELECTED SEEDS-J0923856-2016092704452608-00080.warc.gz</td>
<td>1214329165</td>
<td>9b3b4357b3ee7b7e23be6b53e2dd00c</td>
<td></td>
</tr>
<tr>
<td>ARCHIVED-7520-CRAWL SELECTED SEEDS-J0923856-20160927051113501-00101.warc.gz</td>
<td>1006366555</td>
<td>b1953712927e30a37e65e29f253e5d9</td>
<td></td>
</tr>
<tr>
<td>ARCHIVED-7520-CRAWL SELECTED SEEDS-J0923856-20160927050819253-00039.warc.gz</td>
<td>1054389393</td>
<td>ec88236563e1ab12e6cfa0c2039f60</td>
<td></td>
</tr>
<tr>
<td>ARCHIVED-7520-CRAWL SELECTED SEEDS-J0923856-20160927050625788-00083.warc.gz</td>
<td>1081777753</td>
<td>18783526220236461edf0a406d3677</td>
<td></td>
</tr>
<tr>
<td>ARCHIVED-7520-CRAWL SELECTED SEEDS-J0923856-20160927052780044-00018.warc.gz</td>
<td>1003141217</td>
<td>237078a48618016c62629fa9eaed79d</td>
<td></td>
</tr>
<tr>
<td>ARCHIVED-7520-CRAWL SELECTED SEEDS-J0923856-20160927055357156-00017.warc.gz</td>
<td>1002162783</td>
<td>38811f50be17a4229e8787866d5f7d</td>
<td></td>
</tr>
<tr>
<td>ARCHIVED-7520-CRAWL SELECTED SEEDS-J0923856-20160927050334303-00027.warc.gz</td>
<td>683621504</td>
<td>13362b65e16e002e74ece728561e</td>
<td></td>
</tr>
<tr>
<td>ARCHIVED-7520-CRAWL SELECTED SEEDS-J0923856-201609270540312106-00036.warc.gz</td>
<td>10466614560</td>
<td>5934c37e905b8e79288e313557</td>
<td></td>
</tr>
<tr>
<td>ARCHIVED-7520-CRAWL SELECTED SEEDS-J0923856-20160927041313456-00026.warc.gz</td>
<td>1204629363</td>
<td>54263c2e7390f0b3b3db3fede872870fd</td>
<td></td>
</tr>
<tr>
<td>ARCHIVED-7520-CRAWL SELECTED SEEDS-J0923856-20160927040247028-00026.warc.gz</td>
<td>1027104514</td>
<td>0ab20b43b6f0524b2e6e000606</td>
<td></td>
</tr>
<tr>
<td>ARCHIVED-7520-CRAWL SELECTED SEEDS-J09238499-201609270346046-00020.warc.gz</td>
<td>1144013130</td>
<td>0b4603669e0c42a6042ac41020a0a4</td>
<td></td>
</tr>
<tr>
<td>ARCHIVED-7520-CRAWL SELECTED SEEDS-J09238506-201609270346046-00020.warc.gz</td>
<td>7855668576</td>
<td>0ba8b70277ba7e1a52b0b7e1b12d96dc</td>
<td></td>
</tr>
<tr>
<td>ARCHIVED-7520-CRAWL SELECTED SEEDS-J09238506-20160927034500471-00019.warc.gz</td>
<td>173069308</td>
<td>169f3d350c65d0b66a9a797a825921df</td>
<td></td>
</tr>
<tr>
<td>ARCHIVED-7520-CRAWL SELECTED SEEDS-J09238506-20160927035828972-00007.warc.gz</td>
<td>1316033557</td>
<td>59a4a68e64b9a04a202a1d02a0a4</td>
<td></td>
</tr>
<tr>
<td>ARCHIVED-7520-CRAWL SELECTED SEEDS-J09238506-2016092703185698-00034.warc.gz</td>
<td>1005214927</td>
<td>401f8f149ebab83d0d5ad45e5366f3f3</td>
<td></td>
</tr>
</tbody>
</table>
Local and Global Access
etree.org is a community committed to providing the highest quality live concerts in a lossless, downloadable format. The Internet Archive has teamed up with etree.org to preserve and archive as many live concerts as possible for current and future generations to enjoy. All music in this Collection is from trade-friendly artists and is strictly non-commercial, both for access here and for any further distribution. Artists' commercial releases are off-limits. This collection is maintained by the etree.org community.

Browsing and Finding shows

Search the Entire LMA
Browse All Artists with Recordings in the Live Music Archive
Browse All Artists' Collection Pages in the Live Music Archive (alternate view)
Most Recent Band Activations
Graph of the Number of Shows Added Per Month (For the past 24 months)
Most Recently Added Items
Browse Artists with MP3's
Search for Lossless-Only Recordings (No-MP3)
Web archiving

Save Page Now

Archive-It
Wayback Machine

510 billion web pages saved over time.

Tools

Wayback Machine Availability API
Build your own tools.

WordPress Broken Link Checker
Banish broken links from your blog.

404 Handler for Webmasters
Help users get where they were going.

Subscription Service

Archive-It enables you to capture, manage and search collections of digital content without any technical expertise or hosting facilities. Visit Archive-It to build and browse the collections.

Save Page Now

Capture a web page as it appears now for use as a trusted citation in the future. Only available for sites that allow crawlers.
Loading Saved Page...

http://www.atalm.org/node/1

As it appears live October 7, 2016 9:28:47 PM UTC

INTERNET ARCHIVE wayBackMachine
http://atalm.org

Saved 64 times between December 13, 2010 and October 7, 2016.

PLEASE DONATE TODAY. Your generosity preserves knowledge for future generations. Thank you.
Archive-It

- A web-based application launched in 2006 that allows users to create, manage, access and store collections of web-based digital content.
- A fully hosted solution, including access and storage.
- A suite of tools for selecting and scoping, and cataloging.
- Archived web content includes: html, text, videos, audio, social media, PDF, images, password protected content, static databases and newspapers.
- Browse archived content 24 hours after a capture is complete; full text search is available within 7 days.
How our partners use Archive-It

- Enhance and supplement traditional offline collections
- Support records retention policies
- Capture "at risk" digital content on spontaneous events
- Preserve a website before it goes offline
Content includes:

> Full-text archives of official documents
> Original video and audio recordings of key regional leaders
> Thousands of annual and "state of the nation" reports
> Collections of Latin American elections and political parties
You are viewing an archived web page, collected at the request of University of Texas at Austin Libraries, Latin American Government Documents Archive using Archive-It. This page was captured on 22:37:09 Dec 15, 2008, and is part of the Latin American Government Documents Archive, LAGDA collection. The information on this web page may be out of date. See All versions of this archived page.
Honduras presidential website, 2010 (after coup)
And now for a short demo...
Future collaboration possibilities
Discussion

- Questions on IA / AIT?
- Other questions?
- Needs and use ideas?
- Collaboration ideas?
- Other Partnerships?
Thank you!

Maria Praetzelllis, Program Manager, Web Archiving | maria@archive.org
Internet Archive & Archive-It | @internetarchive & @archiveitorg