What is web archiving?
Why web archive?

- To provide continuing access to web-based material
  - Whole domain crawls (national, institutional etc)
  - Selective archiving

- To enhance collection development
  - Addressing issues of link rot
  - Filling gaps in collection development or enhancing existing areas

- To document events, topics, and regions
  - End of Term collections
  - #blacklivesmatter
  - Charlie Hedbo
PAGE NOT FOUND
DONALD J. TRUMP STATEMENT ON PREVENTING MUSLIM IMMIGRATION

(New York, NY) December 7th, 2015. -- Donald J. Trump is calling for a total and complete shutdown of Muslims entering the United States until our country’s representatives can figure out what is going on. According to Pew Research, among others, there is great hatred towards Americans by large segments of the Muslim population. Most recently, a poll from the Center for Security Policy released data showing “25% of those polled agreed that violence against Americans here in the United States is justified as a part of the global jihad” and 51% of those polled, “agreed that Muslims in America should have the choice of being governed according to Shariah.” Shariah authorizes such atrocities as murder against non-believers who won’t convert, beheadings and more unthinkable acts that pose great harm to Americans, especially women.
Shaping web archives

- **What** to collect?
  - Collection development policy
  - Collaborative archiving

- **When** to do web archiving?
  - Frequency
  - Timeliness

- **Whose** web are we archiving?
  - Ethical and privacy considerations
  - Copyright

More web archives are better
### Some uses for web archives

<table>
<thead>
<tr>
<th>Teaching</th>
<th>Research</th>
<th>Journalism</th>
</tr>
</thead>
<tbody>
<tr>
<td>Archive-It K-12 program</td>
<td>Historians and political scientists</td>
<td>Understanding Media Ecosystem</td>
</tr>
<tr>
<td>iSchool classes - DH, appraisal, digital art curation</td>
<td>Discovery/Legal Issues</td>
<td>Community News Archives</td>
</tr>
<tr>
<td></td>
<td>Computer Science: information retrieval, search algorithms</td>
<td></td>
</tr>
</tbody>
</table>
Working with web archives

- Wayback Machine
- APIs (Wayback/Archive-it-there are ten different API’s)
- Kibana (Visualization tool)
- WarcBase (analysis and visualization)

Challenges

- Training: https://github.com/vinaygoel/ars-workshop
How do we web archive?
### Some of our approaches

<table>
<thead>
<tr>
<th>umd.edu</th>
<th>Organizational websites</th>
<th>Capture on request</th>
</tr>
</thead>
<tbody>
<tr>
<td>Semi-annual crawl in Archive-It</td>
<td>Capture websites of external organizations whose records SCUA or SCPA holds</td>
<td>Capture pages or websites in response to requests, inquiry from organizations or campus departments</td>
</tr>
<tr>
<td>Capture whole domain</td>
<td>● AFL-CIO</td>
<td>● Faculty Voice</td>
</tr>
<tr>
<td></td>
<td>● Preservation Maryland</td>
<td>● Middle States Self Study</td>
</tr>
<tr>
<td></td>
<td>● Liz Lerman Dance Exchange</td>
<td>● Democracy Then &amp; Now</td>
</tr>
<tr>
<td></td>
<td>● Maryland NOW</td>
<td></td>
</tr>
</tbody>
</table>
Some of our approaches

<table>
<thead>
<tr>
<th>Documenting topics</th>
<th>Documenting individuals and organizations</th>
<th>Documenting events</th>
</tr>
</thead>
<tbody>
<tr>
<td>Capture websites that fill out the landscape of a collection area</td>
<td>Capture pages or websites related to a specific individual or organization represented in SCUA collections</td>
<td>Capture pages and websites documenting events that resonate or take place on campus</td>
</tr>
<tr>
<td>● North American labor organizations</td>
<td>● Filipino American Community Archives</td>
<td>● #feartheturtle</td>
</tr>
<tr>
<td>● UMD Greek life</td>
<td>● Arthur Godfrey</td>
<td>● Richard W. Collins III</td>
</tr>
<tr>
<td></td>
<td></td>
<td>● Concussion research ethics scandal</td>
</tr>
</tbody>
</table>
Tools we use

Archive-It (https://archive-it.org/organizations/408)

- Internet Archive subscription service for collection management and access.
- Heritrix - open-source, archival quality web crawler
- Start with seed URL → follow trail of hyperlinks → capture code page by page
- Adjust scope: include or exclude specific URLs, match patterns, doc/data limits

Archive-It crawls are hosted, publicly available, indexed for metadata and full-text search via Wayback Machine; or downloaded locally
Tools we use

WebRecorder ([https://webrecorder.io/](https://webrecorder.io/))

- Rhizome ([http://rhizome.org/](http://rhizome.org/)) tool for capturing dynamic web content
- Records WARC files for playback in browser
- Open a page → interact with the site → save and describe content

WARC files can be hosted & accessible via WebRecorder or downloaded for local storage & ingest

e.g. [https://webrecorder.io/amelish](https://webrecorder.io/amelish)
Current projects

Describe our web archives

- Archive-It Dublin Core ↔ DACS (ArchivesSpace) ↔ MARC
- Contribute feedback to OCLC publication on web archives metadata

Update documentation

- How we use Archive-It 5.0
- How we use WebRecorder
- How we write description
How can we do web archiving better?
New project ideas

Public-facing documentation

- Outreach & promotion: “Who knows that you have web archives?”

Collaboration

- Who else is collecting web-based material? (http://www.docnow.io/ !)
- Seed nomination form: https://goo.gl/YkEbRL
New project ideas

Communicate with website creators


User testing for usability & accessibility
New project ideas

Guide to using web archives for librarians & library users

- What sources of web archives are available at UMD and beyond?
- What tools exist for working with web archives?
- What to be aware of when searching or using web archives?
- What ethical and intellectual property issues are involved?
- How to read a collection development policy
- How to read web archives description
Questions for you:

How do you see web archives playing a role in your work?

What would you like to be able to see or do with web archives?

What web archives would be relevant to the library users you know best?

What would you need to know in order to work with web archives?
Resources

International Internet Preservation Consortium (IIPC) - http://netpreserve.org/

Internet Archive - https://archive.org/

Rhizome - http://oldweb.today/, other projects

Arquivo.pt - http://arquivo.pt/

Library of Congress - https://www.loc.gov/webarchiving/

NARA (e.g. https://clinton1.nara.gov/ & https://obamawhitehouse.archives.gov/)
Resources

List of public web archives: https://github.com/webrecorder/public-web-archives

List of web archiving tools:
http://netpreserve.org/web-archiving/tools-and-software

- OpenWayback - https://github.com/iipc/openwayback/wiki
- Perma.cc - https://perma.cc/
- WebRecorder - https://webrecorder.io/
- Memento / Time Travel - http://timetravel.mementoweb.org/