INTERFERENCE ARCHIVE

Web Crawl Project

Digital Preservation 2014, MIAP
Allie Whalen, Blake McDowell, Genevieve Havemeyer, Karl McCool
The Library
Mission – Structure – Collections

- Volunteer-run library and archive collecting five decades of activism history
- Based in Brooklyn
- Opened December of 2011
- Hosts exhibitions, lectures, workshops, screenings, and performances
- The archive maintains a collection of media, objects, and ephemera from social movements both national and international
- Includes posters, flyers, zines, books, comics, photographs, moving images, audio recordings, T-shirts, and buttons.
- Run by a core collective of ten members

“We consider the use of our collection to be a way of preserving and honoring histories and material culture that is often marginalized in mainstream institutions. As an archive from below, we are a collectively run space that is people powered, with open stacks and accessibility for all.”
The Web Crawl
Finding a curator

• Project was shaped by the collectively run nature of the organization.
• Received input from:
  – Our official curator, Bonnie Gordon
  – Fellow core collective member, Molly Fair
  – Interference born-digital working group member, Anna Perricci
• Different collective members had different knowledge to offer about what Interference-associated web content needed archiving and what did not.
• Gathered answers and information from several people associated with Interference Archive
The Web Crawl

Finding a project
Exploring communications between libraries and ongoing web crawls

Initial plan:
use Archive-It to capture the website for Justseeds Artists’ Cooperative, a separate collective associated with Interference Archive

However...

- Tamiment Library is already archiving the Justseeds website, which the Justseeds group was unaware of
- Tamiment Library policy doesn’t inform organizations that their websites are being archived by the library
- Although Nicole, Bonnie, and Molly all know each other, they didn’t share all the same knowledge about ongoing web archiving and the web archiving needs of Interference Archive and associated websites like Justseeds
- The nonhierarchical, collectivist nature of Interference Archive meant that our group needed to communicate with more than just one member of the collective
The Web Crawl

Our Project:
Archive the Interference Archive website and their social media sites (Blog, Twitter, Facebook, and Vimeo)
The Web Crawl

Our Project:
Archive the Interference Archive website and their social media sites (Blog, Twitter, Facebook, and Vimeo)

Planning and Scoping the Crawl:

• Parameters for crawl were outlined in email correspondence with members of Interference Archive’s Born-Digital Working Group
• Evaluated high priority aspects of website, and what, if anything, was already being archived
• Met with Nicole Greenhouse for more information on status of Tamiment’s crawl of the Justseeds website
• Evaluated Tamiment recommendations to focus on social media and video content, like Vimeo, which they aren’t able to capture.
• Focus just on IA, not including their partner organizations, in order to limit the amount of content captured for this project.
• Good example for IA for future archiving project for born digital works made in collaboration with other organizations.
The Web Crawl
Our Project:
Archive the Interference Archive website and their social media sites (Blog, Twitter, Facebook, and Vimeo)

The Interference Archive Website:
• Well organized
• Contains digital representations of several publications
• Home for the IA Blog, where unique posts are made weekly about ongoing activities and projects.
• A lot of new content and announcements are shared on IA’s social media websites, including links to events and content from partner programs (like Justseeds).

Focus on Social Media:
• Document IA’s participation in the online social movement of digital platforms for political and social activism
• Include social media crawls, in addition to capturing the entire Interference Archive website, because IA’s Blog posts are less frequent than their active Twitter feed.
• Include single crawls of the Justseeds blog and Twitter feed to provide IA and Justseeds with a reference for any differences in how Archive-It captures content, since Tamiment does not capture social media, and there may be link to social media sites via their blog.
The Web Crawl
Technological Implementation

Scope
Original scope of each crawl:
- Interference Archive's website
- Twitter, Vimeo and Facebook pages

Seed URLs
http://interferencearchive.org/
https://twitter.com/InterferenceArc/
http://vimeo.com/81646906
https://www.facebook.com/InterferenceArc/

Frequency

<table>
<thead>
<tr>
<th>Daily</th>
<th>Once a month</th>
<th>Twice a month</th>
</tr>
</thead>
<tbody>
<tr>
<td>Most frequent</td>
<td>Only 1 video</td>
<td>Only 1 video</td>
</tr>
<tr>
<td>Most updated</td>
<td>Reasonable frequency to capture channel page in case usage increases</td>
<td></td>
</tr>
</tbody>
</table>

Twitter: 63,000 URLs per crawl vs. IA Website: 2,000 URLs per crawl

Additional Crawls: Justseeds
- Twitter (One-Time) https://twitter.com/Justseeds/
- Blog (One-Time) http://justseeds.org/blog/
The Web Crawl

Technological Problems and Crawl Analysis

Frequency

• Running most of them far too frequently.
• IA does not have a fast changing website.
• Updates only a few times a month.
• No listing of weekly events or a constant flow of posts on the blog.
The Web Crawl
Technological Problems and Crawl Analysis

Vimeo

- Archive-It’s lack of video capture
- Archive-It does not support capture of videos from a user’s channel page.
- Videos can be captured from their specific URL.
- No supported playback of videos in archived pages viewed on Wayback Machine, but they can be viewed in the Video Reports for each crawl.
- No public with access to archived videos
The Web Crawl
Technological Problems and Crawl Analysis

Twitter

- Twitter appears to be a very friendly URL to crawl.
- Large amount of crawls (63,000) as a result of crawling the IA Twitter account in multiple different languages.
- Archiving all of retweets by different users of any IA tweet.
- “Sign in” Box: When clicking on some of the links on the archived pages the Twitter sign in box pops up.
Facebook

- Initial test and subsequent crawls revealed that the scope of the facebook crawl needed to be adjusted in order for it to be captured.
- We added host constraints to block robots.txt files on:
  - Facebook
  - fbcdn.net
  - akamaihd.net
- To ensure that content would display properly. Archive-It states that crawls of Facebook do not always display well when archived, and that many features are limited, such as the ability to expand comments and capture every ‘like’.
Our Crawls

https://www.archive-it.org/collections/4932