Web Crawling Documentation/Notes

Albany Student Press

- https://www.albanystudentpress.online/
- Wix site scoping rules added via Archive-It help page
- Crawl is breaking and we are not currently getting this site
- Karl said it may be inside engineering issues and not the Wix site– if it’s a priority, they can work on fixing it

Capitol Connection

- Times Union Archive-it Collection
- Weekly one page crawls to get new posts for the week
- scoped in Archives because we weren’t capturing those–also didn’t get new URLs to load at the bottom of the page when we clicked the “More” button–Greg solved with scraping those URLs
- Some one time seeds to capture back articles
- added 664 scraped URLs as seeds to capture URLs in all the blog archives
- scoped out a lot of extra stuff to exclude because we were getting a huge amount of extra MB of data from stuff like Taboola/Aikhimid.net /cloudfront/permutative.app/etc.—also limited Twitter to 100 docs

CSEA

- 2 seeds: bimonthly with Google video blocked (blocked video because it’s larger ) & monthly including Google video (lots of video, but video does not update often)
- in October 2020, Brozzler crawls would duplicate URLs–wrote a help ticket, and they said they didn’t know why it was happening, but it is something they are working on fixing that issue–still not resolved by 4/2021, but we have managed ways to ignore those issues a bit

Empire Center

- 2 seeds: 1 with video and 1 without video

IUE/CWA

- 12/2020: lots of Guardian links, links to other UK websites, and Dublin Core metadata citation

National Coalition to Abolish the Death Penalty

- blog updated about once/month or every other month, but quarterly schedule saves us 6.23 GB/year

NYCLU

- avoid Brozzler for this crawl because it’s such a big site and there’s a problem of duplicated URLs

NYS Democratic Committee

- not getting the embedded Issuu doc "Party Rules" or the YouTube video under "Social" tab
- Archive-It help page has some recommendations for scoping e.issuu.com docs like removing trailing slash and running Brozzler crawl
- need to use Proxy Mode to view the Issuu doc

NYS Right to Life

- recrawled with robots.txt exclusion

NYS Social Workers
• scoped out calendar crawler traps and events
  • some events not captured

SUNY Central Administration
  • all Board of Trustees meetings PDFs--updated between 1-3x per month depending on when meetings are scheduled (also archived by suny.edu in video form)

Times Union Blogs
  • created 3 seeds: 1 for the general blog page and 2 for Capitol Connection (in above collection)
  • problems getting archived blog posts from the drop-down menu
  • scoped out Google video

UAlbany COVID-19
  • put in Archive-It help ticket because the links to the videos are in the March-April 2020 captures on the general Wayback, but not in our captures
  • needed to go into general Wayback Machine from March-April 2020 to get videos that were captured during that time but not crawled in 2021
  • Youtube URLs added as seeds in general UAlbany collection and crawled 1 time
  • working on creating collection in ArchivesSpace

UAlbany News
  • should we crawl Instagram?
  • didn’t get Twitter feed in test crawls
  • didn’t get “Filter” search bar in test crawls
  • some pics not captured and embedded YouTube videos not loading

UAlbany Sports
  • ignored robots.txt at collection level
  • a lot of GB of video

UAlbany Website

WAMC
  • 2 seeds: daily 1 page + seed & monthly seed
  • Scoped 1 page+
  • should we talk to Mark because he gets the audio (that we capture on the website) via another format?

Scholar’s Archive Collections

Campaign Finance Institute (CFI)
  • 2 URLs: www.cfinst.org & https://cfinst.github.io
  • .org wasn’t crawled well, so turned off scoping rule of github and ran crawl again
  • captured .org website, but github needed to be captured with Conifer–still problems with capture with Conifer

No Gun Ri
  • captured the Omeka site, but couldn’t get the interactive Google map–captured a screenshot of the Google map with Conifer
  • Saved site capture March 17, 2021: https://wayback.archive-it.org/12932/20210317014827/https://nogunri.rit.albany.edu/archive/