Streamlined Scoping at North Carolina

Kathleen Kenney

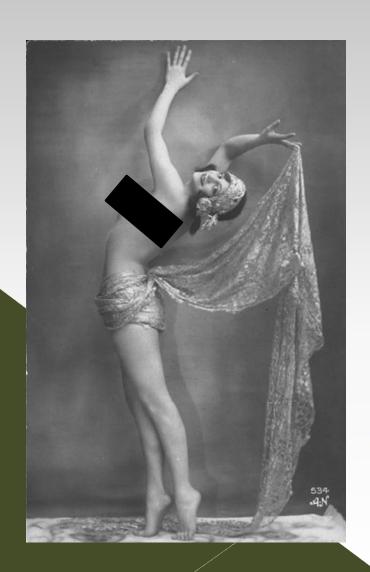
Background

- Mandated by General Assembly to preserve the public record
- Collaboration with NC State Archives
- Partner since September 2005
- O Documents Crawled: 82,539,798
- Data Archived: 6,273.6 GB
- O Total Active Seeds: 489



What is the issue?

- Capturing over 5,000 hosts per crawl
- Eliminate the capture of inappropriate content
- Duty to manage crawl budget
- Determine the most efficient way to eliminate out-of-scope hosts



Benefits of crawl analysis

- Constrain out-of-scope hosts
- Finding new seeds
- Don't want to constrain completely but don't have a budget for adding them as seeds
 - > Local government
 - > State schools, UNC, NCSU
- Possible use by state agencies in the future
 - > Social media
 - > Image servers
 - > Photo sharing



Photo credit: cute-n-tiny.com

- A2 You might want to mention why we do the analysis to:
 - 1) As gov't agency we don't want to include sites that are inappropriate for us to be archiving (either because of their content, copyright issues, or just because they are not related to the gov't)
 - 2) Duty to manage crawl budget to make sure we aren't spending a significant portion of our budget on out of scope materials.

 Author, 10/14/2011
- May want to break this slide into two slides. One about the fact that we do crawl analysis and why and the fact that this created an opportunity for streamlining. And one about benefits of having the analysis process.

 Author, 10/14/2011

The Way We Were

- 1. Download crawl reports
- 2. Import into Excel
- 3. Compare list against previously reviewed urls and delete duplicates
- 4. View each remaining web site to determine if it is in- or out- of scope (3,000+)
- 5. Make indication in Excel to constrain or leave unconstrained
- 6. Batch load constrained URLs to Host Constraints page in Partner Admin tool



A5 Didn't you also streamline 3. using Access? You may want to mention that too.

 $\mbox{I'd leave out the robots part.}$ That will just confuse folks I think. Author, 10/14/2011

North Carolina's Process	Scope-It
Must export host report into another tool	Keeps analysis activity internal to Archive-It interface
Requires bulk load of hosts to be constrained at end of process	Allows for immediate updating of host rules
Can sort host report and organize in any way	Only alpha sorting functionality
	Navigation is difficult if have thousands of hosts (which NC does)
Can identify host as in or out of scope (so only review host once)	Can only identify host as out of scope (leads to re-review same in scope hosts in next crawl)
Provides a visual of host with a link	Provides no visual of host or link (must cut and paste URL into browser to see host site)

Then

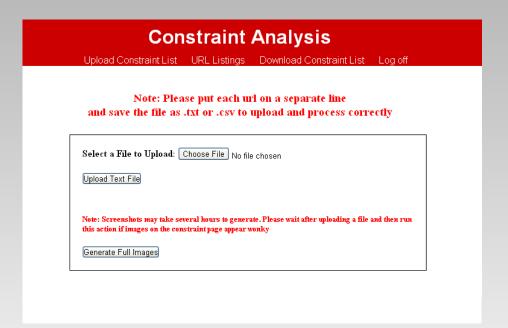
- Download crawl reports
- 2. Import into Excel
- 3. Compare list against previously reviewed urls and delete duplicates
- 4. View each remaining web site to determine if it is in- or out- of scope (3,000+)
- 5. Make indication in Excel to constrain or leave unconstrained
- 6. Batch load constrained URLs to Host Constraints page in Partner Admin tool

Now

- Download crawl reports
- Import into Access and eliminate redundancies with reviewed list by using queries
- Select sites to constrain using Constraint Analysis tool
- 4. Batch load constrained
 URLs to Host Constraints
 page in Partner Admin tool

A1 Didn't you also streamline 3. using Access? You may want to mention that too.

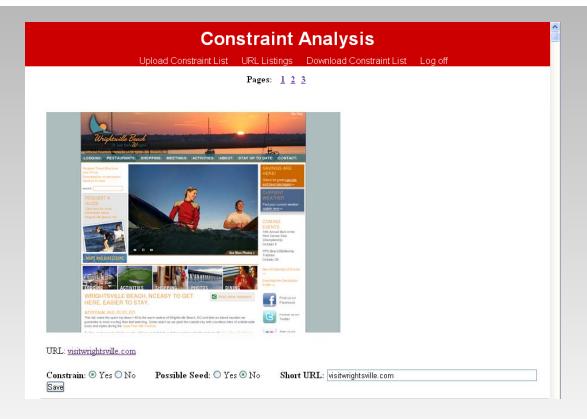
 $\mbox{I'd leave out the robots part.}$ That will just confuse folks I think. Author, 10/14/2011



- Upload a .txt or .csv file with one URL on each line
- A request is sent to a free 3rd party screen scraper service http://wimg.ca, which generates a .png image of the home page.

I'd make it really, really simple for folks by spelling out exactly what they will do and what they will then see. You should know the details in case someone asks (and you can put them on slides if you want), but there probably won't be more than a couple of techie folks in the room.

Author, 10/14/2011



URL listings page

- The links and home page images are shown 100 per page.
- User can select Constrain Yes/No, Possible Seed Yes/No or Shorten the url.
- Click Save if any changes are made.
- After all pages have been reviewed, click "Download Constraint List."

You might want to then show them what the downloaded url list looks like as well. $_{\rm Author,\ 10/14/2011}$ **A8**

constraint_analysis_20110922_0toB - Notepad File Edit Format View Help URL Constrain Seed O.c-ima.com No 001.c-ó-u-n-t.com Yes No 030b577.netsolhost.com No No c-imq.com No No 101d.com Yes No 101distribution.com Yes No. 1040bookkeepingstore.com Yes No 1040taxbiz.com Yes No 123siqnup.com Yes No 124marketingsystem.com Yes No 149.168.102.247 No No 150.216.68.249 No No 152.1.166.29 No No 152.2.63.68 No No 160tracker.com Yes No 173.201.253.58 Yes No 174.120.151.9 No No 184.72.101.167 Yes No 192.41.48.113 No No 198.170.111.238 Yes No 1dental.com Yes No 1monavie.blogspot.com Yes No 2010.bloggies.com Yes No 2011.bloggies.com Yes No 2ndfridayartwalk.com Yes No c-ima.com No 3066411264188899115-a-1802744773732722657-s-sites. Yes No 30for30.espn.com Yes No 31daysalute.com Yes No 31daýsofqlory.therepublik.net Yes No 3432.voxcdn.com Yes No 3d-qlasses-4u.com Yes No 48hoursfestival.com No Yes 4info.net Yes No 4iniuries.com Yes No 4kidstv.com No Yes 50.22.69.226 Yes No

50statesforgood.com

Yes

No

Contact

Find source code at gethub.com

- o kathleen.kenney@ncdcr.gov
- o dean.farrell@ncdcr.gov