Web Archiving Legacy.com: A Case Study

Nicholas Taylor
@nullhandle

History and New Media
February 11, 2013

Google Images: "site:legacy.com"

challenges

- legal
  - copyright
  - robots.txt
- technical
  - scale
  - robots.txt
  - scope

U.S. Copyright Office: "Copyright Act (Title 17, Chapter 1)"

copyright law

fair use

- **the purpose and character of the use**, including whether such use is of a commercial nature or is for **nonprofit educational purposes**
- **the nature of the copyrighted work**
- **the amount and substantiality of the portion used** in relation to the copyrighted work as a whole
- **the effect of the use upon the potential market** for or value of the copyrighted work

ARL Code of Best Practices

- “It is fair use for libraries to develop and facilitate the development of digital databases of collection items to enable **nonconsumptive analysis** across the collection for both scholarly and reference purposes.”
- “It is fair use to **create topically based collections of websites** and other material from the Internet and to make them available for scholarly use.”

ARL: Code of Best Practices in Fair Use for Academic and Research Libraries

robots.txt legal considerations

- unreliable proxy for copyright permissions
- archival crawler ≠ search crawler
- case law suggests it's not legally binding, but has legal value

```
User-Agent: *
Disallow: /music?
Disallow: /widgets/radio?

Disallow: /affiliate/
Disallow: /affiliate_redirect.php
Disallow: /affiliate_sendto.php
Disallow: /affiliatelink.php
Disallow: /campaignlink.php
Disallow: /delivery.php

Disallow: /music/+noredirect/

Disallow: /harming/humans
Disallow: /ignoring/human/orders
Disallow: /harm/to/self

Allow: /
```

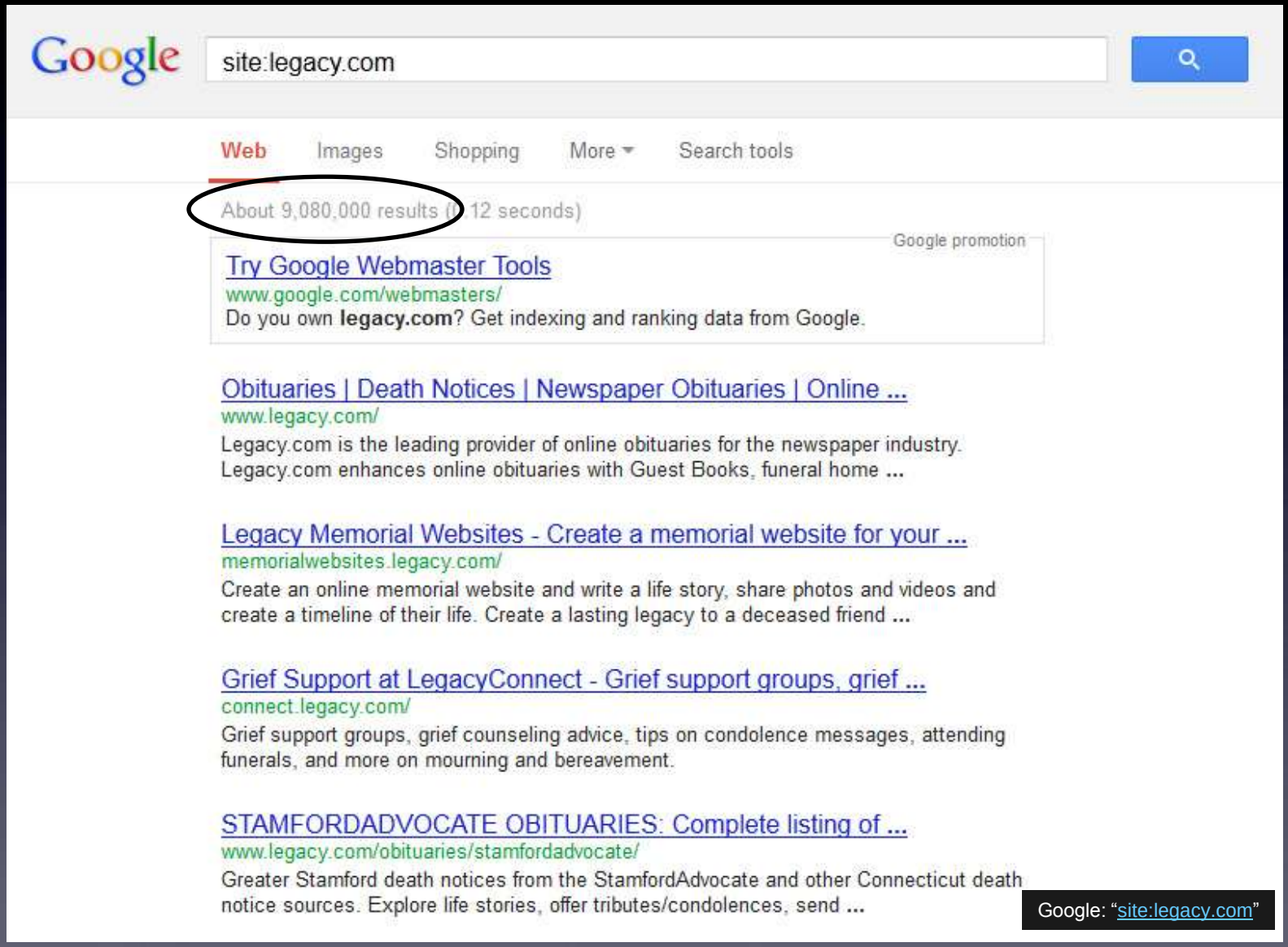
last.fm: robots.txt

TECHNICAL CHALLENGES

Legacy.com

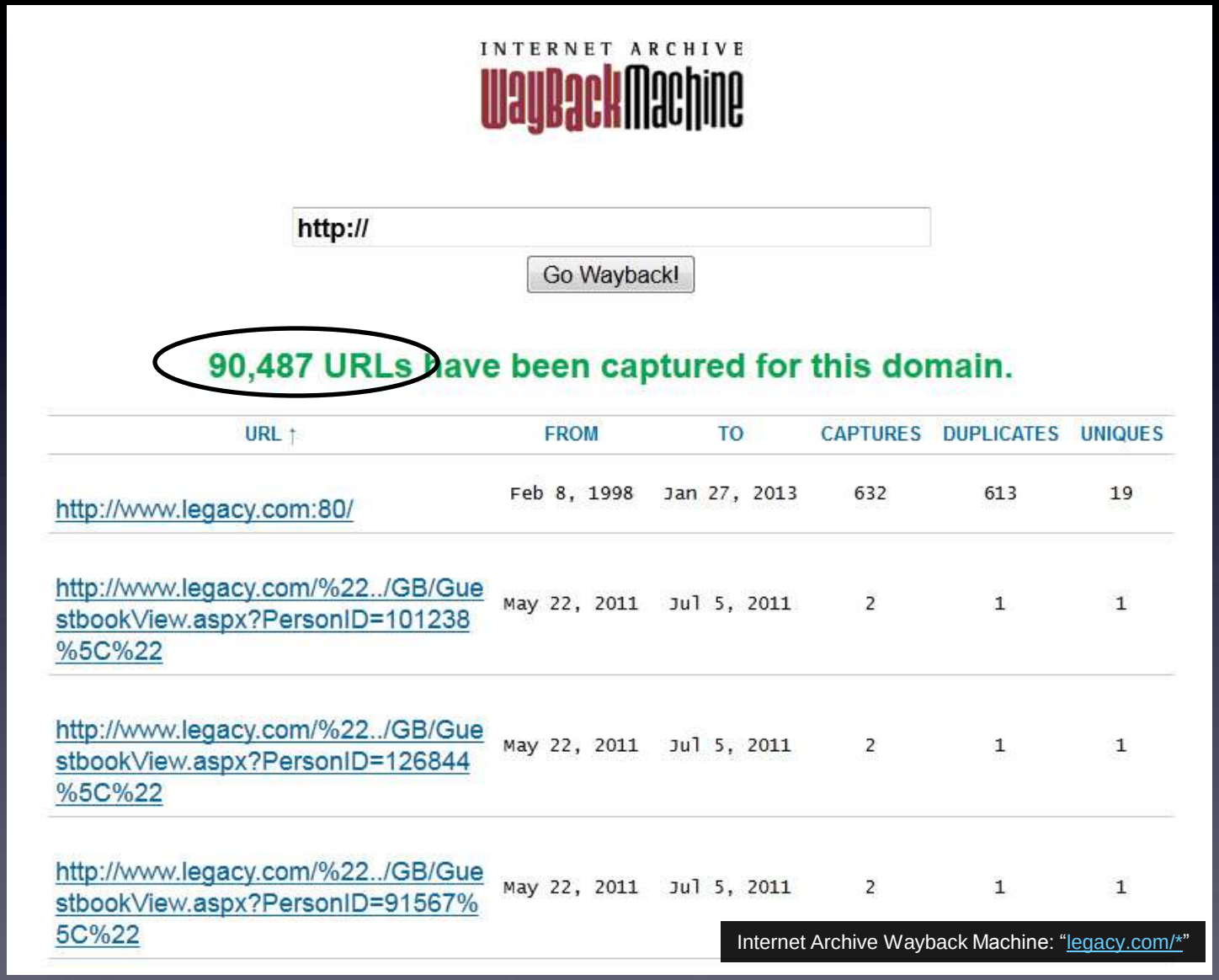
http://www.legacy.com/

hmm, let's temper ambitions



Google: "site:legacy.com"

temper them some more



URL ↑	FROM	TO	CAPTURES	DUPLICATES	UNIQUES
http://www.legacy.com:80/	Feb 8, 1998	Jan 27, 2013	632	613	19
http://www.legacy.com/%22../GB/GuestbookView.aspx?PersonID=101238%5C%22	May 22, 2011	Jul 5, 2011	2	1	1
http://www.legacy.com/%22../GB/GuestbookView.aspx?PersonID=126844%5C%22	May 22, 2011	Jul 5, 2011	2	1	1
http://www.legacy.com/%22../GB/GuestbookView.aspx?PersonID=91567%5C%22	May 22, 2011	Jul 5, 2011	2	1	1

Internet Archive Wayback Machine: “legacy.com/*”

follow robots.txt?

pluses

- courteous behavior
- may ward off the crawler from crawler traps
- less likely to invoke webmaster ire and outright crawler blocking

minuses

- may block the crawler outright
- may miss content vital to purpose of archiving
- may increase time to capture relevant content

many robots directives…

Sitemap: http://www.legacy.com/sitemap.xml User-agent: Googlebot/ Disallow: /Images Disallow: /Multimedia Disallow: /Servic
/*/commemorative-guestbook.aspx Disallow: /guestbook/commemorative-guestbook.aspx Disallow: /guestbook/*/commemorativ
Disallow: /guestbook/*/sponsor-guestbook.aspx Disallow: /guestbooks/thank-you.aspx Disallow: /guestbooks/*/thank-you.aspx D
guestbook-thank-you.aspx Disallow: /guestbook/sign-guestbook-thank-you.aspx Disallow: /guestbook/*/sign-guestbook-thank-yo
thank-you.aspx Disallow: /guestbook/*/photo-guestbook-thank-you.aspx Disallow: /guestbooks/sponsor-guestbook-thank-you.asp
thank-you.aspx Disallow: /obituaries/mymemorialsfacebookfollowwindow.aspx Disallow: /obituaries/*/mymemorialsfacebookfol
Disallow: /ObitNetworkDemo/ Disallow: /guestbooks/commemorative-guestbook.aspx Disallow: /guestbooks/*/commemorative-
guestbook.aspx Disallow: /guestbooks/*/sponsor-guestbook.aspx Disallow: /guestbook/sponsor-guestbook.aspx Disallow: /guestb
/guestbook/*/thank-you.aspx Disallow: /guestbooks/sign-guestbook-thank-you.aspx Disallow: /guestbooks/*/sign-guestbook-than
thank-you.aspx Disallow: /guestbooks/*/photo-guestbook-thank-you.aspx Disallow: /guestbook/photo-guestbook-thank-you.aspx
thank-you.aspx Disallow: /guestbook/sponsor-guestbook-thank-you.aspx Disallow: /guestbook/*/sponsor-guestbook-thank-you.as
Disallow: /Images Disallow: /Multimedia Disallow: /Services/admina.asp Disallow: /Services/adminc.asp Disallow: /ObitNetwor
guestbook.aspx Disallow: /guestbook/*/commemorative-guestbook.aspx Disallow: /guestbooks/sponsor-guestbook.aspx Disallow
/thank-you.aspx Disallow: /guestbooks/*/thank-you.aspx Disallow: /guestbook/thank-you.aspx Disallow: /guestbook/*/thank-you
you.aspx Disallow: /guestbook/*/sign-guestbook-thank-you.aspx Disallow: /guestbooks/photo-guestbook-thank-you.aspx Disallo
Disallow: /guestbooks/sponsor-guestbook-thank-you.aspx Disallow: /guestbooks/*/sponsor-guestbook-thank-you.aspx Disallow:
/mymemorialsfacebookfollowwindow.aspx Disallow: /obituaries/*/mymemorialsfacebookfollowwindow.aspx User-agent: Scoote
/commemorative-guestbook.aspx Disallow: /guestbooks/*/commemorative-guestbook.aspx Disallow: /guestbook/commemorative
guestbook.aspx Disallow: /guestbook/sponsor-guestbook.aspx Disallow: /guestbook/*/sponsor-guestbook.aspx Disallow: /guestbo
/sign-guestbook-thank-you.aspx Disallow: /guestbooks/*/sign-guestbook-thank-you.aspx Disallow: /guestbook/sign-guestbook-th
guestbook-thank-you.aspx Disallow: /guestbook/photo-guestbook-thank-you.aspx Disallow: /guestbook/*/photo-guestbook-thank
guestbook-thank-you.aspx Disallow: /guestbook/*/sponsor-guestbook-thank-you.aspx Disallow: /obituaries/mymemorialsfaceboo
/Services/admina.asp Disallow: /Services/adminc.asp Disallow: /ObitNetworkDemo/ Disallow: /guestbooks/commemorative-gues
/*/commemorative-guestbook.aspx Disallow: /guestbooks/sponsor-guestbook.aspx Disallow: /guestbooks/*/sponsor-guestbook.as
/*/thank-you.aspx Disallow: /guestbook/thank-you.aspx Disallow: /guestbook/*/thank-you.aspx Disallow: /guestbooks/sign-guest
guestbook-thank-you.aspx Disallow: /guestbooks/photo-guestbook-thank-you.aspx Disallow: /guestbooks/*/photo-guestbook-thar
guestbook-thank-you.aspx Disallow: /guestbooks/*/sponsor-guestbook-thank-you.aspx Disallow: /guestbook/sponsor-guestbook-t
/*/mymemorialsfacebookfollowwindow.aspx User-agent: Ask Jeeves Disallow: /Images Disallow: /Multimedia Disallow: /Servic
/*/commemorative-guestbook.aspx Disallow: /guestbook/commemorative-guestbook.aspx Disallow: /guestbook/*/commemorativ

Legacy.com: robots.txt

reformat w/ Notepad++ and regex

- find: “ Disallow:”
- replace: “\nDisallow:”

- find: “:\r\n/”
- replace: “: /”

- manually reformat remaining issues

now we have something legible

User-agent: *
Disallow: /Obituaries/AffiliateAdvertisement.axd
Disallow: /obituaries/rss.ashx
Disallow: /obituaries/*/rss.ashx
Disallow: /obituaries/*/*/rss.ashx
Disallow: */obituaries.aspx?*archive=1
Disallow: /guestbooks/commemorative-guestbook.aspx
Disallow: /guestbooks/*/commemorative-guestbook.aspx
Disallow: /guestbook/commemorative-guestbook.aspx
Disallow: /guestbook/*/commemorative-guestbook.aspx
Disallow: /guestbooks/sponsor-guestbook.aspx
Disallow: /guestbooks/*/sponsor-guestbook.aspx
Disallow: /guestbook/sponsor-guestbook.aspx
Disallow: /guestbook/*/sponsor-guestbook.aspx
Disallow: /guestbooks/thank-you.aspx
Disallow: /guestbooks/*/thank-you.aspx
Disallow: /guestbook/thank-you.aspx
Disallow: /guestbook/*/thank-you.aspx
Disallow: /guestbooks/sign-guestbook-thank-you.aspx
Disallow: /guestbooks/*/sign-guestbook-thank-you.aspx
Disallow: /guestbook/sign-guestbook-thank-you.aspx
Disallow: /guestbook/*/sign-guestbook-thank-you.aspx
Disallow: /guestbooks/photo-guestbook-thank-you.aspx
Disallow: /guestbooks/*/photo-guestbook-thank-you.aspx
Disallow: /guestbook/photo-guestbook-thank-you.aspx
Disallow: /guestbook/*/photo-guestbook-thank-you.aspx
Disallow: /guestbooks/sponsor-guestbook-thank-you.aspx
Disallow: /guestbooks/*/sponsor-guestbook-thank-you.aspx
Disallow: /guestbook/sponsor-guestbook-thank-you.aspx
Disallow: /guestbook/*/sponsor-guestbook-thank-you.aspx
Disallow: /obituaries/mymemorialsfacebookfollowwindow.aspx
Disallow: /obituaries/*/mymemorialsfacebookfollowwindow.aspx

will robots.txt prevent capturing obituaries?

- compare sample urls from multiple news sites to robots.txt directives
  - http://www.legacy.com/obituaries/alamogordonews/obituary.aspx?n=lanita-klingenberg&pid=162926909
  - http://www.legacy.com/obituaries/heraldobserver/obituary.aspx?n=james-f-davis&pid=162902531
  - http://www.legacy.com/obituaries/spartanburg/obituary.aspx?n=louise-hardin&pid=162947461
  - http://www.legacy.com/obituaries/newsminer/obituary.aspx?n=herman-h-demit&pid=162793852
  - http://www.legacy.com/obituaries/dailygazette/obituary.aspx?n=richard-e-martel&pid=162941147
- pattern: http://www.legacy.com/obituaries/*/obituary.aspx

robots.txt won’t prevent capturing obituaries

User-agent: *
Disallow: /Obituaries/AffiliateAdvertisement.axd
Disallow: /obituaries/rss.ashx
Disallow: /obituaries/*/rss.ashx
Disallow: /obituaries/*/*/rss.ashx
Disallow: */obituaries.aspx?*archive=1
Disallow: /guestbooks/commemorative-guestbook.aspx
Disallow: /guestbooks/*/commemorative-guestbook.aspx
Disallow: /guestbook/commemorative-guestbook.aspx
Disallow: /guestbook/*/commemorative-guestbook.aspx
Disallow: /guestbooks/sponsor-guestbook.aspx
Disallow: /guestbooks/*/sponsor-guestbook.aspx
Disallow: /guestbook/sponsor-guestbook.aspx
Disallow: /guestbook/*/sponsor-guestbook.aspx
Disallow: /guestbooks/thank-you.aspx
Disallow: /guestbooks/*/thank-you.aspx
Disallow: /guestbook/thank-you.aspx
Disallow: /guestbook/*/thank-you.aspx
Disallow: /guestbooks/sign-guestbook-thank-you.aspx
Disallow: /guestbooks/*/sign-guestbook-thank-you.aspx
Disallow: /guestbook/sign-guestbook-thank-you.aspx
Disallow: /guestbook/*/sign-guestbook-thank-you.aspx
Disallow: /guestbooks/photo-guestbook-thank-you.aspx
Disallow: /guestbooks/*/photo-guestbook-thank-you.aspx
Disallow: /guestbook/photo-guestbook-thank-you.aspx
Disallow: /guestbook/*/photo-guestbook-thank-you.aspx
Disallow: /guestbooks/sponsor-guestbook-thank-you.aspx
Disallow: /guestbooks/*/sponsor-guestbook-thank-you.aspx
Disallow: /guestbook/sponsor-guestbook-thank-you.aspx
Disallow: /guestbook/*/sponsor-guestbook-thank-you.aspx
Disallow: /obituaries/mymemorialsfacebookfollowwindow.aspx
Disallow: /obituaries/*/mymemorialsfacebookfollowwindow.aspx

will robots.txt prevent capturing guestbooks?

- compare sample urls from multiple news sites to robots.txt directives
  - http://www.legacy.com/guestbooks/alamogordonews/guestbook.aspx?n=lanita-klingenberg&pid=162926909&cid=full
  - http://www.legacy.com/guestbooks/heraldobserver/guestbook.aspx?n=james-davis&pid=162902531&cid=full
  - http://www.legacy.com/guestbooks/spartanburg/guestbook.aspx?n=louise-hardin&pid=162947461&cid=full
  - http://www.legacy.com/guestbooks/newsminer/guestbook.aspx?n=herman-demit&pid=162793852&cid=full
  - http://www.legacy.com/guestbooks/cypresscreek/guestbook.aspx?n=charles-wilson&pid=162376967&cid=full
- pattern: http://www.legacy.com/guestbooks/*/guestbook.aspx

robots.txt won’t prevent capturing guestbooks

User-agent: *
Disallow: /Obituaries/AffiliateAdvertisement.axd
Disallow: /obituaries/rss.ashx
Disallow: /obituaries/*/rss.ashx
Disallow: /obituaries/*/*/rss.ashx
Disallow: */obituaries.aspx?*archive=1
Disallow: /guestbooks/commemorative-guestbook.aspx
Disallow: /guestbooks/*/commemorative-guestbook.aspx
Disallow: /guestbook/commemorative-guestbook.aspx
Disallow: /guestbook/*/commemorative-guestbook.aspx
Disallow: /guestbooks/sponsor-guestbook.aspx
Disallow: /guestbooks/*/sponsor-guestbook.aspx
Disallow: /guestbook/sponsor-guestbook.aspx
Disallow: /guestbook/*/sponsor-guestbook.aspx
Disallow: /guestbooks/thank-you.aspx
Disallow: /guestbooks/*/thank-you.aspx
Disallow: /guestbook/thank-you.aspx
Disallow: /guestbook/*/thank-you.aspx
Disallow: /guestbooks/sign-guestbook-thank-you.aspx
Disallow: /guestbooks/*/sign-guestbook-thank-you.aspx
Disallow: /guestbook/sign-guestbook-thank-you.aspx
Disallow: /guestbook/*/sign-guestbook-thank-you.aspx
Disallow: /guestbooks/photo-guestbook-thank-you.aspx
Disallow: /guestbooks/*/photo-guestbook-thank-you.aspx
Disallow: /guestbook/photo-guestbook-thank-you.aspx
Disallow: /guestbook/*/photo-guestbook-thank-you.aspx
Disallow: /guestbooks/sponsor-guestbook-thank-you.aspx
Disallow: /guestbooks/*/sponsor-guestbook-thank-you.aspx
Disallow: /guestbook/sponsor-guestbook-thank-you.aspx
Disallow: /guestbook/*/sponsor-guestbook-thank-you.aspx
Disallow: /obituaries/mymemorialsfacebookfollowwindow.aspx
Disallow: /obituaries/*/mymemorialsfacebookfollowwindow.aspx

will robots.txt prevent capturing thumbnail photos?

- compare sample urls from multiple news sites to robots.txt directives
  - https://cache.legacy.com/legacy/images/cobrands/heraldobserver/Photos/79932432-502b-4c51-b9a1-d3f80f5f273f.jpg
  - https://cache.legacy.com/legacy/images/cobrands/newsminer/Photos/86f6d696-d4d7-419c-9313-447db5ec0268.jpg
  - https://cache.legacy.com/legacy/images/cobrands/cypresscreek/Photos/G286053_1_20130116.jpg
  - https://cache.legacy.com/legacy/images/Cobrands/BaxterBulletin/Photos/BBL012735-1_20130208.jpg
  - https://cache.legacy.com/legacy/images/Portraits/James-DePreist-dead-162938549port.jpgx?w=117&h=151&option=1
- pattern: https://cache.legacy.com/legacy/images/

robots.txt won’t prevent capturing thumbnail photos

User-agent: *
Disallow: /Obituaries/AffiliateAdvertisement.axd
Disallow: /obituaries/rss.ashx
Disallow: /obituaries/*/rss.ashx
Disallow: /obituaries/*/*/rss.ashx
Disallow: */obituaries.aspx?*archive=1
Disallow: /guestbooks/commemorative-guestbook.aspx
Disallow: /guestbooks/*/commemorative-guestbook.aspx
Disallow: /guestbook/commemorative-guestbook.aspx
Disallow: /guestbook/*/commemorative-guestbook.aspx
Disallow: /guestbooks/sponsor-guestbook.aspx
Disallow: /guestbooks/*/sponsor-guestbook.aspx
Disallow: /guestbook/sponsor-guestbook.aspx
Disallow: /guestbook/*/sponsor-guestbook.aspx
Disallow: /guestbooks/thank-you.aspx
Disallow: /guestbooks/*/thank-you.aspx
Disallow: /guestbook/thank-you.aspx
Disallow: /guestbook/*/thank-you.aspx
Disallow: /guestbooks/sign-guestbook-thank-you.aspx
Disallow: /guestbooks/*/sign-guestbook-thank-you.aspx
Disallow: /guestbook/sign-guestbook-thank-you.aspx
Disallow: /guestbook/*/sign-guestbook-thank-you.aspx
Disallow: /guestbooks/photo-guestbook-thank-you.aspx
Disallow: /guestbooks/*/photo-guestbook-thank-you.aspx
Disallow: /guestbook/photo-guestbook-thank-you.aspx
Disallow: /guestbook/*/photo-guestbook-thank-you.aspx
Disallow: /guestbooks/sponsor-guestbook-thank-you.aspx
Disallow: /guestbooks/*/sponsor-guestbook-thank-you.aspx
Disallow: /guestbook/sponsor-guestbook-thank-you.aspx
Disallow: /guestbook/*/sponsor-guestbook-thank-you.aspx
Disallow: /obituaries/mymemorialsfacebookfollowwindow.aspx
Disallow: /obituaries/*/mymemorialsfacebookfollowwindow.aspx

what to set as seed url(s)?

- seed url is where the web crawler starts
- my goal: pick a sub-section whose html content I could crawl exhaustively
- New Mexico Newspapers

basic crawler operation

1. start at seed url
2. extract all links
3. put links in a queue
4. compare link against scope
   a. if out of scope, don't follow
   b. if within scope, capture and return to 2.
5. repeat until crawl exhausted or terminated by operator

HTTrack overview

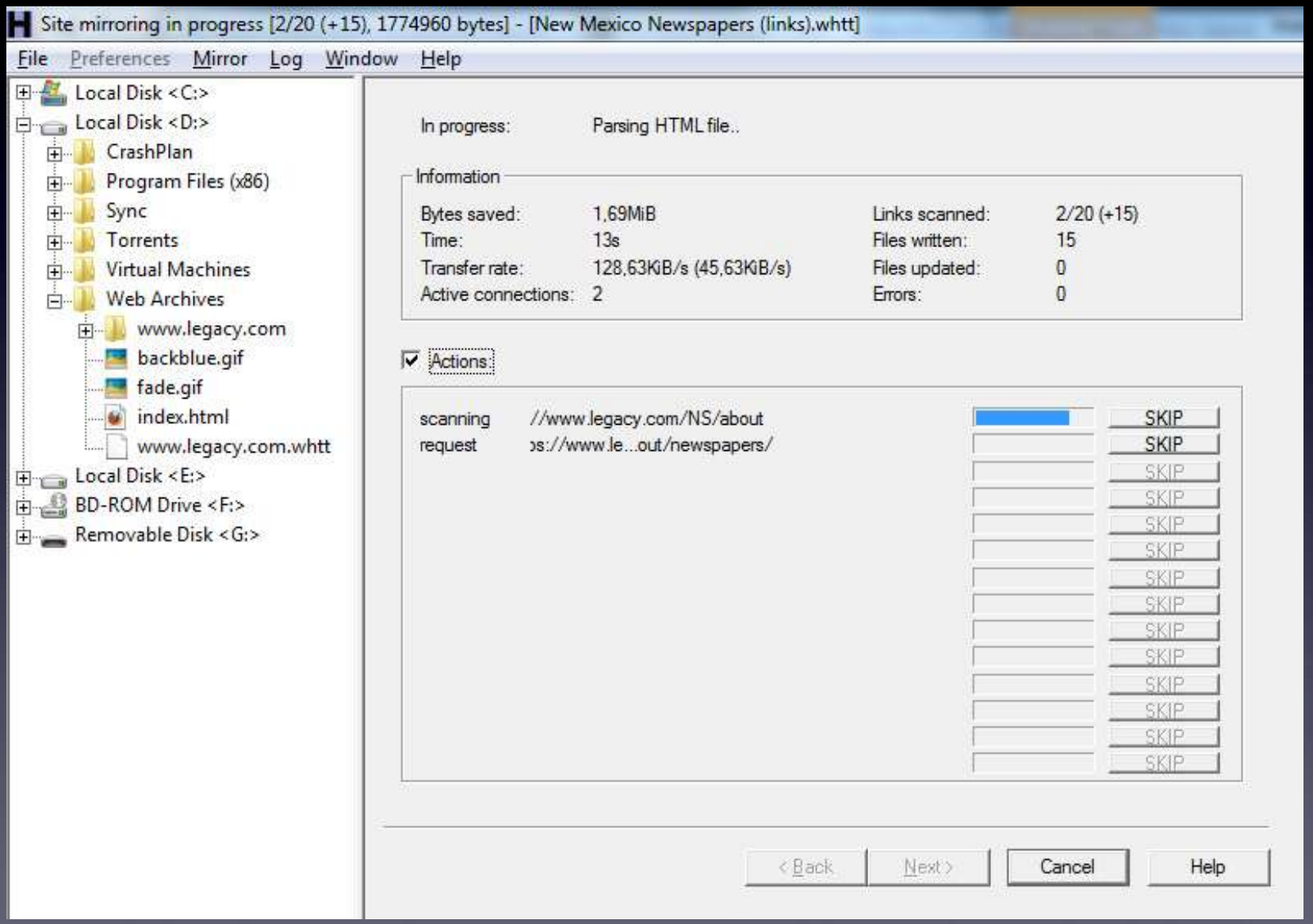
http://www.httrack.com/

- small-scale website copier
- recreates remote website as local filesystem hierarchy
- Windows GUI and CLI
- OSX/Linux web service and CLI

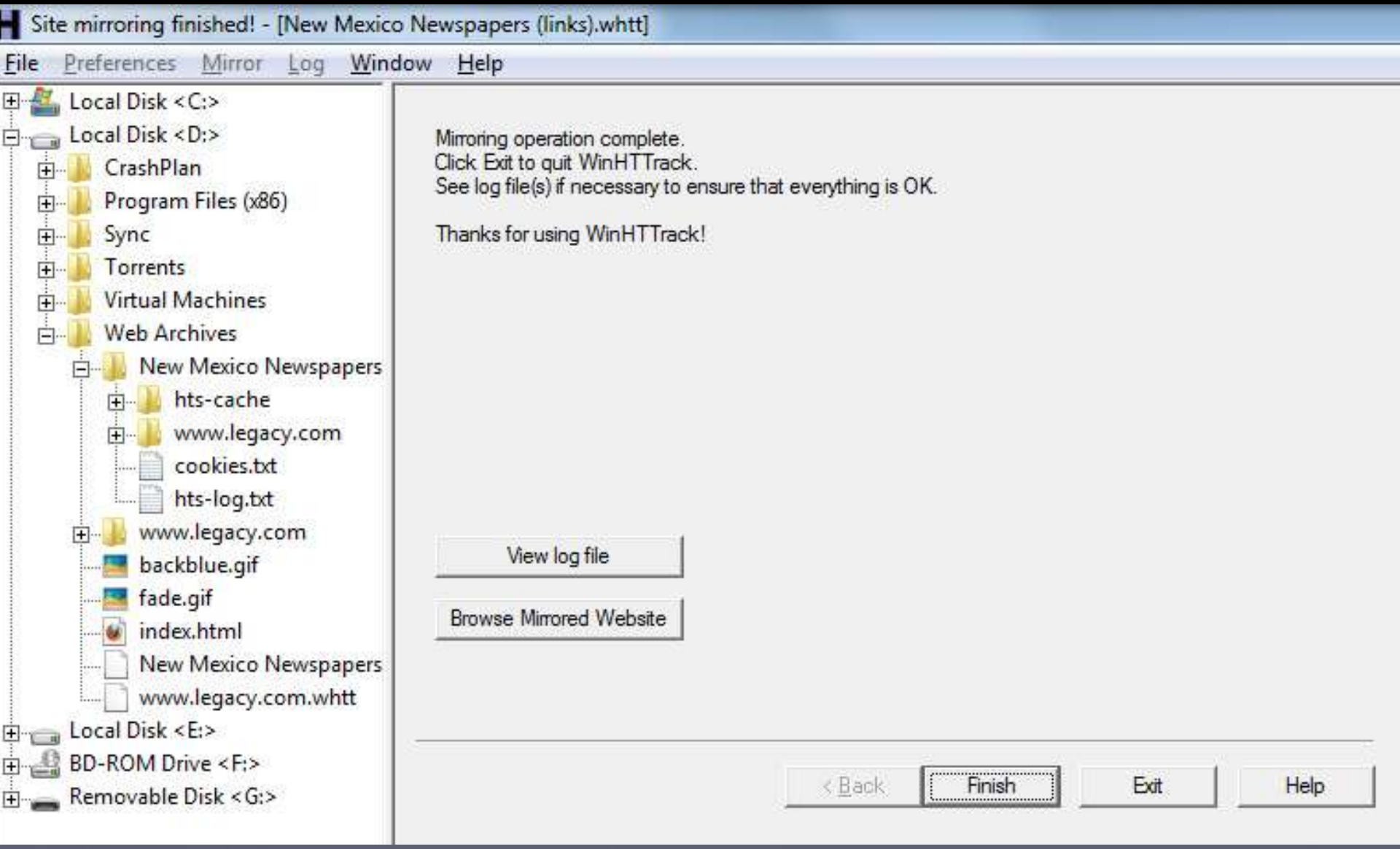
start by scanning links

- run HTTrack
- New project name: `New Mexico Newspapers (links)`
- Mirroring Mode: `Download web site(s)`
- Web Addresses (URL): https://www.legacy.com/NS/about/newspapers/?sid=39
- Set options:
  - Scan rules: `-mime:*/* +mime:text/html`
  - Limits > Maximum mirroring depth: `2`
  - Experts Only > Primary Scan Rule (scan mode): `Just scan`
  - Experts Only > Travel mode: `Can both go up & down`
  - Experts Only > Global travel mode: `Stay on the same domain`
- click `Next`, then `Finish` to run configured crawl

link scanning in progress



link scanning completed



examine crawl log in Notepad++

- `~\New Mexico Newspapers (links)\hts-log.txt`
- links off from seed url:
  - www.legacy.com/ns/
  - www.legacy.com/memorial-sites/sandy-hook-school-tragedy/
  - https://www.legacy.com/NS/
  - www.legacy.com/obituaries/alamogordonews/
  - www.legacy.com/obituaries/demingheadlight/
  - www.legacy.com/obituaries/daily-times/
  - www.legacy.com/obituaries/currentargus/
  - www.legacy.com/obituaries/lascrucesbulletin/
  - www.legacy.com/obituaries/lcsun-news/
  - www.legacy.com/obituaries/lasvegasoptic/
  - www.legacy.com/obituaries/lamonitor/
  - www.legacy.com/obituaries/ruidosonews/
  - www.legacy.com/obituaries/santafenewmexican/
- these inform scope

a list of urls may be all you need

- Voyant Tools: online text analysis platform
- examine word frequency and distribution
- accepts urls or uploaded files
  - urls are easier
  - can only upload 1 file at a time

scoping

- defines what crawler should/should not crawl
- base on extracted seed urls and earlier robots.txt analysis
- exclude:
  - www.legacy.com/ns/ (and everything “below” in path)
  - www.legacy.com/memorial-sites/ (and everything “below” in path)
  - https://www.legacy.com/NS/ (and everything “below” in path)
  - cache.legacy.com/ (and everything “below” in path)
  - non-html files
- include:
  - www.legacy.com/obituaries/ (and everything “below” in path)
  - www.legacy.com/guestbooks/ (and everything “below” in path)
  - html files

configure the crawl

- New project name: `New Mexico Newspapers (html)`
- Mirroring Mode: `Download web site(s)`
- Web Addresses (URL): https://www.legacy.com/NS/about/newspapers/?sid=39
- Set options:
  - Scan rules: (see next slide)
  - Links: `Get HTML files first!`
  - Experts Only > Primary Scan Rule (scan mode): `Store html files`
  - Experts Only > Travel mode: `Can both go up & down`
  - Experts Only > Global travel mode: `Stay on the same domain`

scan rules (scoping)

- follow HTTrack scan rule syntax:

  -mime:*/*

  +mime:text/html

  -www.legacy.com/ns/*

  -www.legacy.com/memorial-sites/*

  -https://www.legacy.com/NS/*

  -cache.legacy.com/

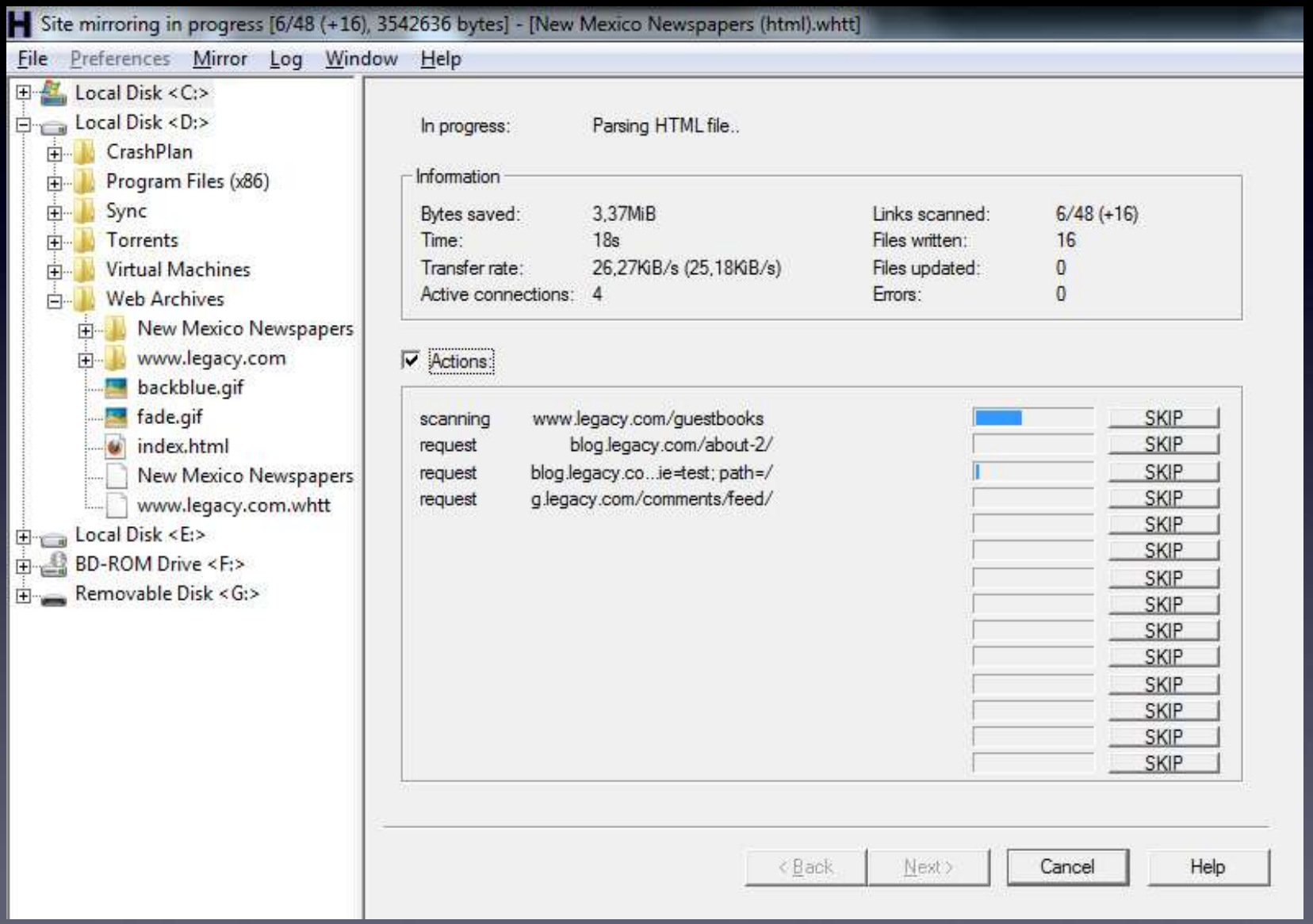
  +www.legacy.com/obituaries/*

  +www.legacy.com/guestbooks/*

optional configuration parameters

- Limits > Site size limit (B)
  - prevents overcrawling w/ misconfigured crawl
- Flow Control > Number of connections
  - parallelizes link retrieval, hastening crawl
- Spider > Spider
  - toggle adherence to robots.txt directives
- Log, Index, Cache > Make a word database
  - creates a word count index at crawl completion

mirroring in progress



watch the crawl log

- `~\New Mexico Newspapers (html)\hts-log.txt`
- open in Notepad++
- scroll to bottom
- go to `File` menu
- select `Reload from Disk` to see latest downloads
- confirm it's crawling what you expect
- if not, cancel, reconfigure, run again w/ resume mode
- I decided to re-run crawl with additional exclude filters:

  -blog.legacy.com/*

  -*connect.legacy.com/*

  -media2.legacy.com/*

  -memorialwebsites.legacy.com/*

crawl results

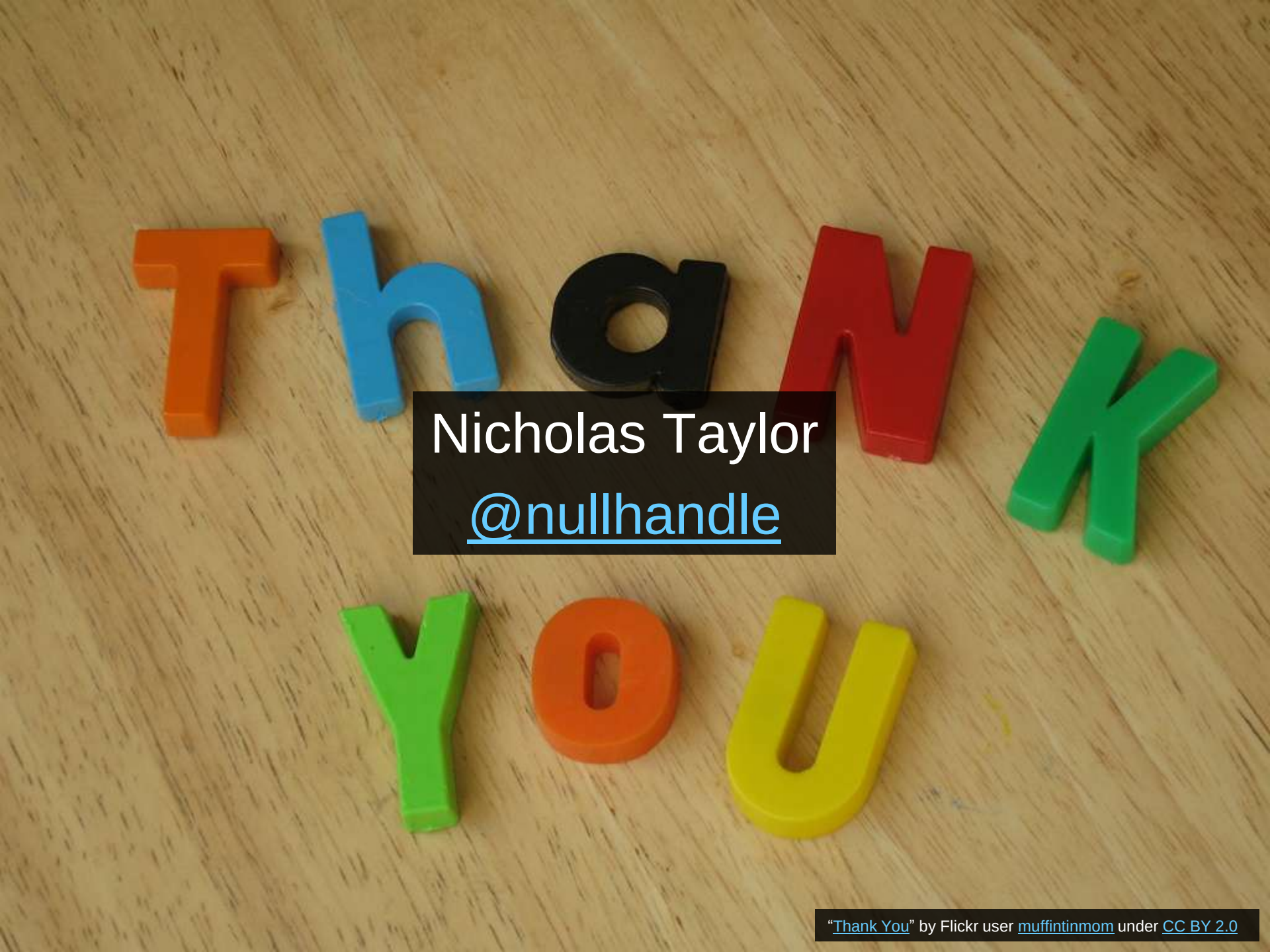
- crawl terminated by operator after 2.5 hours
- 19281 links scanned
- 12264 files written or updated
- 826 MB
- rough performance benchmark for trying to capture only obituaries and guest books for 13 newspapers (w/ 4 threads)

data for analysis

- concatenate html files, then upload to Voyant Tools
- alternatively, these 2 files are “spreadsheet-able” data including file size, http response code, mime type, date of capture, url, and url of document where discovered:

```
~\New Mexico Newspapers (html)\hts-cache\new.txt
~\New Mexico Newspapers (html)\hts-cache\old.txt
```

fwiw, we may not need to archive, after all



Nicholas Taylor

@nullhandle