



# Web Archiving Legacy.com: A Case Study

Nicholas Taylor  
[@nullhandle](mailto:nicolas@nullhandle.com)

[History and New Media](#)  
February 11, 2013

Google Images: "[site:legacy.com](http://site.legacy.com)"

# challenges

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



[“looking up”](#) by Flickr user [lovestruck](#), under [CC BY-NC-SA 2.0](#)



# LEGAL CHALLENGES

U.S. Copyright Office: "[Copyright Act \(Title 17, Chapter 1\)](#)"

# copyright law

Is copyright a little fuzzy?



# 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.”



# 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: /
```





# TECHNICAL CHALLENGES

"When I was a kid, I dreamed of you." by Flickr user [moonbird](#) under [CC BY-NC-ND 2.0](#)



# Legacy.com

Comfortable rooms, friendly service, HDTVs and more

BOOK NOW

LAQUINTA  
Hotels & Suites

[About](#) • [Company Blog](#) • [Contact Us](#) • [Subscribe With Us](#) • [Saturday, February 08, 2013](#)

## Legacy.com®

Where life stories live on™

Visit us on [f](#) [t](#) [in](#) [v](#)

**Obituaries & Guest Books**  
Celebrate a loved one's life

**Memorial Sites**  
Remember shared legacies

**LegacyConnect**  
Find grief support & advice

**ObitMessenger**  
Get free obituary alerts

### SEARCH OBITUARIES

**By newspaper**  
Select a state to browse up-to-date obituaries from hundreds of local newspapers in the U.S. or find obituaries from our international newspaper partners

[Learn More >](#)

**By name or keyword**  
Search millions of obituaries from more than 500 newspapers worldwide

First Name  Last Name

Keyword  Past 3 Days

United States  All States

### FEATURED OBITUARIES

**James DePreist**  
One of the first African American conductors of a major orchestra.  
[Obituary](#) | [Guest Book](#)

**Guy Tozzoli**  
Supervised the development of the World Trade Center...  
[Obituary](#) | [Guest Book](#)

**Stuart Freeborn**  
Pioneering makeup artist behind Yoda and Chewbacca in the "Star Wars" films.  
[Obituary](#) | [Guest Book](#)

**MEMORIAL SITES** [View All >](#)

**Civil Rights**  
Honoring those who worked to ensure civil rights for all

**Wars in Iraq and Afghanistan**  
Paying tribute to U.S. veterans of the Iraq-Afghan wars

**2013**  
Remembering famous figures lost in 2013

**LIVING WITH LOSS**

*If You Want To Be Happy...*  
By Nancy Weil

Grief is real and painful, but it doesn't have to define you. If you want to be happy, here are some simple steps you can take to help you get there...

[Learn More >](#)

**SYMPATHY & SUPPORT**

[Helpful Links](#)  
[Planning an Obituary](#)  
[Funeral Home Directory](#)  
[Advice & Support](#)  
[Gift Shop](#)  
[Flowers](#)  
[Candles](#)

**FUNERAL ETIQUETTE**

### LEGENDS & LEGACIES

[View All >](#)

**Alex Haley's Roots**  
20K2013  
Alex Haley's epic tale "Roots" gave

**Different Stripes Aligned The World**  
20K2013

**Anne-Marie Lindberg's Gift**  
2-7-2013

**Virginia**  
DRIVER'S LICENSE

[Compensation.org](#)

# hmm, let's temper ambitions

The image is a screenshot of a Google search interface. At the top, the Google logo is on the left, and a search bar contains the text "site:legacy.com". To the right of the search bar is a blue button with a magnifying glass icon. Below the search bar, there are tabs for "Web", "Images", "Shopping", "More", and "Search tools". The "Web" tab is selected and highlighted with a red underline. Below the tabs, the search results are displayed. The first result is a "Google promotion" box that says "Try Google Webmaster Tools" with a link to "www.google.com/webmasters/" and the text "Do you own legacy.com? Get indexing and ranking data from Google." Below this, there are four organic search results, each with a blue link, a green URL, and a brief description. The first organic result is "Obituaries | Death Notices | Newspaper Obituaries | Online ..." with URL "www.legacy.com/" and a description about Legacy.com being a leading provider of online obituaries. The second is "Legacy Memorial Websites - Create a memorial website for your ..." with URL "memorialwebsites.legacy.com/" and a description about creating an online memorial website. The third is "Grief Support at LegacyConnect - Grief support groups, grief ..." with URL "connect.legacy.com/" and a description about grief support groups and counseling. The fourth is "STAMFORDADVOCATE OBITUARIES: Complete listing of ..." with URL "www.legacy.com/obituaries/stamfordadvocate/" and a description about Greater Stamford death notices. In the bottom right corner, there is a black box with white text that says "Google: 'site:legacy.com'".

Google

site:legacy.com

Web Images Shopping More Search tools

About 9,080,000 results (0.12 seconds)

Google promotion

[Try Google Webmaster Tools](#)  
[www.google.com/webmasters/](http://www.google.com/webmasters/)  
Do you own **legacy.com**? Get indexing and ranking data from Google.

[Obituaries | Death Notices | Newspaper Obituaries | Online ...](#)  
[www.legacy.com/](http://www.legacy.com/)  
Legacy.com is the leading provider of online obituaries for the newspaper industry. Legacy.com enhances online obituaries with Guest Books, funeral home ...

[Legacy Memorial Websites - Create a memorial website for your ...](#)  
[memorialwebsites.legacy.com/](http://memorialwebsites.legacy.com/)  
Create an online memorial website and write a life story, share photos and videos and create a timeline of their life. Create a lasting legacy to a deceased friend ...

[Grief Support at LegacyConnect - Grief support groups, grief ...](#)  
[connect.legacy.com/](http://connect.legacy.com/)  
Grief support groups, grief counseling advice, tips on condolence messages, attending funerals, and more on mourning and bereavement.

[STAMFORDADVOCATE OBITUARIES: Complete listing of ...](#)  
[www.legacy.com/obituaries/stamfordadvocate/](http://www.legacy.com/obituaries/stamfordadvocate/)  
Greater Stamford death notices from the StamfordAdvocate and other Connecticut death notice sources. Explore life stories, offer tributes/condolences, send ...

Google: "site:legacy.com"

# temper them some more

Go Wayback!

**90,487 URLs** have been captured for this domain.

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

Internet Archive Wayback Machine: "[legacy.com/](http://www.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: /Service  
/\*/commemorative-guestbook.aspx Disallow: /guestbook/commemorative-guestbook.aspx Disallow: /guestbook/\*/commemorative-  
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-ye  
thank-you.aspx Disallow: /guestbook/\*/photo-guestbook-thank-you.aspx Disallow: /guestbooks/sponsor-guestbook-thank-you.as  
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-tha  
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/admind.asp Disallow: /ObitNetwork  
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 Disallow  
Disallow: /guestbooks/sponsor-guestbook-thank-you.aspx Disallow: /guestbooks/\*/sponsor-guestbook-thank-you.aspx Disallow: /  
/mymemorialsfacebookfollowwindow.aspx Disallow: /obituaries/\*/mymemorialsfacebookfollowwindow.aspx User-agent: Scooter  
/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/admind.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-tha  
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: /Service  
/\*/commemorative-guestbook.aspx Disallow: /guestbook/commemorative-guestbook.aspx Disallow: /guestbook/\*/commemorative-

# 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-klingsberg&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](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=ames-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](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/cobrand/heraldobserver/Photos/79932432-502b-4c51-b9a1-d3f80f5f273f.jpg>
  - <https://cache.legacy.com/legacy/images/cobrand/newsminer/Photos/86f6d696-d4d7-419c-9313-447db5ec0268.jpg>
  - [https://cache.legacy.com/legacy/images/cobrand/cypresscreek/Photos/G286053\\_1\\_20130116.jpg](https://cache.legacy.com/legacy/images/cobrand/cypresscreek/Photos/G286053_1_20130116.jpg)
  - [https://cache.legacy.com/legacy/images/Cobrand/BaxterBulletin/Photos/BBL012735-1\\_20130208.jpg](https://cache.legacy.com/legacy/images/Cobrand/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

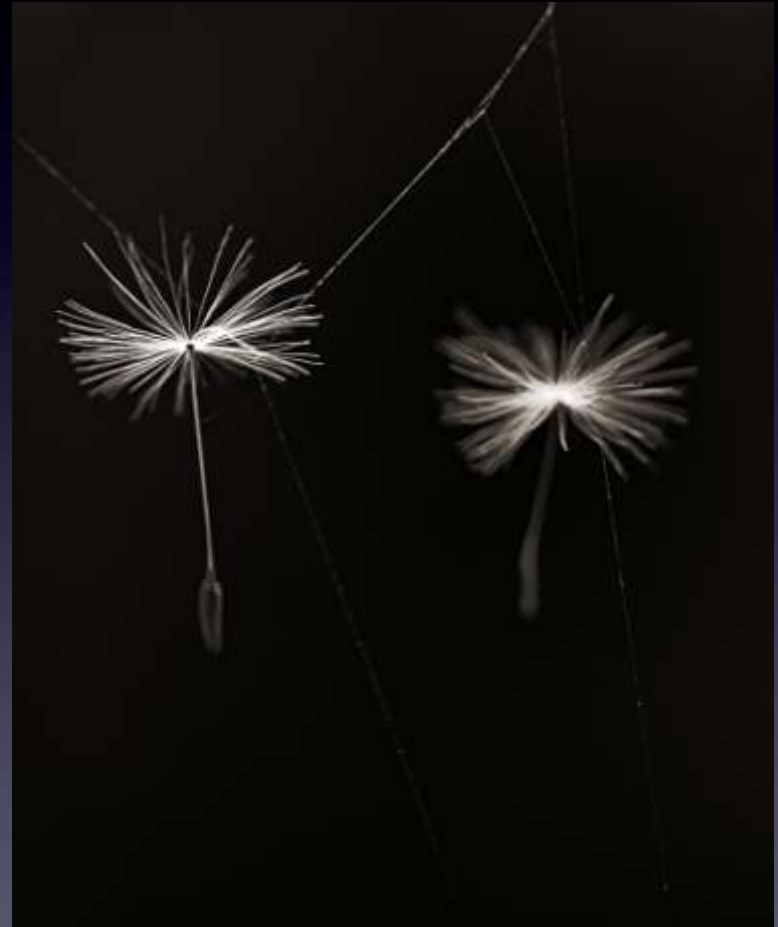
Disallow: /guestbook/\*/sponsor-guestbook-thank-you.aspx

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 subsection whose html content I could crawl exhaustively
- [New Mexico Newspapers](#)



["I was somebody falling for one who was not somebody tired of dreaming."](#)  
by Flickr user [Neal](#) under [CC BY-NC-ND 2.0](#)

# 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

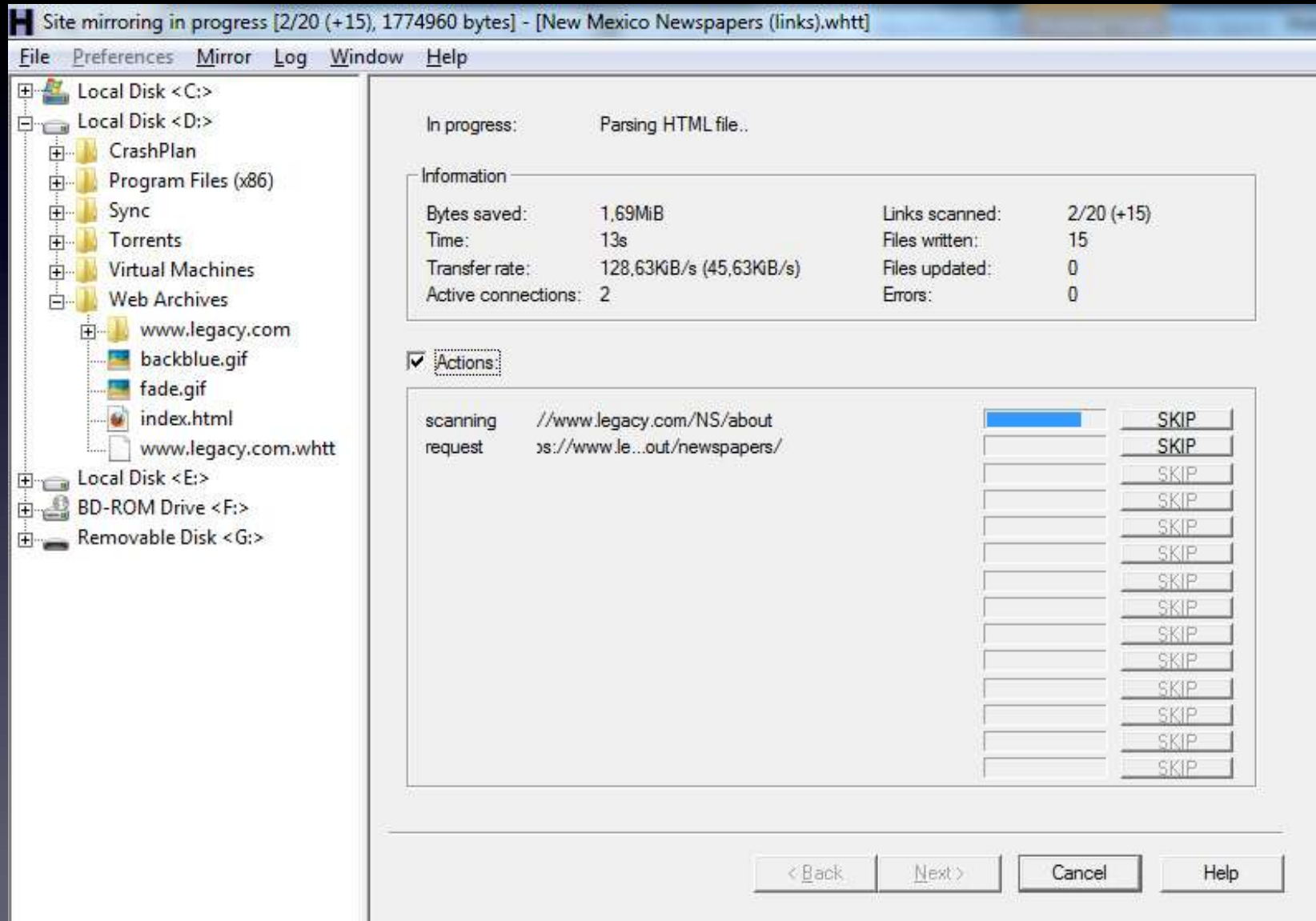




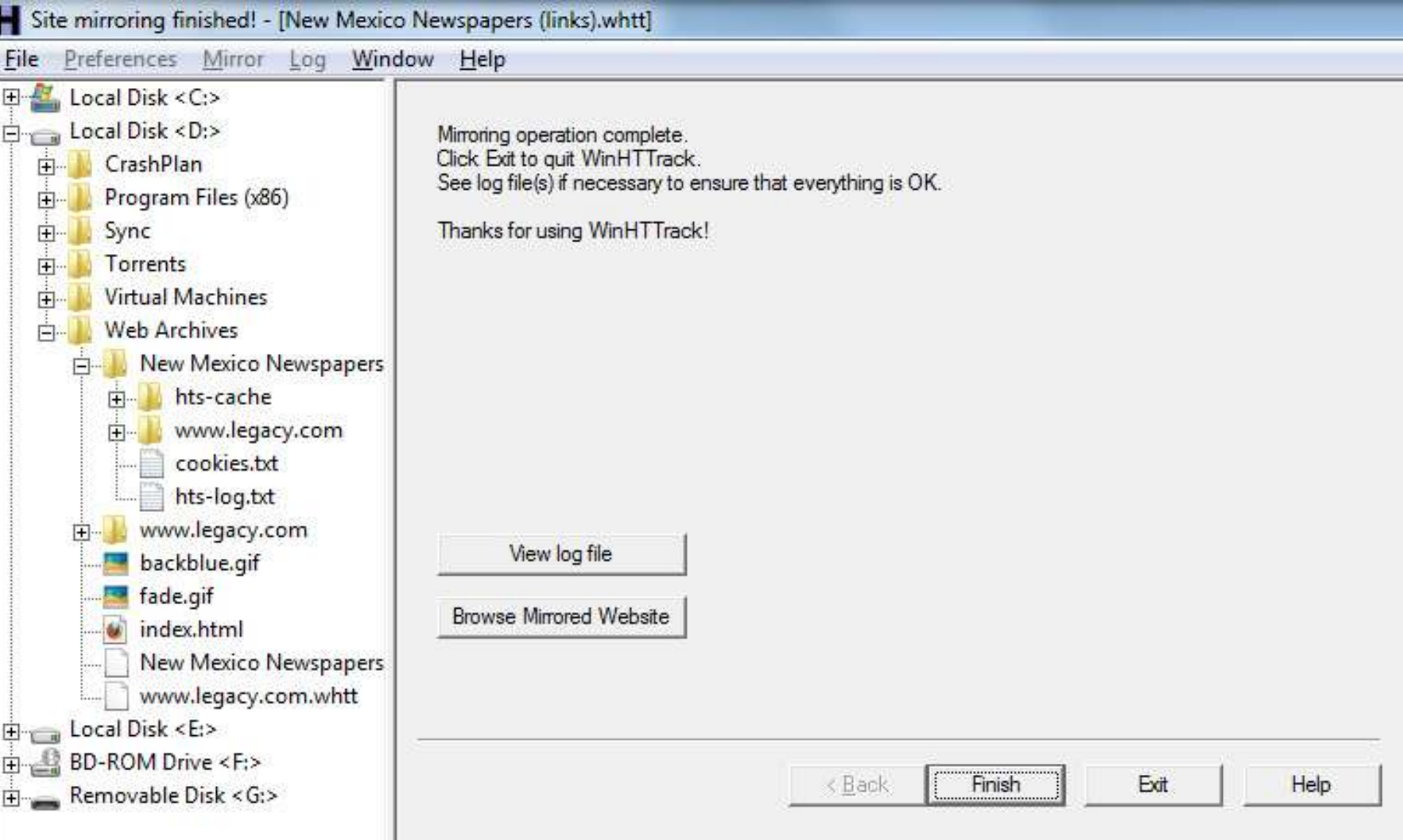
# start by scanning links

- 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/](http://www.legacy.com/ns/)
  - [www.legacy.com/memorial-sites/sandy-hook-school-tragedy/](http://www.legacy.com/memorial-sites/sandy-hook-school-tragedy/)
  - <https://www.legacy.com/NS/>
  - [www.legacy.com/obituaries/alamogordonews/](http://www.legacy.com/obituaries/alamogordonews/)
  - [www.legacy.com/obituaries/demingheadlight/](http://www.legacy.com/obituaries/demingheadlight/)
  - [www.legacy.com/obituaries/daily-times/](http://www.legacy.com/obituaries/daily-times/)
  - [www.legacy.com/obituaries/currentargus/](http://www.legacy.com/obituaries/currentargus/)
  - [www.legacy.com/obituaries/lascrucesbulletin/](http://www.legacy.com/obituaries/lascrucesbulletin/)
  - [www.legacy.com/obituaries/lcsun-news/](http://www.legacy.com/obituaries/lcsun-news/)
  - [www.legacy.com/obituaries/lasvegasoptic/](http://www.legacy.com/obituaries/lasvegasoptic/)
  - [www.legacy.com/obituaries/lamonitor/](http://www.legacy.com/obituaries/lamonitor/)
  - [www.legacy.com/obituaries/ruidosonews/](http://www.legacy.com/obituaries/ruidosonews/)
  - [www.legacy.com/obituaries/santafenewmexican/](http://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/](http://www.legacy.com/ns/) (and everything “below” in path)
  - [www.legacy.com/memorial-sites/](http://www.legacy.com/memorial-sites/) (and everything “below” in path)
  - <https://www.legacy.com/NS/> (and everything “below” in path)
  - [cache.legacy.com/](http://cache.legacy.com/) (and everything “below” in path)
  - non-html files
- include:
  - [www.legacy.com/obituaries/](http://www.legacy.com/obituaries/) (and everything “below” in path)
  - [www.legacy.com/guestbooks/](http://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

**H** Site mirroring in progress [6/48 (+16), 3542636 bytes] - [New Mexico Newspapers (html).whtt]

File Preferences Mirror Log Window Help

Local Disk <C:>  
Local Disk <D:>  
  CrashPlan  
  Program Files (x86)  
  Sync  
  Torrents  
  Virtual Machines  
  Web Archives  
    New Mexico Newspapers  
    www.legacy.com  
      backblue.gif  
      fade.gif  
      index.html  
      New Mexico Newspapers  
      www.legacy.com.whtt  
Local Disk <E:>  
BD-ROM Drive <F:>  
Removable Disk <G:>

In progress: Parsing HTML file..

Information

Bytes saved:	3,37MiB	Links scanned:	6/48 (+16)
Time:	18s	Files written:	16
Transfer rate:	26,27KiB/s (25,18KiB/s)	Files updated:	0
Active connections:	4	Errors:	0

☒ Actions:

scanning	www.legacy.com/guestbooks	<div></div>	SKIP
request	blog.legacy.com/about-2/	<div></div>	SKIP
request	blog.legacy.co...ie=test; path= /	<div></div>	SKIP
request	g.legacy.com/comments/feed/	<div></div>	SKIP
		<div></div>	SKIP
		<div></div>	SKIP
		<div></div>	SKIP
		<div></div>	SKIP
		<div></div>	SKIP
		<div></div>	SKIP
		<div></div>	SKIP
		<div></div>	SKIP
		<div></div>	SKIP
		<div></div>	SKIP
		<div></div>	SKIP
		<div></div>	SKIP

< Back Next > Cancel Help



# 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

# Guest Book

[Home](#) • [View Obituary](#)

 Remember this Guest Book

 Index View

 Print Entries



## GARRETT LEWIS

*This Guest Book will remain online permanently.*

[Sign Guest Book](#) • [Add Photo to Gallery](#) • [Light a Candle](#)

Add a message to the Guest Book

February 09, 2013

Garrett Lewis was a multi-talented actor, singer, dancer and artist. I has the pleasure of working with him on Broadway in "Vintage '60" and in Abbe Lane's nightclub act. His "design" eye took him into set decoration where he received multiple Oscar nominations. He was also the best friend anyone could ever have: kind, funny, caring and loyal. My wife and I will never forget him.

~ Larry Billman, *California*

Legacy.com: [Garrett Lewis Guest Book](#)



Nicholas Taylor  
[@nullhandle](#)