

What Makes a Good Web Archive?

Brenda Reyes Ayala, Ph.D Student
Brenda.Reyes@unt.edu
University of North Texas Libraries
Best Practices Exchange Conference
Montgomery, Alabama
November 19, 2014

What is Web Archiving?

Web archiving: the process of storing and maintaining Internet resources to preserve them as a historical, informational, legal, or evidential record.

Many state and federal archives, agencies, and universities in the United States archive web content

Examples:

- › [The Internet Archive](#)
- › [UNT Cybercemetery](#)
- › [Columbia University Human Rights Web Archive](#)
- › [Library of Congress Web Archive](#)

Quality in a Web Archive

Ideally, an archived web site is fully functional and identical in every way to the original.

However, many factors make perfect quality impossible:

- Rich media such as video, audio, and scripts are difficult to capture and render
- Many sites do not allow crawlers or limit their scope.
Ex: social media sites
- A lot of content is behind pay walls or requires log in information to be accessed
- Database-driven sites cannot be captured optimally with current technologies
- Content is missing from the archived site that makes it difficult to understand or use

High-quality Archived Site

Wayback - External links, forms, and search boxes may not function within this collection. Url: <http://frackfreedenton.com/ten-reasons-to-ban-fracking-in-denton/> time: 13:15:44 Nov 06, 2014 [[hide](#)]



Donate Volunteer Stay Updated  

VOTE FOR THE BAN

HOME FRACKING IN DENTON READ THE BALLOT SCIENTIFIC STUDIES RESOURCES WHO WE ARE

10 Reasons to Ban Fracking in Denton

#1: Fracking is bad news for property rights

Fracking has been shown to [decrease nearby residential property values](#) and can [make it more difficult](#) to get a home mortgage and insurance. Forty-three families in Denton have [sued EagleRidge Energy](#) for \$25 million in damage, nuisance, and trespass. The right to property comes with an obligation to [respect others' rights to their property](#). Fracking companies are abusing their rights by harming others.

#2: Fracking poisons our neighborhoods

[Evidence is mounting](#) about the health risks of fracking, and air samples from a Denton neighborhood near gas wells [showed benzene](#) - a carcinogen - at unsafe levels. Fracking [is a nasty business](#). Would you want [this](#) near your home? In Denton, fracking is allowed [250 feet](#) from homes and playgrounds.

#3: Fracking is dangerous

All of the [270 gas wells in Denton](#) are susceptible to accidents like [this explosion](#) that caused an entire Texas town to be evacuated. One of EagleRidge Energy's wells in Denton experienced a [blowout for over 14 hours](#) that [spewed](#) thousands of gallons of [undisclosed and toxic chemicals](#) into the environment.

#4: Fracking is a uniquely invasive industry

Fracking is the [only industry allowed to operate in residential areas](#) and it is also the only industry permitted by law to release [non-disclosed](#) and [unmonitored](#) toxic chemicals into the environment.

#5: Fracking harms air quality

Our Denton Blog



[Frack Free Denton statement in response to industry lawsuit to stop the ban](#)

November 5, 2014



[We WON! A few thoughts from president Cathy McMullen](#)

November 5, 2014



[Health board presentation: Health Risks of Oil and Gas Development](#)

November 3, 2014



[The Fracking Kings](#)

November 2, 2014

[View Post Archive](#)


[Tweets by @frackfreedenton](#)

Content has been captured and can be played back. Appearance resembles the original. External links work, which is important for a blog.

Medium-quality Archived Site

[An easy way to use Twitter via text messenger.](#)

Follow Following Unfollow Blocked Unblock Pending Cancel

 **DentonRC** @DentonRC Nov 4

City of #Denton releases statement on proposition to ban hydraulic fracturing on behalf of Mayor Watts pic.twitter.com/chpxAvdl27

0 replies 44 retweets 28 favorites


Reply

Retweet 44 Retweeted 44

Favorite 28 Favorited 28

More

- Embed Tweet

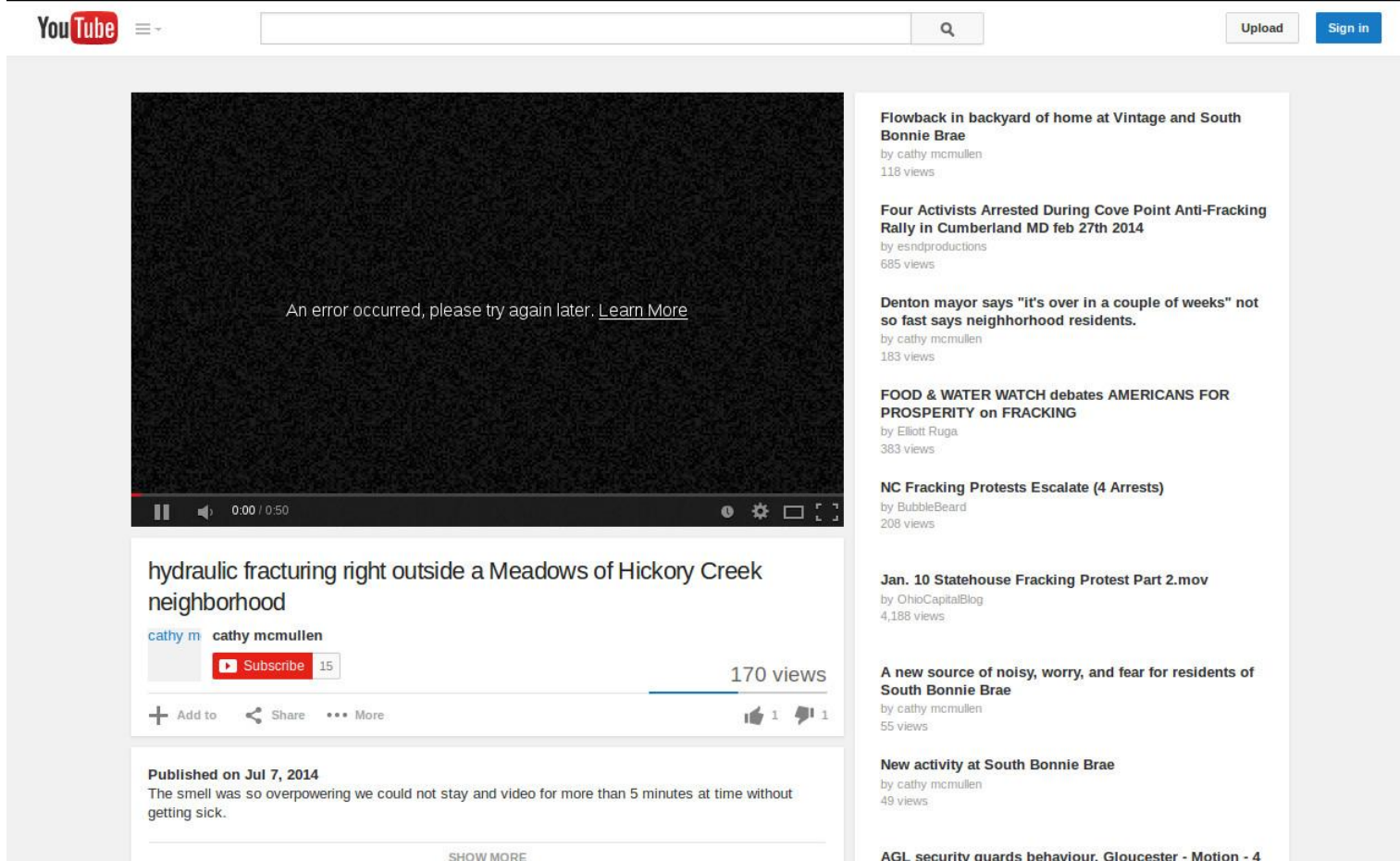
 **Denton Record-Chronicle**
20 minutes ago

City of Denton just sent the following statement to the media in light of early voting results regarding the proposition to ban hydraulic fracturing on behalf of Mayor Chris Watts:

"As I have stated numerous times, the democratic process is alive and well in Denton. Hydraulic fracturing, as determined by our citizens, will be prohibited in the Denton city limits. The City Council is committed to defending the ordinance and will exercise the legal remedies that are available to us should the ordinance be challenged. The City Council is committed to continuing the review of our gas well ordinance to ensure the utmost health, safety, and welfare of our residents, and we will continue to work with industry representatives to ensure full compliance with our gas well drilling ordinance."

Style information is missing but basic intellectual content is still present.

Low-quality Archived Site



The image shows a screenshot of a YouTube video player. The video player area is mostly black with the text "An error occurred, please try again later. [Learn More](#)" centered. Below the player, the video title is "hydraulic fracturing right outside a Meadows of Hickory Creek neighborhood" by user "cathy mcmullen". The video has 170 views and a "Subscribe" button with 15 subscribers. The description states: "Published on Jul 7, 2014. The smell was so overpowering we could not stay and video for more than 5 minutes at time without getting sick." To the right of the video player is a list of related videos:

- Flowback in backyard of home at Vintage and South Bonnie Brae** by cathy mcmullen, 118 views
- Four Activists Arrested During Cove Point Anti-Fracking Rally in Cumberland MD feb 27th 2014** by esndproductions, 685 views
- Denton mayor says "it's over in a couple of weeks" not so fast says neighborhood residents.** by cathy mcmullen, 183 views
- FOOD & WATER WATCH debates AMERICANS FOR PROSPERITY on FRACKING** by Elliott Ruga, 383 views
- NC Fracking Protests Escalate (4 Arrests)** by BubbleBeard, 208 views
- Jan. 10 Statehouse Fracking Protest Part 2.mov** by OhioCapitalBlog, 4,188 views
- A new source of noisy, worry, and fear for residents of South Bonnie Brae** by cathy mcmullen, 55 views
- New activity at South Bonnie Brae** by cathy mcmullen, 49 views
- AGL security guards behaviour. Gloucester - Motion - 4**

Content has been captured, but cannot be played back correctly.

QA process at UNT Libraries

1. Crawl engineer configures and launches the crawl
2. Student looks at archived sites and records any quality problems in a spreadsheet
3. Crawl engineer does follow-up QA: identifies the issues leading to the quality problems and decides how they can be addressed
4. Crawl engineer configures and launches a patch crawl

QA spreadsheet

1. Identifying info
 - 1.1. Name & URL of site(s)
 - 1.2. Present on live web? Y/N
 - 1.3. Scope of the crawl
 - 1.4. Priority
 - 1.5. Depth of site that was checked
 - 1.6. Does streaming audio/video work correctly?
 - 1.7. Do navigational menus work properly?
 - 1.8. Does the site's appearance resemble the original?
 - 1.9. Are there parts of the site missing that should have been captured?
 - 1.10. Robots.txt rules (ignored/followed)
 - 1.11. List of tools used to do QA

QA spreadsheet (cont.)

2. Quality problems & their severity (high/low)
 - 2.1. Missing content
 - 2.2. Wrong representation
 - 2.3. Other errors
3. Subdomains not crawled
 - 3.1. Ignored subdomains
 - 3.2. Ignored parts of subdomains
4. Should be added to crawl
 - 4.1. Link to be added
 - 4.2. URL that contained the link

Advantages and Disadvantages

- Requires time and effort
- Impossible to navigate an entire website and its outlinks
- Provides an opportunity to investigate the underlying causes of the quality problems & adjust software accordingly
- Results in rich, detailed documentation about every captured resource

QA at Archive-It

1. Run a test crawl of sites to be archived
2. Identify possible problems and adjust crawl settings if necessary
3. Run a QA report on archived content. This lets you know why a resource was not archived
4. Browse archived sites manually
5. Conduct a patch crawl if necessary

QA at the Internet Archive (LOC)

1. Precrawl. Any possible problems are communicated to the LOC web archiving team
2. Production crawl. Generate WAT files for the crawl. A WAT file contains metadata for each WARC file and is extremely useful for data analysis
3. Automated QA. Perform browser analysis and link analysis on WAT files. Add all the missing content to a “to-be-crawled” list
4. Patch crawl. Identify whether the quality problem is a replay issue or a capture issue. Crawl seeds in list
5. Human QA. Curators browse the archived content

Aspects of Quality

- Can be measured horizontally (a particular archived site is of high quality) or vertically (an entire collection is of high quality)
- Horizontal & vertical measurements can differ (a high-quality archived site inside a medium-quality archive)
- Can have both objective and subjective dimensions

Aspects of Quality (cont.)

- Correspondence: a one-to-one correspondence between the original resource and the archived resource
- Completeness: archived resource contains all its constituent elements
- Coherence: archived resource integrates diverse elements in a logical and consistent manner
- Integrity: The data elements are uncorrupted and error-free

References

Reyes Ayala, Brenda; Phillips, Mark Edward & Ko, Lauren.
Current Quality Assurance Practices in Web Archiving.
UNT Digital Library. [http://digital.library.unt.edu/ark:
/67531/metadc333026/](http://digital.library.unt.edu/ark:/67531/metadc333026/).