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

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

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