# Judging Visual Correspondence in Web Archives using Similarity Measurements

Brenda Reyes Ayala[1] Ella Hitchcock[1] James Sun[1]

[1]University of Alberta

Workshop on Web Archiving and Digital Libraries, Joint Conference on Digital Libraries (JCDL) June 6, 2019

# Overview


1. Introduction
2. Research Questions
3. Methodology, Data Collection, Analysis
4. Findings and Discussion
5. References

The visual representation of information on the web has become an important part of judging the quality of archived web pages [3].

I define visual correspondence in a web archive as “the similarity in appearance between the original website and the archived website” [5].

# Why quality matters in a web archive

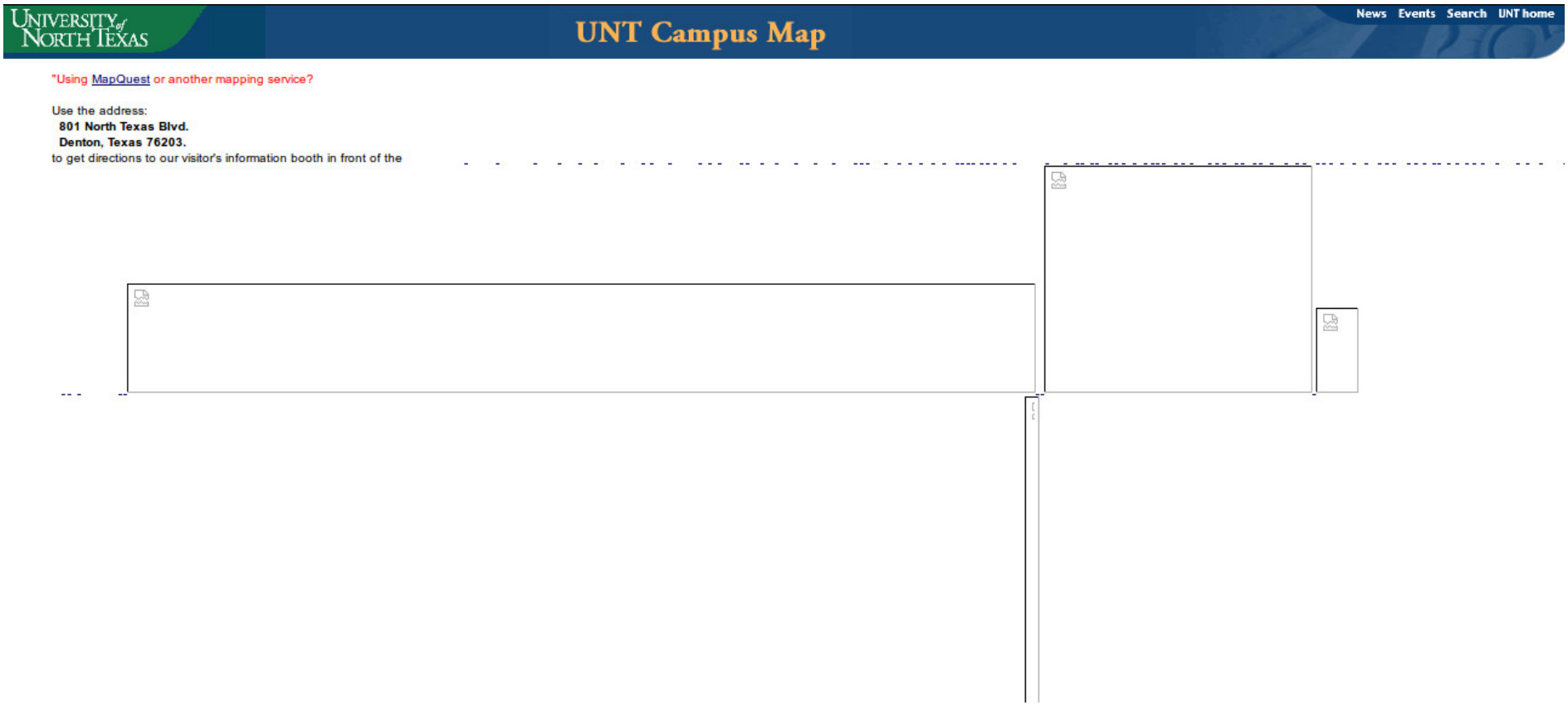



**Figure:** Screenshot of an archived version of the UNT Campus Map from 2004. Retrieved from `http://web.archive.org/web/20040722064240/http://www.unt.edu/pais/map/campusmap.htm`

# Research Questions and Purpose

How effective are different similarity measures at measuring the visual correspondence between an archived website and its live counterpart?

We examine how the visual correspondence of an archived website can be measured using popular image similarity measures. Using these measures we evaluate how visual correspondence can be used as an indication of overall archive quality.

# The dataset used

1. “Idle No More” [6]: topical web archive that preserves websites related to “Idle No More”, a Canadian political movement encompassing environmental concerns and the rights of indigenous communities.
2. Western Canadian Arts collection [7]: preserves the born digital resources created by filmmakers in Western Canada.
3. British Library’s OA web archive [4]: UK websites that can be made available online according to British legal deposit laws.

# Generating the screenshots

Created set of tools called "wa screenshot compare", currently freely available as a Github repository at
`https://github.com/reyesayala/wa_screenshot_compare`

1. Take a seedlist as input and generate screenshots of the live websites using Pyppeter (a Python port of the Puppeteer screenshot sofware) and and a headless instance of the Chrome browser.
2. Generates list of all archived versions of the live sites that are available from the University of Alberta's Archive-It collection.
3. Takes screenshots of the archived websites (with or without the banners).

# Characteristics of Web Archive Collections Used for Similarity Judgments

We categorized as "lost", those websites that returned an HTTP status code other than 200 and were not redirects.

| Collection | No. Seeds | No. Seeds Still Available | % Collection Still Available |
|---|---|---|---|
| Idle No More | 196 | 182 | 92.86 |
| Western Canadian Arts | 101 | 95 | 94.06 |
| UK Open Access | 659 | 516 | 78.30 |

# Issues encountered during the screenshot process

Not a trivial process despite our initial assumptions:

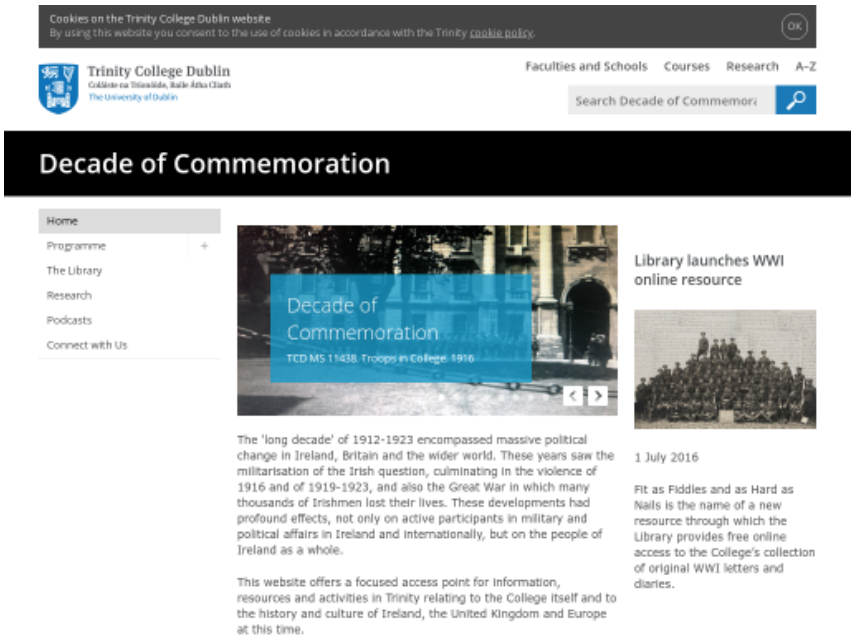
- The use of banners to indicate to users that they are viewing an archived website. Appended the text “id_” to the url of the archived websites, but this approach often breaks the CSS styling of the archived site, resulting in a screenshot that was even farther from the actual appearance of the archived website.
- Link rot. Highlights the importance of conducting visual quality assessments early in the web archiving process, while the websites collected are still online and accessible for comparison.

# Calculating similarity

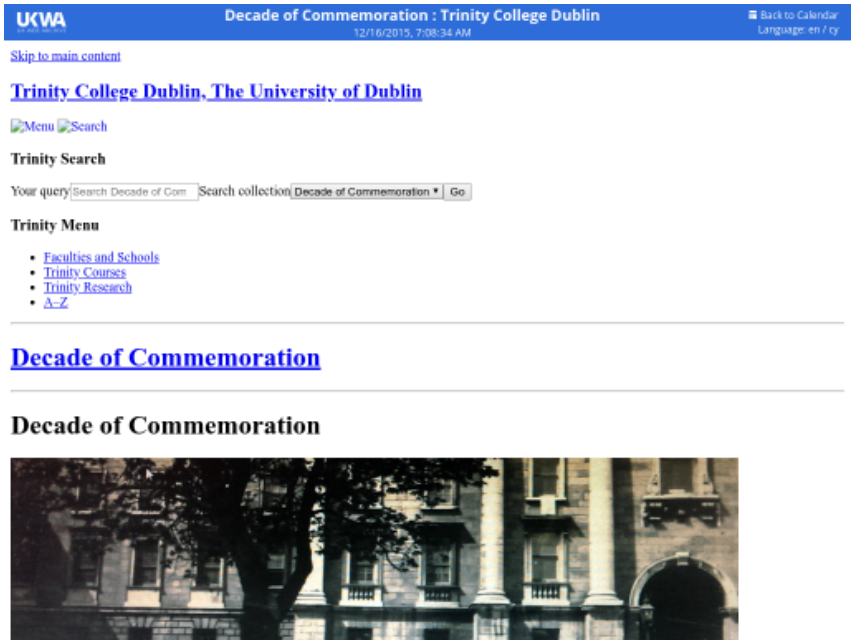
Based on popular image similarity measures: Structural Similarity Index (SSIM), Mean Squared Error (MSE), and “vector distance”, which produces the distance between the RGB values of each screenshot. We changed this metric slightly by subtracting every result from 100, thus giving us the percentage similarity between a pair of images.

- SSIM: calculates similarity on a scale of [-1,1]. 1 is perfect similarity.
- MSE: calculates similarity on a scale of [0, $\infty$]. 0 is perfect similarity.
- Vector distance: calculates similarity on a scale [0-1]. 1 is perfect similarity.

# Example: A "medium" quality archived website




Screenshot of current, live website




Screenshot of archived website

**Figure:** Comparison of images for the website "Trinity College Dublin: Decade of Commemoration". SSIM = 0.51, MSE = 61536.53, Vector Distance = 59.87

# Example: A "low" quality archived website


Screenshot of current, live website


Screenshot of archived website

**Figure:** Comparison of images for the website of the play "Nye & Jennie". SSIM = 0.28, MSE = 169603.88, Vector Distance = 8.83

# Correlation between similarity measures in web archives

Performed a correlation analysis on all our similarity scores for the three web archives collections to determine if there were relationships between different similarity measures.

| Collection | SSIM - MSE | MSE - Vector | SSIM - Vector |
|---|---|---|---|
| Idle No More | -0.61 | -0.97 | 0.61 |
| Western Canadian Arts | -0.72 | -0.98 | 0.78 |
| UK Open Access | -0.63 | -0.97 | 0.69 |
| All | -0.65 | -0.97 | 0.71 |

# Correlation between similarity measures in web archives (2)

- Moderate negative correlation between SSIM and MSE score.
- Very strong negative correlation between MSE and Vector distance.
- Moderate-to-strong (depending on the collection) correlation between SSIM and vector distance scores.

Almost perfect negative relations between MSE and vector distance suggests that one measure might be easily substituted for another. Because we found MSE scores relatively difficult to interpret, we recommend the use of vector distance or SSIM as measures of similarity.

# Hot off the presses...

Q: Does taking screenshots without the banner have a significant impact on the similarity scores?
For UK OA collection, performed the Wilcoxon signed-rank test (paired samples) on the similarity scores for images with and without the banner. 225 pairs of images.

| Similarity measure | Median similarity score | |
|---|---|---|
| | with banner | without banner |
| SSIM | 0.77 | 0.78 |
| MSE | 39547.61 | 39604.71 |
| Vector distance | 68.13 | 68.49 |

# Hot off the presses...(2)

Q: Does taking screenshots without the banner have a significant impact on the similarity scores?

| | Significance ($p$) |
|---|---|
| SSIM w/o banner vs SSIM w/ banner | $< 0.000$ |
| MSE w/o banner vs MSE w/ banner | 0.004 |
| Vector w/o banner vs Vector w/ banner | $< 0.000$ |

A: There were statistically significant differences between screenshots with the banner and those without the banner.

# Conclusions and next steps

Image similarity metrics can be successfully applied in order to measure the visual correspondence (and thus visual quality) of archived websites. Our results indicated that these metrics were able to successfully distinguish between website captures of poor quality and those of higher quality.
Next steps:

- Explore other measures of similarity.
- Improve code to increase performance.
- Conduct experiments to find which similarity measures most closely match up with human judgments of visual correspondence in a web archive.

[1] British Library. (2015, December). Trinity College Dublin: Decade of Commemoration. Retrieved from https://www.webarchive.org.uk/wayback/archive/1/https://www.
commemoration/

[2] British Library. (2018, June). Nye and Jennie: A working class tale of life, labour and love. Retrieved from https://www.webarchive.org.uk/wayback/archive/1/https://www.

[3] Gyllstrom, K., Eickhoff, C., de Vries, A.P. & Moens, M. (2012). The downside of markup: Examining the harmful effects of CSS and Javascript on indexing today's web. *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, 1990-1994. doi: 10.1145/2396761.2398558

[4] Jackson, A. (2019). UKWA Manual QA dataset. Retrieved from https://github.com/iipc/qa2019/tree/master/ukwa-manual-qa-dataset

[5] Reyes Ayala, B. (2018). *A grounded theory of information quality in web archives*. (Doctoral dissertation). Retrieved from https://digital.library.unt.edu/ark:/67531/metadc1248497/

[6] University of Alberta. (n.d). Idle No More collection. Retrieved from https://archive-it.org/collections/3490

[7] University of Alberta. (n.d). Western Canadian Arts collection. Retrieved from https://archive-it.org/collections/6296