Towards a better QA process: Automatic detection of quality problems in archived websites using visual comparisons

Dr. Brenda Reyes Ayala[1],

[1]Associate Professor
School of Library and Information Studies, University of Alberta
Edmonton, Alberta, Canada
brenda dot reyes at ualberta dot ca

February 20, 2025

21st Conference on Information and Research Science Connecting to Digital and Library Science (IRCDL 2025), Udine, Italy

Overview

The context of web archives

- Web archiving is the practice of preserving web content.
- Carried out by institutions such as libraries, governments, and universities to preserve digital cultural heritage.
- Day-to-day tasks for archiving the web:
  1. Appraisal and Selection: institutions decide specifically which websites they want to collect.
  2. Scoping: institutions may opt to archive portions of a website, whole sites, or even entire web domains.
  3. Data Capture: institutions fine-tune how they want to capture their data through decisions about crawl (capture) frequency and types of files to archive or not archive.
  4. Storage and Organization: This step includes a temporary or long-term storage plan for the archived data.
  5. **Quality Assurance and Analysis**: institutions review what they have archived and how well the resulting collection satisfies the goals they set at the beginning of the life cycle.

The context of web archives, II

- The Quality Assurance (QA) process at most institutions is still a manual one, requiring web archivists to manually inspect hundreds or thousands of archived websites to compare them to the original, live websites (Reyes Ayala, Phillips, & Ko, 2014).
- This poses significant burdens to an institution in terms of time and resources.
- Javascript, AJAX, and Cascading Style Sheets (CSS) has complicated the process of web archiving, making it difficult to create archived websites that are as close as possible to the original website (J. F. Brunelle, Kelly, Weigle, & Nelson, 2016).

Objectives

- Can we automate part of the QA process?
- We examine how image similarity measures can be used to detect problems with the quality of archived websites.
- Previously, we introduced the use of image similarity metrics to detect problems with visual correspondence in archived websites (Reyes Ayala, Hitchcock, & Sun, 2019).
- We follow up our original research by examining more visual similarity metrics and determining if these measures match human judgments of the quality of archived websites.

Research Questions

1. How do different image similarity measures perform at measuring the visual correspondence between an archived website and its live counterpart?
2. Are similarity measures able to detect low-quality archived websites in a way that is consistent with human judgments of quality?

Archived websites with missing elements

- It is common for archived websites to have missing elements; however, not all missing elements are created equal.
  item (J. Brunelle, Kelly, SalahEldeen, Weigle, & Nelson, 2015) examined the importance of missing elements or resources and their impact on the quality of archived websites in their paper.
- Missing embedded resources results in a “damaged” archived website.
- Authors proposed a new metric to assess damage that is based on three factors: the MIME type, size, and location of the embedded resource (J. Brunelle et al., 2015).

Replay quality

- (Kiesel et al., 2018) focused on the reproduction (replay) quality of archived websites.
- Introduced the Webis Web Archiver tool, which relied on emulating user interactions with a web page while recording all network traffic.
- The authors defined reproduction quality as thus: “the more individual users that scroll down a web page are affected in their perception or use of the web page by visual differences between the original web page and its reproduction, the smaller the reproduction quality for that web page.”
- Reproduction quality was assessed on a 5-point Likert scale from no effect (score 1) to unusable reproduction (score 5).

Reference rot and soft 404s

Reference rot, which has two components, as identified by (Jones et al., 2016):

1. Link rot: The resource identified by a URI vanishes from the web. As a result, a URI reference to the resource ceases to provide access to referenced content.
2. Content drift: The resource identified by a URI changes over time. The resources content evolves and can change to such an extent that it ceases to be representative of the content that was originally referenced.

Soft 404s when websites redirect failed URLs to a site's homepage, masking the standard 404 return code (Meneses, Furuta, & Shipman, 2012). As a result, web resources that have been lost may appear to still exist.

Dataset

Used the following web archive collections:

1. Idle No More (INM): pages related to Idle No More, a Canadian political movement encompassing environmental concerns and the rights of indigenous communities (University of Alberta, n.da).
2. Fort McMurray Wildfire 2016 (FMW): pages related to the Fort McMurray Wildfire of 2016 in the province of Alberta, Canada (University of Alberta, 2016).
3. Western Canadian Arts (WCA): websites created by filmmakers in Western Canada (University of Alberta, n.db).
4. Government of Canada (GoC): Canadian government pages (Libraries and Archives of Canada, 2016).

Process

1. Generate the screenshots
2. Address reference rot
3. Calculate similarity
4. Compare measures
5. Evaluation

Generate the screenshots and address reference rot

- Created a set of tools available as a Github repository [1].
- Uses Pyppeter (a Python port of the Puppeteer screenshot sofware) and and a headless instance of the Chrome browser [2].
- Takes screenshot of both the live website and its archived version.
- Research Assistants inspected each of the live websites and compared them to their archived versions.
- If a website had drifted or had experienced link rot, it was removed from the dataset.

[1] https://github.com/reyesayala/wa_screenshot_compare
[2] https://github.com/miyakogi/pyppeteer

Calculate similarity

The software calculated similarity between live and archived screenshots several popular image similarity measures:

1. Structural Similarity Index (SSIM)
2. Mean Squared Error (MSE)
3. Normalized Mean Square Error (NMSE)
4. Perceptual Hash (P-Hash)
5. Peak Signal to Noise Ratio (PSNR)
6. Percentage Difference

Example of live/archived screenshot and its similarity measures

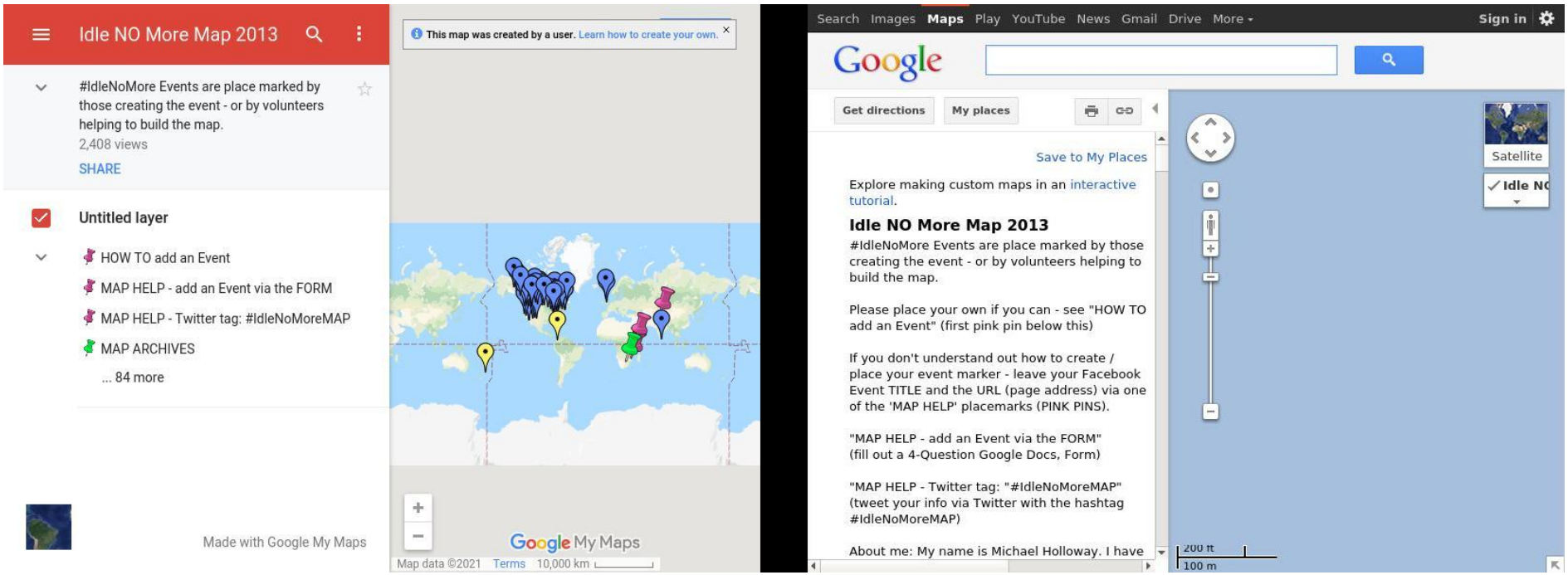

Figure: Comparison of screenshots of the live website (left) and the archived website (right) for "Idle No More Map 2013". SSIM = 0.39, MSE = 3646.68, Percentage similarity = 84.88, P-hash = 32, NRMSE = 0.25, PSNR = 12.51

Compare measures

- Our ideal similarity measure would be
  1. Easy to understand for a non-expert audience.
  2. Illustrate meaningful differences between low and high-quality websites.
- MSE, PSNR, and P-hash did not meet the first criterion because they have no proper upper bound.
- A correlation analysis indicated a strong negative correlation (-0.94) between percentage similarity and NRMSE, indicating NRMSE could be substituted for percentage difference.
- We continued our analysis with NRMSE and SSIM.

Research Question #2

We still needed to see if

Our calculated similarity measures matched how humans assessed quality in archived websites, and thus how web archivists might evaluate the quality of an archived website during the QA process.

Evaluation

- Used Amazon Mechanical Turk (AMT) to solicit opinions from human judges. AMT provides a large pool of participants who receive payment for completing tasks.
- 221 image pairs, 68 from the INM collection, 18 from the WCA collection, 73 from the FMW collection, and 62 from the GoC collection. Each image pair was judged by two unique participants each, for a total of 442 judgments.
- Most study participants would likely be unfamiliar with web archiving, and might struggle to understand the terminology and processes that are specific to archiving websites.
- To avoid confusion, we instead asked the following question: *We saved this website for you. Did we do a good job of saving the website?* This approach was directly informed by that of (J. Brunelle et al., 2015).

Evaluation II

On the left-hand side, we presented a screenshot of the live URL; on the right-hand side, we presented a screenshot of the archived URL. Participants had the following options:

1. Yes, the copy looks exactly like the original (perfect similarity, coded as “high quality”)
2. Yes, the copy looks a little worse than the original, but the differences are small (small differences, coded as “high quality”)
3. No, the copied website looks worse than the original (low quality, coded as “low quality”)
4. Other, please describe (other)

Cohen’s κ, a measure of inter-rater agreement for categorical scales, was run to determine agreement between two participants (Laerd Statistics, 2015). There was substantial agreement between the two respondents’ judgments, $\kappa = .76, p < .001$.

Significance testing

- A one-way multivariate analysis of variance (MANOVA) test was run to determine if there were differences in SSIM and NRMSE scores between archived websites judged to be of high quality and archived websites judged to be of low quality.
- Result: scores for high quality archived websites and low quality archived websites were statistically significantly different: $F(2, 222) = 44.95, p < .001$
- We conducted follow-up univariate ANOVAs
- Result: Both SSIM scores ($F(1, 223) = 10.53, p = .001$; partial $\eta^2 = 0.05$) and NRMSE scores ($F(1, 223) = 89.52, p < .001$; partial $\eta^2 = 0.29$) were statistically significantly different between high quality and low quality archived websites

Discussion

How do different image similarity measures perform at measuring the visual correspondence between an archived website and its live counterpart?

We found that SSIM, NRMSE, MSE, PSNR, percentage difference, and P-hash measures are are all suitable for measuring image similarity. However, we discarded MSE, PSNR, and P-hash scores because they are more difficult to interpret. We also saw that NRMSE had a very strong correlation to percentage difference, and could thus take its place. SSIM and NRMSE scores are thus recommended for detecting visual quality problems in archived websites.

Discussion II

Are similarity measures able to detect low quality archived websites in a way that is consistent with human judgements of quality?

Our statistical tests indicated that both SSIM and NRMSE scores produced significantly different scores for high-quality and low-quality archived websites. Thus, they are able to distinguish between high-quality and low-quality archived websites.

Conclusion

Looking ahead, the QA process for institutions engaged in web archiving could become the following:

1. After an initial crawl of the seedlist, the code presented here is run to take screenshots of both the archived websites and their live counterparts.
2. A measure of similarity is calculated that indicates the visual correspondence between the archived website and its original version.
3. Archived websites with similarity above a certain threshold are classed as high quality and left alone.
4. Archived websites with lower similarity scores are flagged for manual examination by a web archivist.
5. After manual examination, the web archivist can opt to re-crawl the website in order to increase its quality.

Conclusion II

- Our research was informed at every step by our understanding of the QA process in web archives, and of the needs and constraints of web archivists.
- Having methods such as the one presented here can allow institutions or researchers to effectively detect low-quality content without needing to manually inspect each archived website.

References I

Brunelle, J., Kelly, M., SalahEldeen, H., Weigle, M. C., & Nelson, M. L. (2015). Not all mementos are created equal: measuring the impact of missing resources. *International Journal on Digital Libraries*, 1-19. doi: 10.1007/s00799-015-0150-6

Brunelle, J. F., Kelly, M., Weigle, M. C., & Nelson, M. L. (2016, June). The impact of javascript on archivability. *International Journal on Digital Libraries*, *17*(2), 95–117. Retrieved from http://dx.doi.org/10.1007/s00799-015-0140-8 doi: 10.1007/s00799-015-0140-8

Jones, S. M., Van de Sompel, H., Shankar, H., Klein, M., Tobin, R., & Grover, C. (2016, 12). Scholarly context adrift: Three out of four uri references lead to changed content. *PLOS ONE*, *11*(12), 1-32. Retrieved from https://doi.org/10.1371/journal.pone.0167475 doi: 10.1371/journal.pone.0167475

References II

Kiesel, J., Kneist, F., Alshomary, M., Stein, B., Hagen, M., & Potthast, M. (2018, October). Reproducible web corpora: Interactive archiving with automatic quality assessment. *Journal of Data and Information Quality*, *10*(4). Retrieved from https://doi.org/10.1145/3239574 doi: 10.1145/3239574

Laerd Statistics. (2015). *Cohen's kappa using spss statistics.* Retrieved from https://statistics.laerd.com/

Libraries and Archives of Canada. (2016, February). *Government of canada 2016 collection.* Retrieved from https://archive-it.org/collections/7084

References III

Meneses, L., Furuta, R., & Shipman, F. (2012). Identifying “soft 404” error pages: Analyzing the lexical signatures of documents in distributed collections. In P. Zaphiris, G. Buchanan, E. Rasmussen, & F. Loizides (Eds.), *Theory and practice of digital libraries* (pp. 197–208). Berlin, Heidelberg: Springer Berlin Heidelberg.

Reyes Ayala, B., Hitchcock, E., & Sun, J. (2019). Using image similarity metrics to measure visual quality in web archives. In M. Klein, Z. Xie, & E. A. Fox (Eds.), *Proceedings of the 2019 web archiving & digital libraries workshop (WADL 2019), june 6, 2019, urbana-champaign, illinois, USA* (pp. 12–14). Virginia Tech University Libraries. Retrieved from `https://vtechworks.lib.vt.edu/bitstream/handle/10919/97987/WADL2019.pdf#page=12`

References IV

Reyes Ayala, B., Phillips, M. E., & Ko, L. (2014). *Current quality assurance practices in web archiving* (Research Report). Retrieved from http://digital.library.unt.edu/ark:/67531/metadc333026/

University of Alberta. (2016, March). *Fort McMurray wildfire 2016 collection.* Retrieved from https://archive-it.org/collections/7368

University of Alberta. (n.da, March). *Idle No More collection.* Retrieved from https://archive-it.org/collections/3490

University of Alberta. (n.db, March). *Western Canadian Arts collection.* Retrieved from https://archive-it.org/collections/6296