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

Brenda Reyes Ayala[1] (Associate Professor)

[1]University of Alberta, Faculty of Education, 11210 87 Ave, Edmonton AB T6G 2G5, Canada

Abstract

For web archivists, Quality Assurance (QA) is a lengthy manual process that involves inspecting hundreds or thousands of archived websites to see if they have been captured correctly, i.e., resemble the original. This paper describes how this process can be automated by using image comparison measures to detect quality problems in archived websites. To this end, we created a suite of Python tools to 1) create screenshots of live websites and their archived counterparts, and 2) calculate the image similarity between the screenshots. We tested our code on four web archive collections to test the efficacy and usefulness of six different image similarity measures. We compared their scores to human judgments of the quality of archived websites obtained from Amazon Mechanical Turk (AMT). Our results show that the Structural Similarity Index (SSIM) and the Normalized Root Mean Square (NRMSE) scores are able to distinguish between high and low-quality archived websites. Our research at every step was informed by the specific needs and challenges of web archivists. Having methods such as the one presented here can allow cultural heritage institutions or researchers to more quickly and effectively detect low-quality content and produce high-quality web archives.

Keywords

web archiving, quality assurance, similarity, web archives

1. Introduction

Web archiving is the practice of preserving web content. It is usually carried out by institutions such as libraries, governments, and universities for the purpose of preserving their digital cultural heritage. The Internet Archive's Archive-It subscription service (AIT) [1] described the following 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.

As has been noted by [2], 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. This can pose significant very burdens to an institution in terms of time and resources.

IRCDL 2025: 21st Conference on Information and Research Science Connecting to Digital and Library Science, February 20-21 2025, Udine, Italy

brenda.reyes@ualberta.ca (B. Reyes Ayala)
https://reyesayala.github.io/ (B. Reyes Ayala)
0000-0002-9342-3832 (B. Reyes Ayala)

The widespread use of Javascript, AJAX, and Cascading Style Sheets (CSS) has complicated the process of web archiving [3], making it difficult to create archived websites that are as close as possible to the original website. A high-quality archived website is one with a high degree of visual correspondence, defined as the "the similarity in appearance between the original website and the archived website", as initially defined by [4].

This paper examines how image similarity measures can be used to detect problems with the quality of archived websites in an automated way. In a previous work [5], we introduced the use of image similarity metrics to detect problems with visual correspondence in archived websites. In this paper, 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. We are interested in answering the following 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 judgements of quality?

2. Previous Work

[6] addressed quality problems that could affect the coherence of a web archive, such as off-topic web pages. Off-topic web pages are those that have, over time, moved away from the initial scope of the page. This can occur because the page has been hacked, its domain has expired, or the service has been discontinued. The authors compiled three different archived collections and experimented with several methods of detecting these off-topic webpages and with how to define threshold that separates the on-topic from the off-topic pages. According to their results, the cosine similarity method proved the best at detecting off-topic web pages. The authors also experimented with combining several similarity measures in an attempt to increase performance. The combination of the cosine similarity and word count methods yielded the best results [6].

It is common for archived websites to have missing elements; however, not all missing elements are created equal. [7] examined the importance of missing elements or resources and their impact on the quality of archived websites in their paper. Embedded resources are files such as images, videos, or CSS stylesheets, that are referenced in a website. They play a key role in ensuring the website looks and operates in the correct way. Missing embedded resources results in a "damaged" archived website. The authors proposed a new metric to assess damage that is based on three factors: the MIME type, size, and location of the embedded resource [7].

When websites disappear from the web, it is known as *reference rot*, which has two components, as identified by [8]:

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 resource's content evolves and can change to such an extent that it ceases to be representative of the content that was originally referenced.

Content drift also occurs when a website redirects failed URLs to a site's homepage, thus causing it to mask the standard 404 return code that occurs when there is a failure to access a web resource, known as *soft 404s*[9].

[10] focused on the reproduction (replay) quality of archived websites. To this end, they introduced the Webis Web Archiver tool, which relied on emulating user interactions with a web page while recording all network traffic. In order to evaluate their tools, the researchers recruited human evaluators through Amazon's Mechanical Turk to assess web pages. 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 to account for different levels of perceived severity, ranked from no effect (score 1) to unusable reproduction (score 5).

The lack of adequate technologies to address quality problems in web archives was highlighted by [11] in their 2019 paper. The authors stated that current web archiving technologies were optimized to either: 1) operate at scale or 2) provide high-quality archival captures, but not both. To address this imbalance, they introduced the Memento Tracer framework, which aimed to achieve both quality and quantity, by allowing the curator to determine the desired components of a web resource that should be archived. Klein et al. acknowledged that quality in web archives is often subjective, and thus focused on the extent to which URIs that should be captured are actually captured. The authors "expect that a high-quality archival record to contain at least the same number of URIs as its live website version" [11].

3. Methodology

3.1. The Dataset

We used four different web archives in order to apply the similarity metrics, three from the University of Alberta and one from the Government of Canada. All were created using the Archive-It subscription service [12]. The first three collections were created by the University of Alberta Libraries in an effort to preserve western Canadian cultural heritage on the web [13] and the last one was created by Library and Archives of Canada (LAC) as part of their efforts to preserve Canadian government websites.

1. Idle No More (INM): websites related to "Idle No More", a Canadian political movement encompassing environmental concerns and the rights of indigenous communities [14].
2. Fort McMurray Wildfire 2016 (FMW): websites related to the Fort McMurray Wildfire of 2016 in the province of Alberta, Canada [15].
3. Western Canadian Arts (WCA): websites created by filmmakers in Western Canada [16].
4. Government of Canada (GoC): Canadian government websites [17].

3.2. Generating the screenshots

3.2.1. The screenshot module

In order to measure the visual correspondence of an archived website to its live counterpart, we created a set of tools available as a Github repository [1]. Written in Python, these tools take a *seedlist* (list of URLs) 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]. The software then takes screenshots of the archived versions of these websites without the characteristic banner added by the AIT service, which increases accuracy of the comparison. The procedure is the following:

1. Read settings file with the seedlist
2. For each seed:
   a) Check if the website still exists. If it does, take a screenshot and store it as an image file
   b) Get the URLs for all of the archived versions of the seed preserved by AIT. These are the archived URLs
   c) Take a screenshot of each of the archived URLs and store it as an image file
3. Write the URLs and their corresponding image file names to a CSV file

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

3.3. Addressing Reference Rot

Reference rot proved to be the most serious challenge to our approach, as many of the original websites in our collections have suffered link rot or content drift. As [4] pointed out, visual correspondence can **only** be measured if the original website still exists, otherwise there can be no comparison. We categorized as "lost", those websites that returned an HTTP status code other than 200 and were not redirects (link rot).

Additionally, a qualitative analysis of our screenshots from the live web indicated that the collections suffered significant content drift. This posed a problem for our experiment, as screenshots of content that has drifted are not appropriate for quality comparisons by human judges. In order to determine the amount of content drift in our web archive collections, Research Assistants inspected each of the live websites and compared them to their archived versions. If a website had drifted according to the definition in [8], it was removed from the dataset and not used for the rest of the experiment.

3.4. Calculating similarity

3.4.1. The similarity module

The software we created then put the remaining sets of screenshots through a similarity analysis based on several popular image similarity measures: Structural Similarity Index (SSIM), Mean Squared Error (MSE), Normalized Mean Square Error (NMSE), Perceptual Hash (P-Hash), and Peak Signal to Noise Ratio (PSNR), as well as an additional measure called *percentage difference*. We chose the first five due to their popularity in the image comparison community and their accessibility, and the last one was added due to its intuitiveness and ease of implementation. We used the implementation of these measures available from the scikit-image library in Python [3]. The scikit-image library employs a representation of an image as a matrix of MxN dimensions, with each of its elements representing a pixel's Red, Blue, and Green (RGB) values as integers from 0 to 255. The procedure for calculating similarity measures is the following:

1. Reads CSV file with list of live screenshots and their corresponding archived screenshots
2. Reads settings file with desired similarity measures
3. For each (live screenshot, archived screenshot) pair:
    - Check if the live screenshot is blank (all black or all white). If blank, write file name of screenshot to list of blank screenshots
    - Check if images are equal in size. If not, crop the images to be of equal size. This is a requirement for most image similarity calculations
    - Calculate similarity scores
    - Output all similarity scores to a CSV file

3.4.2. Structural Similarity Index (SSIM)

The Structural Similarity Index introduces a structural component to the comparison process by taking into account the pixel vector positions in both images, comparing the properties of pixels in both images one pixel at a time, and preserving the position of the pixels as they are compared [18]. The similarity between two images x and y is thus defined as:

$$SSIM(x, y) = l(x, y)^{\alpha} \cdot c(x, y)^{\beta} \cdot s(x, y)^{\gamma} \tag{1}$$

where l, c, and s are functions of luminance, contrast, and structure, respectively, and α, β, and γ are their weights. SSIM calculates similarity on a scale of [-1,1], where 1 is perfect similarity and negative values occur when the image has been inverted.

[3] https://scikitimage.org/docs/stable/api/skimage.metrics.html#skimage.metrics.adapted_rand_error

3.4.3. Mean Square Error (MSE) and Normalized Root Mean Square Error (NRMSE)

MSE is the mean of the square of the difference in pixels between two images, defined as:

$$MSE = \frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1} (Y_i, j - X_i, j)^2 \tag{2}$$

$(Y_i, j - X_i, j)$ is the difference. between the RGB values of corresponding pixels X and Y, indexed by i, j in the original, "ground truth" image and the test image [19]. MSE calculates similarity on a scale of [0, ∞], where 0 is perfect similarity. MSE does not take into account the vector position of the pixels, meaning that if for example two images contained the same number of measured pixel values the images would be deemed similar, no matter the position of the pixels in the image [20]

The Root Mean Square Error (RMSE) is simply the square root of the MSE score for two images. If the RMSE is normalized to give it an upper bound, then the result is the NMRSE. The Python implementation in the scikit library uses the averaged Euclidean norm of the original, true image [4].

$$NRMSE = \frac{RMSE \cdot \sqrt{N}}{||X||} \tag{3}$$

where N is the size of the original image, and $||X||$ is the Frobenius norm of the original image. NRMSE uses a scale of [0,1] and as with MSE, 0 indicates perfect similarity.

3.4.4. Peak Signal to Noise Ratio (PSNR)

PSNR is another similarity measure based on MSE, and is defined as:

$$PSNR = 10 \cdot \frac{MAX^2}{MSE} \tag{4}$$

where MAX is the maximum value a pixel can take, in our case, 255 [19]

3.4.5. P-hash

P-hash [21] is a similarity measure that uses the features of an image to generate a distinct fingerprint. This way, two images can be compared by comparing their hashes. It calculates similarity on a scale [0, ∞]. 0 is perfect similarity.

3.4.6. Percentage Difference

We added a fourth measure, called “percentage difference” [22], which calculates the distance between the RGB values of each screenshot. The greater the distance, the greater the difference between the two images, and thus, the greater the difference between the two websites. We changed this metric slightly by subtracting every result from 100, thus giving us the *percentage similarity* between a pair of images. It uses a scale [0-100], where 100 is perfect similarity.

3.5. Comparison of Measures and Correlation Analysis

In comparing and contrasting our different similarity measures, we kept in mind two specific criteria. Our ideal similarity measure would be 1) easy to understand for a non-expert audience, and 2) illustrate meaningful differences between low and high-quality websites.

As an example, Figure 1 shows the live and archived versions of the site "Idle No More Map (2013)." The archived version has lost its background image, labels, and markers, rendering it ineffective as an interactive map. This low quality is reflected in the low similarity scores. After calculating the similarity scores for each of the image pairs in our collections, we discarded three measures, P-hash, PSNR, and

[4] https://scikit-image.org/docs/stable/api/skimage.metrics.html#skimage.metrics.adapted_rand_error

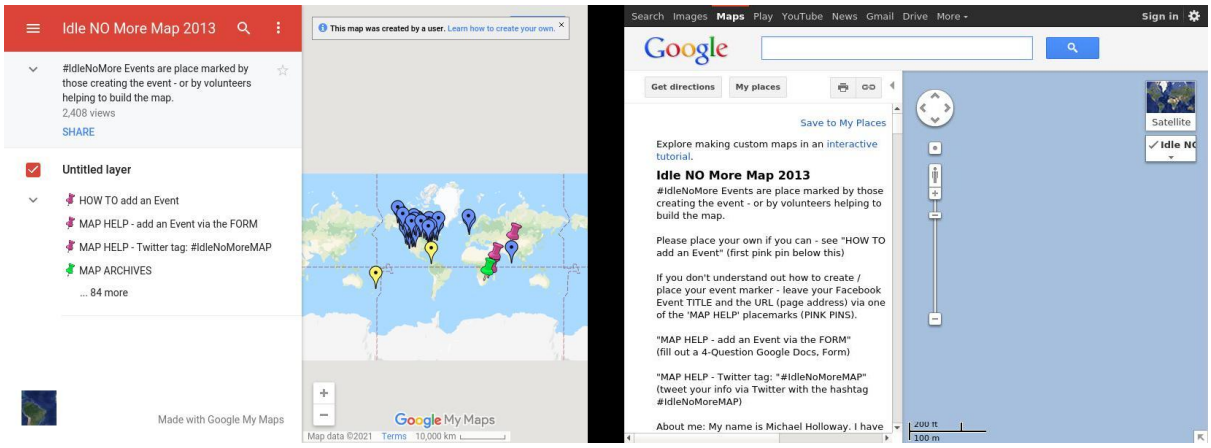

Figure 1: Comparison of screenshots of the live website (left) and the archived website (right) for "Idle No More Map 2013". SSIM = 0.39, MSE = 3646.68, Percentage similarity = 84.88, P-hash = 32, NRMSE = 0.25, PSNR = 12.51

Table 1
Pearson correlations between different similarity measures in web archives

	Percentage similarity	NRMSE
SSIM	0.45	-0.48
Percentage similarity	-	**-0.94**

MSE, because they have no proper upper bound, and thus were more difficult to interpret than other measures. PSNR also had the added disadvantage that, since it is based on MSE, when $MSE = 0$ (indicating two images are identical), PSNR becomes equal to ∞, which again might cause confusion to a non-expert audience. Taken together, MSE, PSNR, and P-hash did not meet the first criterion.

In order to determine if there were relationships between the remaining similarity measures, we performed a correlation analysis on all our similarity scores for the web archives collections. The results are shown in Table 1. The correlation coefficients indicate that there is a strong negative correlation between percentage similarity and NRMSE, a moderate positive correlation between SSIM and percentage similarity, and a moderate negative correlation between percentage similarity and NRMSE. Since strong correlations suggest that some measures might be easily substituted for another, it led us to narrow down our list further by discarding percentage similarity. We thus went ahead with our analysis with percentage SSIM and NRMSE, which proved promising for meeting our second criterion.

4. Evaluation

In order 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, we used Amazon Mechanical Turk (AMT) [23] to solicit opinions from human judges. AMT enables researchers to conduct user studies by providing a large pool of participants who receive payment for completing certain tasks such as participating in surveys, providing “ground-truth” data, and moderating content [23]. Researchers (requesters) create Human Interface Tasks (HITs) that interested participants (workers) can sign up for. As [24] explained, "requesters specify the amount to be paid for a HIT, how many unique workers per HIT, how much time to allot to workers, and when the HIT will no longer be available for work (expire)". In order to complete this experiment, we sought and obtained permission from our institution’s Research Ethics Office.

We realized that 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, such as crawling, capture, and replay. To avoid confusion, we chose to avoid mentioning the subject of web

archiving, and 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 [7], where the authors were trying to assess if respondents could evaluate archived websites that were "damaged" due to missing resources.

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, similar to the image in Figure 1. To answer the question, participants could choose from the following options:

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

The responses we designed for our participants were also informed by our knowledge of the way web archivists perform QA. As previous work on this subject attests, web archivists aim to preserve a website as best as possible, even if it is not perfect. Since perfect quality is impossible or impractical, the strategy for a web archivist is to settle for "good enough" quality [2]. Thus an archived website with some small quality problems might still be classified by an expert as "good enough" even if it is not perfect. Accordingly, we gave our respondents the flexibility to note two types of high-quality archived websites: those that are exactly like the original and those that are slightly different. We added the "other" response to allow for additional flexibility in case the participant thought the other options did not fit the example being shown.

In order to create more consistency in responses, we provided a website with a training manual to the AMT participants. In our training manual, we described in detail each of the options. Each option had at least one accompanying example to illustrate a proper response for the question. We aimed to keep the language and length of the training manual short and easily comprehensible to accommodate all participants regardless of language skills, educational backgrounds, and time availability.

To further improve the quality of our experiment, we implemented a number of features recommended by [24] in his overview of Mechanical Turk experiments. We required participants to have had at least a 70% lifetime approval rate from all requesters/HITs and added two attention-checking questions to ensure that participants were reading the questions. We only accepted (i.e. paid for) responses from participants who passed these attention-checks

Our final dataset in Mechanical Turk consisted of 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 judgements. Results were interpreted using the following procedure. Images classified as "perfect similarity" and "small differences" were coded as being "high quality", and those judged to be much worse than the original were coded as being "low quality". Images classified as "other" retained the same coding. Cohen's κ, a measure of inter-rater agreement for categorical scales, was run to determine if there was agreement between two participants on whether an archived website was of high quality, low quality, or had another issue [25]. There was substantial agreement between the two respondents' judgments, $\kappa = .76, p < .001$.

5. Results and Discussion

Despite the high level of inter-rater agreement, there were still 27 images where raters disagreed. In order to resolve these conflicts, we studied how inter-rater conflicts are addressed in human-subjects research. One popular approach, when faced with conflicting ratings, is to have a third, experienced rater look at the relevant data and make a final decision. The expert rater approach has been widely used across many disciplines to resolve disagreements [26], [27], [28]. Accordingly, we had a researcher examine the image and determine its final rating. Seven images labeled as "other" were also re-classified by a subject-matter expert according to their quality. In our final dataset, all image pairs were judged to be either "high quality" or "low quality." This became our "ground truth" dataset.

Since the similarity scores for SSIM and NRMSE were normally distributed, we conducted a statistical analysis using tests of significance. 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. The results on the combined dependent variable indicate that scores for high quality archived websites and low quality archived websites were statistically significantly different, $F(2, 222) = 44.95, p < .001$; Wilks' $\lambda = 0.71$; Pillai's trace $= 0.29$, partial $\eta^2 = 0.29$. We conducted follow-up univariate ANOVAs, which showed that 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, using a Bonferroni α adjusted level of .025.

In comparing both measures, we must remember than for SSIM, higher scores denote higher quality, while the reverse is true for NRMSE. Data are expressed as mean $\pm$ standard deviation. SSIM scores were higher for high quality websites (0.55 ± 0.22), than for low quality websites (0.46 ± 0.15). NRMSE scores were lower for high quality websites (0.22 ± 0.08) than for low quality websites (0.51 ± 0.36). Though the confidence intervals overlap, this is not unusual, since two means can overlap and yet be statistically significantly different from one another at the $\alpha = 0.05$ level [29], [30].

These results indicate that both SSIM and NRMSE scores are appropriate for distinguishing between high and low-quality archived websites, hence the final choice of measure can be left to the web archivist and their institutional needs. However, NRMSE scores might have a slight advantage due to the greater difference between the means of high and low quality websites, which perhaps suggests a higher discriminatory power.

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

In our research, 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 have no proper upper bound, and are thus 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.

5.2. 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 different scores for high-quality and low-quality archived websites. Thus, they are able to distinguish between high-quality and low-quality archived websites.

There are two important things to note. First, we deliberately did not supply specific threshold values for SSIM or NRMSE scores because these are highly contextual in nature and will depend on the institution that does the web archiving, the resources it can dedicate to the process, and the nature of what it is collecting. For example, an institution collecting hundreds of thousands of websites at once might choose to assign a relatively low threshold value, such as 0.6 for SSIM, because its web archivists might not be able to manually check and fix all archived sites classified as problematic. Conversely, an institution that works with a small number of archived websites might prefer a very high threshold, such as 0.85 because its collecting goal is for each archived website to be of the highest quality. It is up to the institutions involved in web archiving to set their thresholds for what would constitute a "good", "bad", or "good enough" archived website.

Second, this research focuses only on the *visual appearance* of an archived website and its visual similarity to the original website. It is not intended to focus on the quality of the user's interaction with the archived website [10], or on an archived website's completeness [7], which are other important aspects of web archive quality. However, problems with the visual appearance of an archived website can be indicative of those two issues. For example, if an archived website that is missing an image for a

video frame, such as a YouTube embedded component, it could indicate that either the video will not replay in the Wayback Machine (lack of interaction), or that the video itself has not been captured (lack of completeness).

6. Conclusions and Future Research

This paper describes how image similarity measures can be successfully applied in order to measure the visual correspondence (and thus visual quality) of archived websites, and that these similarity scores are able to distinguish between high and low-quality archived websites. 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. It is a step towards automating the QA processes for any institution involved in web archiving. 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.

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. In the future, we would live to explore further the discriminatory power of similarity metrics and establish case studies for their use.

7. Acknowledgments

The research in this paper was supported in part by funding from the Social Sciences and Humanities Research Council of Canada (SSHRC)

References

[1] Archive-It Team, The web archiving lifecycle model, 2013. URL: http://ait.blog.archive.org/files/2014/04/archiveit_life_cycle_model.pdf.

[2] B. Reyes Ayala, M. E. Phillips, L. Ko, Current Quality Assurance Practices in Web Archiving, Research Report, 2014. URL: http://digital.library.unt.edu/ark:/67531/metadc333026/.

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

[4] B. Reyes Ayala, Correspondence as the primary measure of information quality for web archives: a human-centered grounded theory study, International Journal on Digital Libraries 23 (2022) 19–31. URL: https://link.springer.com/10.1007/s00799-021-00314-x. doi:10.1007/s00799-021-00314-x.

[5] B. Reyes Ayala, E. Hitchcock, J. Sun, 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, Virginia

Tech University Libraries, 2019, pp. 12–14. URL: https://vtechworks.lib.vt.edu/bitstream/handle/10919/97987/WADL2019.pdf#page=12.
[6] Y. AlNoamany, M. C. Weigle, M. L. Nelson, Detecting Off-Topic Pages in Web Archives, volume 9316, Springer International Publishing, Cham, Switzerland, 2015, pp. 225–237.
[7] J. Brunelle, M. Kelly, H. SalahEldeen, M. C. Weigle, M. L. Nelson, Not all mementos are created equal: measuring the impact of missing resources, International Journal on Digital Libraries (2015) 1–19. doi:10.1007/s00799-015-0150-6.
[8] S. M. Jones, H. Van de Sompel, H. Shankar, M. Klein, R. Tobin, C. Grover, Scholarly context adrift: Three out of four uri references lead to changed content, PLOS ONE 11 (2016) 1–32. URL: https://doi.org/10.1371/journal.pone.0167475. doi:10.1371/journal.pone.0167475.
[9] L. Meneses, R. Furuta, F. Shipman, 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, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012, pp. 197–208.
[10] J. Kiesel, F. Kneist, M. Alshomary, B. Stein, M. Hagen, M. Potthast, Reproducible web corpora: Interactive archiving with automatic quality assessment, Journal of Data and Information Quality 10 (2018). URL: https://doi.org/10.1145/3239574. doi:10.1145/3239574.
[11] M. Klein, H. Shankar, L. Balakireva, H. Van de Sompel, The memento tracer framework: Balancing quality and scalability for web archiving, in: A. Doucet, A. Isaac, K. Golub, T. Aalberg, A. Jatowt (Eds.), Digital Libraries for Open Knowledge, Springer International Publishing, Cham, 2019, pp. 163–176.
[12] Archive-It, Learn more, 2020. URL: https://archive-it.org/learn-more.
[13] University of Alberta Library, Digital preservation services, n.d. URL: https://www.library.ualberta.ca/digital-initiatives/preservation.
[14] University of Alberta, Idle No More collection, n.d. URL: https://archive-it.org/collections/3490.
[15] University of Alberta, Fort McMurray wildfire 2016 collection, 2016. URL: https://archive-it.org/collections/7368.
[16] University of Alberta, Western Canadian Arts collection, n.d. URL: https://archive-it.org/collections/6296.
[17] Libraries and Archives of Canada, Government of canada 2016 collection, 2016. URL: https://archive-it.org/collections/7084.
[18] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, et al., Image quality assessment: from error visibility to structural similarity, IEEE transactions on image processing 13 (2004) 600–612.
[19] J. Søgaard, L. Krasula, M. Shahid, D. Temel, K. Brunnström, M. Razaak, Applicability of Existing Objective Metrics of Perceptual Quality for Adaptive Video Streaming, in: Electronic Imaging, volume 28, 2016, pp. 1–7. URL: https://library.imaging.org/ei/articles/28/13/art00010. doi:10.2352/ISSN.2470-1173.2016.13.IQSP-206, iSSN: 2470-1173 Issue: 13 Journal Abbreviation: ei.
[20] A. M. Eskicioglu, P. S. Fisher, S.-Y. Chen, Image quality measures and their performance, The 1994 Space and Earth Science Data Compression Workshop (1994) 57–67. URL: http://ntrs.nasa.gov/search.jsp?R=19940023754.
[21] Z. Tang, Y. Dai, X. Zhang, Perceptual hashing for color images using invariant moments, Applied Mathematics & Information Sciences 6 (2011) 643S–650S.
[22] Rosettacode.org, Percentage difference between images, 2018. URL: https://rosettacode.org/wiki/Percentage_difference_between_images#Python.
[23] Amazon, Amazon mechanical turk, 2018. URL: https://www.mturk.com.
[24] S. M. Jones, Building the better crowdsourced study - literature on mechanical turk, 2019. URL: https://ws-dl.blogspot.com/2019/08/2019-08-14-building-better-crowdsourced.html.
[25] Laerd Statistics, Cohen's kappa using spss statistics, 2015. URL: https://statistics.laerd.com/.
[26] J. A. Penny, R. L. Johnson, The accuracy of performance task scores after resolution of rater disagreement: A Monte Carlo study, Assessing Writing 16 (2011) 221–236. URL: https://linkinghub.elsevier.com/retrieve/pii/S1075293511000298. doi:10.1016/j.asw.2011.06.001.
[27] E. M. Voorhees, Variations in relevance judgments and the measurement of re-

trieval effectiveness, Information Processing & Management 36 (2000) 697–716. URL: https://www.sciencedirect.com/science/article/pii/S0306457300000108. doi:https://doi.org/10.1016/S0306-4573(00)00010-8.

[28] J. Belur, L. Tompson, A. Thornton, M. Simon, Interrater Reliability in Systematic Review Methodology: Exploring Variation in Coder Decision-Making, Sociological Methods & Research 50 (2021) 837–865. URL: http://journals.sagepub.com/doi/10.1177/0049124118799372. doi:10.1177/0049124118799372.

[29] S. Greenland, S. J. Senn, K. J. Rothman, J. B. Carlin, C. Poole, Goodman, S. N., D. G. Altman, Statistical tests, p values, confidence intervals, and power: a guide to misinterpretations, European journal of epidemiology 31 (2016) 337–350. doi:https://doi.org/10.1007/s10654-016-0149-3.

[30] P. C. Austin, J. E. Hux, A brief note on overlapping confidence intervals, Journal of Vascular Surgery 36 (2002) 194–195. URL: https://www.sciencedirect.com/science/article/pii/S0741521402000307. doi:https://doi.org/10.1067/mva.2002.125015.