Quality matters: A new approach for detecting quality problems in web archives

Brenda Reyes Ayala[1] Jennifer McDevitt[1] James Sun[1] Xiaohui Liu[1]

[1] University of Alberta

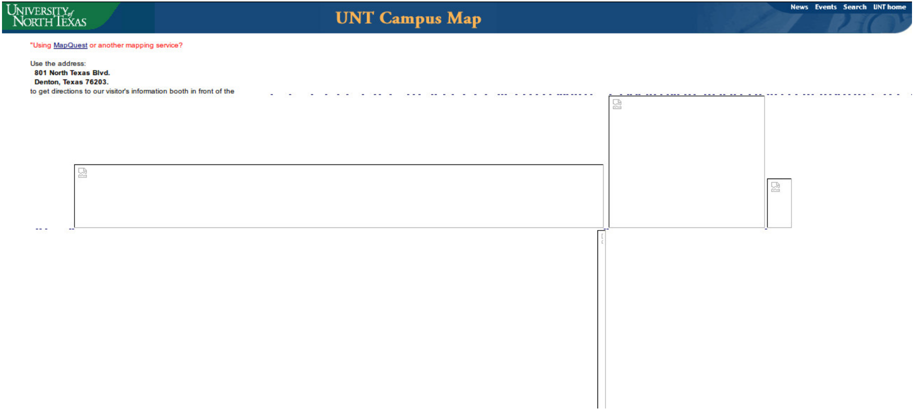
Canadian Association for Information Science (CAIS)
September 24, 2020

Contact: brenda dot reyes at ualberta dot ca

Overview

1. Introduction
2. Research Questions
3. Literature Review
4. Methodology
5. References

Why quality matters in a web archive



"Using MapQuest or another mapping service?

Use the address:
801 North Texas Blvd.
Denton, Texas 76203.
to get directions to our visitor's information booth in front of the

Figure: Screenshot of an archived version of the UNT Campus Map from 2004.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`

Context and research problem

- It is common to see archived websites with no images or media or with broken links.
- In order to detect these problems, web archivists must engage in an onerous process of quality assurance (QA) where they manually inspect hundreds or even thousands of archived websites.
- As the number of archived websites rises, quality problems grow to such a scale that human intervention is no longer enough to detect and fix them.

Research Problem

Unless we learn how to address quality problems in a web archive, we will soon be facing an incomplete digital historical record, at a time when the record is crucial.

Research questions and purpose

1. Can an image similarity metric successfully distinguish between high-quality archived websites and lower-quality archived websites?
2. How effective are different similarity measures at measuring the visual correspondence between an archived website and its live counterpart? Which measure yields the best performance?

Purpose

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.

Loss of Canadian web content

In their work McNally, Wakaruk, and Davoodi (2015) [8] examined the extensive removal of Canadian government web content and its impact on researchers, who would no longer have access to historical Canadian government web content essential for scrutinizing government policy and activities.

They stated that web archiving programs were performing a crucial role in maintaining their role as stewards of government information.

Definition of quality in web archives

According to Masanés (2006) [7], quality in a web archive is made up of the following elements:

1. the completeness of material (linked files) archived within a target perimeter
2. the ability to render the original form of the site, particularly regarding navigation and interaction with the user

Framework for data quality in web archiving

Denev, Mazeika, Spaniol, and Weikum (2011) [4] introduced the Sharp Archiving of Website Captures (SHARC) framework for data quality in web archiving, which included two measures of data quality:

1. Blur: expected number of page changes that a time-travel access to a site capture would accidentally see, instead of the ideal view of a instantaneously captured, "sharp" site. This value needed to be minimized in order to achieve a high-quality capture.
2. Coherence: number of unchanged and thus coherently captured pages in a site snapshot. Coherence needed to be maximized in order to achieve a high-quality capture.

Impact of missing elements in web archive quality

Brunelle, Kelly, SalahEldeen, Weigle, and Nelson (2015) [3] examined the importance of missing elements and their impact on the quality of archived websites.

- Embedded resources are files such as images, videos, or stylesheets, that are present and referenced in a website.
- They play a key role in ensuring the website looks and operates in the correct way.
- Crawl engineers might calculate a percentage of missing embedded resources in an archived website and use it to estimate the overall quality of the site.
- Proposed a new metric to assess the "damage" to an archived website caused by missing embedded resources. Based on three factors: the MIME type, size, and location of the embedded resource.

Core facets of IQ for web archives

Reyes Ayala [9] created a model of information quality for web archives.

1. **Correspondence**: similarity between the original and archived websites. Good correspondence requires equivalence, or at least a close resemblance, between the two
   - Visual
   - Interactional
   - Completeness
2. **Relevance**: pertinence of the contents of an archived website to the original. Archived websites must not contain off-topic content (topic relevance) or content in quantity or volume that is unexpected or excessive (size relevance)
   - Topic
   - Size
3. **Archivability**: intrinsic properties of a website that make it more difficult to archive.

Process

1. Obtain a dataset of archived websites and related data. Clean the dataset and extract relevant data.
2. Write and test code that implements metrics for quality facets of visual correspondence.
3. Conduct analyses to understand similarities and differences between the various metrics.
4. Solicit judgements of quality from human subjects.
5. Perform evaluation by analyzing which metrics most closely correspond to human judgements of IQ in a web archive.

Step 1. Obtain a dataset

The dataset used

1. “Idle No More” [10]: 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 [11]: preserves the born digital resources created by filmmakers in Western Canada.
3. British Library’s OA web archive [6]: UK websites that can be made available online according to British legal deposit laws.

Step 2. Write and test code

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).

Step 2. Write and test code

Issues encountered during the screenshot process

Not a trivial process despite our initial assumptions:

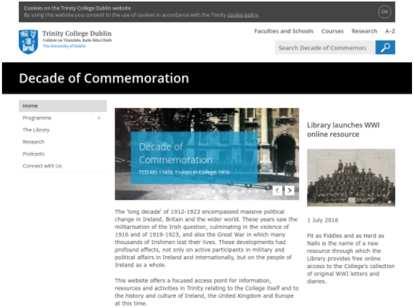
- The use of banners to indicate to users that they are viewing an archived website. Appended the text “id_” to the url of the archived websites, but this approach often breaks the CSS styling of the archived site, resulting in a screenshot that was even farther from the actual appearance of the archived 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.

Step 3. Conduct analyses to understand similarities and differences between the various metrics.

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, ∞]. 0 is perfect similarity.
- Vector distance: calculates similarity on a scale [0,1]. 1 is perfect similarity.

Example: A "medium" quality archived website



Cookies on the Trinity College Dublin website
By using this website you consent to the use of cookies in accordance with the Trinity cookie policy.

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Decade of Commemoration

Decade of Commemoration
TCD MS 11438. Troops in College. 1916

Library launches WWI online resource

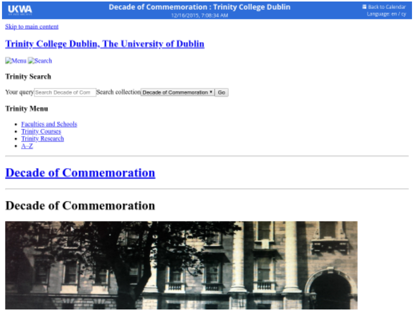
The 'long decade' of 1912-1923 encompassed massive political change in Ireland, Britain and the wider world. These years saw the militarisation of the Irish question, culminating in the violence of 1916 and of 1919-1923, and also the Great War in which many thousands of Irishmen lost their lives. These developments had profound effects, not only on active participants in military and political affairs in Ireland and internationally, but on the people of Ireland as a whole.

This website offers a focused access point for information, resources and activities in Trinity relating to the College itself and to the history and culture of Ireland, the United Kingdom and Europe at this time.

1 July 2016

Fit as Fiddles and as Hard as Nails is the name of a new resource through which the Library provides free online access to the College's collection of original WWI letters and diaries.

Screenshot of current, live website



Decade of Commemoration : Trinity College Dublin
12/16/2015, 7:08:34 AM

Skip to main content

Trinity College Dublin, The University of Dublin

Menu Search

Trinity Search

Your query Search Decade of Com Search collection Decade of Commemoration Go

Trinity Menu

- Faculties and Schools
- Trinity Courses
- Trinity Research
- A–Z

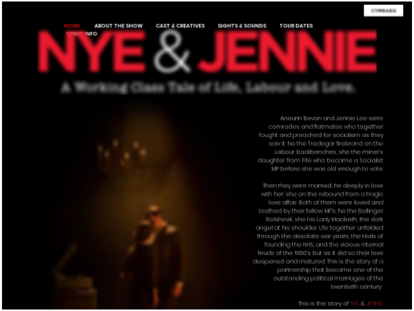
Decade of Commemoration

Decade of Commemoration

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 = 0.60

Example: A "low" quality archived website



NYE & JENNIE

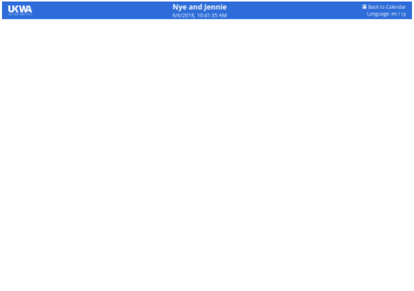
A Working Class Tale of Life, Labour and Love.

Aneurin Bevan and Jennie Lee were comrades and flatmates who together fought and preached for socialism as they saw it; he the Tredegar firebrand on the Labour backbenches, she the miner's daughter from Fife who became a Socialist MP before she was old enough to vote.

Then they were married; he deeply in love with her; she on the rebound from a tragic love affair. Both of them were loved and loathed by their fellow MP's; he the Bollinger Bolshevik; she his Lady Macbeth, the dark angel at his shoulder. Life together unfolded through the desolate war years, the trials of founding the NHS, and the vicious internal feuds of the 1950's. But as it did so their love deepened and matured. This is the story of a partnership that became one of the outstanding political marriages of the twentieth century.

This is the story of NYE & JENNIE.

Screenshot of current, live website



Nye and Jennie

6/6/2018, 10:41:35 AM

Screenshot of archived website

Figure: Comparison of images for the website of the play "Nye & Jennie". SSIM = 0.28, MSE = 169603.88, Vector Distance = .08

Step 4. Solicit judgements of quality from human subjects.

Rubric for judging the visual correspondence of the archived screenshots

Quality Judgement	Description
High Quality	Intellectual content, images, and styling are preserved in both screenshots; the screenshots look almost identical. Images can be missing if not integral to webpage content (i.e. ads)
Medium Quality	Intellectual content is preserved without styling and lack of style elements does not impede readability.
Low Quality	Little or no content is preserved.
No Comparison	Script failed to take a screenshot, server connection failed, technical issues occurred, etc. OR link rot has occurred (link not connecting for whatever reason. For example, blog has moved, page no longer exists, etc.), or page has changed substantially so it no longer resembles archived screenshot (website redesign, etc.)

Step 4. Solicit judgements of quality from human subjects

We enlisted the help of two University of Alberta librarians with previous web archiving experience and two student RAs. Quality judges met several times to discuss the process and rubric involved, and one of the researchers was responsible for checking the judgements for accuracy.

Step 5. Perform evaluation to see which metrics most closely correspond to human judgements

1. MSE measure proved to be the most difficult to interpret, as it has no proper upper bound. For this reason, we discarded MSE from our image similarity metrics.
2. Qualitative inspection of the scores also revealed that the vector distance would sometimes yield inaccurate results, as some screenshots that were completely blank had scores in the 0.50-0.60 range.
3. For these reasons, we decided on SSIM as the best image similarity measure for the task.

Step 5. Perform evaluation to see which metrics most closely correspond to human judgements

Preparing the data for statistical testing

We removed the screenshots labeled "No Comparison" from the dataset and examined the remaining similarity scores. We found that neither SSIM, nor vector similarity were able to fully distinguish between low and medium-quality archived websites, therefore we merged these two categories into a single category labeled "Low/Medium Quality."

Step 5. Perform evaluation to see which metrics most closely correspond to human judgements

Statistical testing

1. Created a balanced random sample of 100 screenshots, with 50 images deemed to be "High Quality" and 50 images that were "Low/Medium Quality."
2. Ran a Mann-Whitney U test (chosen because the distribution was not a normal one) to determine if there were differences in similarity scores between high-quality archived websites and low-quality archived websites.
3. Median SSIM scores were statistically significantly higher for high-quality websites (0.95) than for websites judged to be medium or low-quality (0.65), $U = 279, z = -6.69, p < .001$. Therefore we conclude that the SSIM scores for high-quality archived websites are statistically significantly higher than those of low-quality websites.

Conclusions and findings

Our results indicated that

- Similarity metrics are able to successfully distinguish between website captures of poor quality and those of higher quality.
- Structural Similarity Index metric (SSIM) was most able to successfully measure visual correspondence between an archived website and its live counterpart

Next steps

This is only the first step in developing a comprehensive toolkit for automated or semi-automated quality assurance processes in web archives, which will in turn help web archivists create better web archives in the future.

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

[3] 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

[4] Denev, D., Mazeika, A., Spaniol, M., & Weikum, G. (2011, March). The SHARC framework for data quality in web archiving. *The VLDB Journal*, 20 (2), 183-207. doi:10.1007/s00778-011-0219-9

[5] 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

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

[7] Masanés, J. (2006). Web archiving. Berlin; New York: Springer.

[8] McNally, M.B, Wakaruk, A., & Davoodi, D. (2015). Rotten by design: Shortened expiry dates for government of Canada web content. *Proceedings of the Annual Conference of CAIS, Canada*. doi: https://doi.org/10.29173/cais909

[9] 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/`

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

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