Quality Preservation

Emerging Quality Assurance Practices in the Library of Congress Web Archives

Meghan Lyon
Library of Congress
USA
mlyon@loc.gov

Grace Bicho
Library of Congress
USA
grth@loc.gov

Abstract – Building sustainable quality assurance practices is a challenge for today's preservationists, who want to be sure that content preserved in web archives is not only the correct content, but in working order. This often means that archived web content should be replayed via Wayback rendering software in good fidelity when compared to the original website. The exponentially growing scale of web archives necessitates a multipronged approach to identify what is (and is not) being preserved, and where improvements can be made. This paper will explore actions that can take place iteratively throughout the web archiving life cycle, as part of a larger system of review where multiple individuals can contribute, including non-technical Library staff and subject matter experts. The processes described are part of a novel workflow in the Library of Congress Web Archiving Program.

Keywords – Web Archives, Quality Assurance, Workflows, Human-centered digital practitioners

Conference Topics – From Theory to Practice, Sustainability, We're All in this Together

I. Introduction

The Library of Congress Web Archiving Program manages an ever-growing archive of over 3.5 Petabytes (PB) of content archived from the web since 2000. The archive comprises over 180 event and thematic collections, nearly 31,000 cataloged web archives, and approximately 15,000 seed URLs ("websites") actively crawling at any given time. The Library's technical Web Archiving Team (WAT) is responsible for managing the program from start to finish, which includes leading the assessment of archive quality, even though the WAT does not select content for the archive.

Assessing the quality of web archives is a notoriously difficult endeavor for the web archiving community, given the sheer chaos of file formats present in the archive, the quickly increasing scale, and persistent replay issues with the current suite of access tools, which will always lag behind new technologies used to build the live web. However, it is seen as due diligence by the WAT to confirm capture of selected content for the Library's collection. WAT also approaches quality assessment as an act of sustainability, within the feedback loop of the Library's ongoing captures, in order to scope capture to *only* content that has been selected for the collection, according to the Library of Congress Collection Policy Statements [1]. Finally, performing quality assessment allows the WAT to provide a reasonable expectation of the usability of the archive for those building and using the collection [2].

This paper presents a detailed explanation of the Library of Congress Web Archiving Team's practical approach to quality assessment of the web archive, including computer-mediated methods, according to Dr. Brenda Reyes Ayala's theoretical framework for performing quality assessment on archived web content [3].

II. Theoretical Framework

The "human-centered grounded theory" [3] is the first of its kind to provide a theoretical framework for increasing web archivists' confidence

iPRES 2023: The 19th International Conference on Digital Preservation, Champaign-Urbana, IL, US.

https://hdl.handle.net/2142/121122

in quality assurance (QA) methods in the face of the enormous scale of managing web archives. The grounded theory includes three dimensions used to assess quality of the web archive: Archivability, Relevance, and Correspondence.

A. Theoretical Definitions

1) *Archivability:* "the degree to which the intrinsic properties of a website make it easier or more difficult to archive."
2) *Relevance:* "the pertinence of the contents of an archived website to the original website. Reference [3] defines two measures of relevance: topic relevance and size relevance."
3) *Correspondence:* "the degree of similarity, or resemblance, between the original website and the archived website." Reference [3] defines three measures of correspondence: *visual correspondence, interactional correspondence, and completeness.*

III. Archivability

Archivability is the most difficult dimension to assess completely as website-building frameworks are constantly changing, and web archiving technology is slow to adapt. The WAT works with its vendor, who performs the data capture component (known as web "harvesting" or "crawling") of the web archiving life cycle [4], to begin assessing archivability. The WAT also takes on the responsibility of communicating archivability to nominators—non-technical Library staff responsible for selecting content for the archive—in order to manage expectations of what is possible to archive.

A. Vendor Collaboration

The Library's crawl vendor works continuously to improve the captures of selected content and to determine which web development technologies make crawling difficult. Before a harvest begins, the vendor first uses a technology, such as Wappalyzer [5], to scan a website for frameworks, programming languages, web servers, and anything else that may impede capture. Based on the results, the vendor can decide which crawling technology is best suited to harvest each website. Once a crawl finishes, the WAT can provide feedback about how well the technology worked, and can suggest movement among various crawl technologies. This collaborative feedback loop is critical in identifying challenges with archivability.

B. Known Challenges

Over time, working with the vendor and assessing crawls, the WAT has built up a list of common challenges with certain platforms or websites. In order to manage expectations of crawling and archive replay for nominators, the team provides a table of guidelines, on an internal Wiki, called "Web Archiving Known Challenges." Nominators are then able to consult the list at any time, particularly during initial content selection or while assessing crawl quality of their selected content.

IV. Relevance

According to [3], the core category of relevance is split into two dimensions: topic and size relevance. Topic relevance measures the closeness of a web archive to the original, live website or part of a website. This curatorial measurement is largely outside of the scope of practice for the WAT. The second dimension of size relevance, or how closely a web archive's size correlates to the live website, is within scope for WAT, the technical team tasked with assessing quality of incoming web harvests.

Since it is difficult to determine the size of any given website, it is also difficult to determine whether the size of the archived version matches the live website. Some web archiving programs run test crawls to determine archivability and accuracy of crawl instructions, and are able to determine approximate website size at that point. However, the Library only crawls at ongoing, regular intervals, providing the ability to compare the size of archived versions over time, as well as identify websites that appear unreasonably small or unreasonably large, given the number and types of resources it takes to make up a website.

Using reports generated by the crawler software and crawl vendor, the WAT devised a method for assessing the relative size of each seed (or website URL at which the crawl is set to begin harvesting). The reports utilized are: the Heritrix crawler standard seeds-report.txt report [6], including the response codes and HTTP status of each seed at the time of harvest, and a bespoke report of the number of hops traversed (or depth) and number of raw bytes collected (or bytes) per seed by the end of each crawl.

The above data points are collated by the WAT into a spreadsheet and are matched with collection data from the program's curatorial database per seed URL. From there, the WAT can easily sort by the response codes, depth, and bytes, or by a particular collection or crawl frequency. Various sorting highlights initially the websites with extremely low bytes and depth that had obvious crawl issues. From there, the WAT staff performing QA can triage the investigation of seeds with low- to mid-range bytes and depth as an indication of difficulty crawling some or all parts of the seed. Resolutions of these investigations can look like switching the crawl technology for a particular seed, updating crawl instructions (or "scopes") for the web crawler, or removing the seed from crawl altogether.

In this way, the WAT leans into the iterative flow of the Library's unique crawling ecosystem, using relative size of the seeds in a crawl and over time to highlight acute seed issues.

V. Correspondence

The WAT is responsible for overseeing the capture of approximately 15,000 seeds at any given time. Regarding the assessment of quality for those seeds, archivability and size relevance help immensely to highlight seed issue needles in the archive haystack. To look deeper into the quality of each site at scale, subject expertise and the measures of correspondence come into play, a process which the WAT calls "capture assessment."

A. Capture Assessment: Data Collection

For the Library, all three correspondence categories: visual correspondence, interactional correspondence, and completeness, rely on the nominator's knowledge of the live website for comparison. To gather actionable information about quality from nominators and other staff supporting review of the content–referred to as "reviewers" in the context of performing capture assessment–WAT has translated the three categories into a rubric to be measured. For each category, a numeric range is instituted from 1 (worst) to 5 (perfect), which the reviewer can use to ascribe a numeric value for that category for a single capture of a seed.

The visual correspondence score can range from appearing "unrecognizable" (1) to appearing "perfect" (5). The WAT's prompt elaborates, "similarity in appearance between the original website and the archived website" [3] by asking reviewers: *If you were to look at the archived page and the page on the live web side by side, how similar would they look?*

Similarly, the interactional correspondence category includes the definition, "the degree to which a user's interaction with the archived site is similar to that of the original" [3], alongside a series of questions meant to flesh out the concept, such as: *Do the navigation buttons function? Is there an endless scrolling feature or interactive visualization?,* and *Does it work in the archive?* The interactional correspondence score can range from inability "to interact with any features of the archived website" (1) to ability "to interact with all features of the archived website" (5).

Completeness, "the degree to which an archived website contains all of the components of the original", asks reviewers to *get a holistic sense of the archive. What overall patterns emerge as you navigate around the archived site?* We ask reviewers to rank the whole capture to say "no content missing" (5), "some content missing" (4), "half content missing" (3), "most content missing" (2), and "all content missing." (1)

If the rating of any category is any less than 5, the WAT provides a checklist of common issues that communicate the issue they are seeing with that capture, including a free-text "other" box for any unlisted issues. Some of the common issues in the checklist include: *Missing images*, *Missing documents*, *Missing style*, *Paywall or login impedes use*, *Page elements disappear*, and *Issues with interactive content*.

An introduction to the work of Reference [3], a rubric for correspondence scores, and a Specific Issue checklist is presented to the reviewer within a Confluence form. When the form is submitted, WAT gets an email with the results and can act on identified issues. However, in order to streamline review of capture assessments, WAT exports the form results at regular intervals, integrating work reviewing the capture assessments with bi-weekly work-planning sessions within the team's Scrum workflow [7].

B. Capture Assessment: Action steps

Individual tickets are created, per capture assessment form response, in a workflow organizer (Jira) and assigned at random to WAT staff. Before importing into Jira, the form response data undergoes a transformation via Python script. This step fulfills the dual purpose of: 1) formatting the form responses into an order suitable for bulk-import to Jira tickets and 2) averages the 1-5 correspondence ratings. The average of the three correspondence ratings dictates the priority level of the Jira ticket:

1) *Blocker:* a score of 1 in any category
2) *Critical:* average correspondence score less than or equal to 2
3) *High:* average correspondence score greater than 2 and less than or equal to 3.5
4) *Medium:* average correspondence score greater than 3.5 and less than 5
5) *Minor:* average correspondence score of 5, exactly, indicating a perfect capture

Prioritization of quality assurance is critical in web archives, which have endless opportunities for improvement, but real human limits. Assigning Blocker to a given capture assessment ticket indicates to the WAT that a crawled seed requires attention immediately. A Medium score, on the other hand, is indicative of something wrong, which can often be righted with a small adjustment by WAT, such as updating the crawl instructions.

C. Early Results

Six months into the effort to put theory into practice, WAT is beginning to see preliminary results. Over 193 captures of seed URLs have been assessed by 15 unique reviewers across 13 collections (some collections had multiple reviewers and some unique reviewers assessed captures from more than one collection). An average correspondence score of 3.86 has emerged. By priority, roughly 30% of tickets land in the Blocker, Critical or High priorities with the remaining 70% at the Medium and Minor levels. During February 2023, the WAT averaged 7.5 days to complete processing of new capture assessments.

The majority of assessments (54%) were performed on content collected as part of a multi-disciplinary, cross-divisional collecting effort geared toward collecting publications via web archiving. This collection is unique to the Web Archiving Program in that it has acquisitions staff assigned to the collection who act as liaisons between staff with recommending authorities and the WAT. In keeping with the collection's focus, the most widespread specific issue discovered in this collection is *Missing documents* (38% of all reported issues for the collection), followed by *Missing content* (other) and *Missing links* (11% each).

Of the 310 specific issues reported across the assessed collections, the highest counts of specific issues checked were *Missing images* (21%), *Missing documents* (19%), and *Missing style* (13%), which is a common formatting error where CSS is either not captured or improperly rendered.

It is helpful for reviewers to indicate when they see something "missing" that they expect to be present in the archived capture. Reviewers with language and subject expertise highlight areas of the site most critical to collect. When these specifics are pointed out, WAT can investigate further to verify whether something is truly missing from the archive versus un-navigable from a given starting point, thereby ensuring capture of content selected for the Library's collection.

Investigation often begins by consulting the live site for the URL in question, or a representative URL of the larger issue, i.e., an image URL if *Missing images* was checked. With a URL in hand, WAT can pinpoint examination of the resource via Wayback replay or the archive indexes to better understand whether the URL is truly absent in the archive. WAT can then compare the document URL path with existing scopes in the Library curatorial workflow tool. At this point it becomes possible to detect whether the issue is a crawl directive error or something more problematic in respect to the composition of the live site and rendering behaviors in use. The crawl vendor can be consulted to investigate the crawl logs to confirm a point of failure.

Results of capture assessment processing and subsequent investigative work are relayed back to the reviewers via email and are also included in comments within the Library's curatorial tool. These comments allow future stewards of the permanent collections to take stock of capture quality at a given time and collate known quality issues of a given seed.

VI. Conclusions

After implementing practical methods to satisfy each component of the grounded theory for web archives QA, the WAT has found that each practice provides a unique view into the quality of the web archive, with little overlap. After the first six months, it appears that staff performing capture assessment are reviewing captures not normally highlighted during the semi-automated size relevance assessments performed by WAT. This indicates the importance of maintaining an ecosystem of quantitative and qualitative methods to assess quality, particularly as the collection continues to grow.

The emerging average correspondence score of 3.86 is a positive take away for the WAT. Results of web archiving at-scale can never be perfect, and this score indicates to us that captures are generally good. Correspondence ratings broken down by category are also positive indicators: 69% of captures scored a 4 or 5 on Completeness, about 64% scored 4 or 5 on Visual Correspondence, and 72% received a rating of 4 or 5 in Interactional Correspondence; only about 7% scored a 1 (lowest score) in any of the 3 correspondence categories. An anecdotal, positive takeaway of capture assessment is the WAT's ability to act in many cases to resolve or clarify "missing" elements.

VII. Ongoing Work

As the Library continues to work closely with its crawl vendor on QA, and particularly issues relating to archivability, the WAT is exploring other areas for improvement in the capture assessment and QA processes. There are some technical hurdles related to available tools for the workflow. WAT's first question in the capture assessment form, "is this the right website?" is meant to address the issue of link drift. If a capture is not intellectually consistent with the entity targeted for harvest, often this means that there is content drift on the live web. When checked "no", the form is supposed to end, however it defaults to all 5's (minor priority) and has affected 4 assessments out of the 193, at this point. This can be resolved by making the default ratings all "1" however this creates extra work for reviewers rating perfect captures, as they will have to manually click "5", "5", "5". Not having a default selection is not an option in the available tool.

Plans are underway to include employing technicians in the Library's Digital Content Management Section to complete capture assessments. As nominators have a small percentage of time for their web archiving duties, the technicians will be able to review a larger swath of the archive in a shorter time period. This practice will remove subject expertise, to some degree, but as they complete capture assessments, the technicians will gain familiarity with the collections. Data dashboards are also currently in development that can merge and visualize capture assessment results and technical crawl data (bytes, hops, etc.) for seed URLs and collections over time.

The Library's Web Archiving Program exists in a state of continual improvement, and the team will streamline features of the described workflows, as possible. Parts of the size relevance assessment workflow are scheduled to be automated further, such as generating the crawl report spreadsheet via continuous integration pipeline, thereby allowing WAT staff to press a button versus running a command line Python script. Against the scale of the archive, these small components of workflow preparation add up and the WAT will continue to leverage automation as much as possible.

1. REFERENCES

[1] Library of Congress Collection Policy Statements, https://www.loc.gov/acq/devpol/

[2] Bicho, G., Lyon, M. (May 24, 2022) Building a Sustainable Quality Assurance Lifecycle at the Library of Congress, presentation at the International Internet Preservation Consortium (IIPC) General Assembly and Web Archiving Conference. https://digital.library.unt.edu/ark:/67531/metadc1983138/

[3] Reyes Ayala, B. Correspondence as the primary measure of information quality for web archives: a human-centered grounded theory study. Int J Digit Libr 23, 19–31 (2022). https://doi.org/10.1007/s00799-021-00314-x

[4] Bragg, M., & Hanna, K. (2013). The web archiving life cycle model. https://ait.blog.archive.org/files/2014/04/archiveit_life_cycle_model.pdf

[5] Wappalyzer, https://www.wappalyzer.com/

[6] Seeds (seeds-report.txt), Heritrix, https://heritrix.readthedocs.io/en/latest/operating.html?highlight=seeds-report.txt#seeds-seeds-report-txt

[7] Bicho, G. (2021). The Library of Congress Web Archiving Team Goes Agile. The Signal: Digital Happenings at the Library of Congress. https://blogs.loc.gov/thesignal/2021/01/wat-goes-agile/