IFLA Section
News Media

IFLA International News Media Conference 2024

High Fidelity Web Archiving of News Sites and New Media with Browsertrix

Tessa Walsh
Webrecorder, Montreal, Canada.
E-mail address: tessa@webrecorder.org

Henry Wilkinson
Webrecorder, Toronto, Canada.
E-mail address: henry@webrecorder.org

Ilya Kreymer
Webrecorder, San Francisco, United States of America.
E-mail address: ilya@webrecorder.org

Abstract:

This paper discusses how Webrecorder's free and open source browser-based web archiving tools such as Browsertrix can and have been used by libraries and archives to create and provide access to high fidelity web archives of online news sites, social media, digital publications, digital humanities projects, and other historically difficult to preserve forms of online news media. Emphasis is placed on recently developed assistive quality assurance (QA) tools implemented in Browsertrix that allow users to assess the quality of captured content with the assistance of automatically calculated metrics such as screenshot and text comparison between the site as visited by a browser during crawling and its replay from the captured archive. This exciting new development builds on existing features which differentiate Webrecorder's browser-based crawling from alternative web archiving methods, such as the use of browser profiles to archive material behind log-ins and on personalized social media feeds, ad and cookie blocking features, and a suite of extendable behaviors that drive the browser during capture, allowing for autoscroll as well as automated navigation of certain social media sites. The paper discusses how these features enable librarians to easily and effectively preserve and provide access to news media, referencing several recent collaborations between Webrecorder, libraries, journalists, and others invested in high fidelity archiving of important and often complex online content.

Keywords: web archiving, quality assurance, webrecorder, browsertrix, digital preservation

Introduction

Webrecorder's free and open source browser-based web archiving tools Browsertrix and ArchiveWeb.page enable libraries, archives, journalists, and other interested parties to create and provide access to high-fidelity web archives of paywalled news sites, algorithmically-generated feeds, interactive web applications, and other traditionally difficult to preserve forms of new(s) media. This paper discusses how Webrecorder tools such as Browsertrix differ from other web archiving technologies and how they can and have been used by those interested in preserving cultural heritage to preserve news media on the web, regardless of its format. Core to this ability are features such as browser profiles, which enable users to save a browser's state, including cookies that store login sessions to websites, for use during crawling; ad and cookie blocking features; and a set of extendable behaviors that drive the browser to auto-scroll and navigate some social media sites automatically. Recently developed quality assurance (QA) features in Browsertrix push the state of the art in web archiving further, by enabling users to assess the quality of web crawls with the assistance of automatically calculated metrics such as screenshot and text comparison between the site as visited by a browser during crawling and its replay from the captured archive. Taken together, these features make it easier than ever for librarians and archivists to focus their attention on core practices such as selection, description, making informed preservation decisions, and providing access, and reduce the time spent worrying that important cultural heritage online will be difficult or impossible to preserve.

Background

To fully understand how Browsertrix and browser-based web archiving differ from and improve upon traditional web archiving tools, it is important to ensure that readers have sufficient background in the state and historical development of web archiving broadly. This background aims to give sufficient context for the remainder of the paper rather than being a comprehensive history of the field.

Web archiving

The widest known and perhaps earliest wide-scale institutional effort to preserve pages on the web is the Internet Archive. Founded in 1996 by Brewster Kahle alongside Alexa Internet, the Internet Archive preserves culturally important web pages, which have been available publicly via the Wayback Machine since 2001 (Webster, 2017). As of the writing of this page, the Internet Archive's Wayback Machine includes over 866 billion web pages, which are searchable by URL and keyword, with individual snapshots of captured web pages navigable via a calendar-based interface (Internet Archive, 2024).

Web archiving programs were also developed starting in the 1990s by national libraries and archives predominantly interested in capturing their own national web domains, often reflecting an evolution of legal deposit requirements being expanded to include online resources. Early

projects to begin preserving important websites were launched at Library and Archives Canada in 1994-1995 via the Electronic Publications Pilot Project, the National Library of Australia via the PANDORA project in 1996, and the Royal Library of Sweden via the Kulturarw3 project in 1996, among others. Legal frameworks developed alongside these projects, with Denmark among other nations expanding legal deposit laws to include non-print material as early as 1997 (ibid). In the United States, the Library of Congress began collecting culturally important web content, with selection driven by curation rather than legal deposit frameworks, through the MINERVA (Mapping the INternet Electronic Resources Virtual Archive), which would eventually develop into the Library of Congress Web Archive (Grotke, 2011). Web archiving programs and the legal frameworks that support and in some cases mandate them have continued to evolve in the decades since, resulting in a number of large scale national web archives with holdings in the petabytes.

The International Internet Preservation Coalition (IIPC) was established in 2003 by an agreement at the National Library of France, originally with twelve participating institutions . Since its founding, IIPC members have coordinated efforts to "fund and participate in projects and working groups to accomplish the goals of the IIPC," in order to affect change and push the field forward at a faster pace than each individual would be able to accomplish alone. From its founding in 2003, the IIPC has expanded to more than 50 member institutions from 35 countries as of 2024, and has made significant strides in identifying and establishing best practices and standards for web archiving, funding development for web archiving tools, advocating for legal frameworks friendly to the preservation of important online content, encouraging and facilitating use of web archives by researchers, and putting on annual conferences to facilitate conversation amongst members and the web archiving profession more broadly. The membership of the IIPC has also diversified as it has grown, including not only the Internet Archive and national libraries but also university web archiving programs and web archiving tool developers (IIPC, 2024).

Early web archiving primarily involved the use of crawlers such as the Internet Archive's Heritrix, which navigate websites and domains similarly to how search engine crawlers find and index new pages, archiving the pages in bulk based on the content returned by the website's servers. Such crawlers, Heritrix included, are still widely used, particularly for large crawls such as those of an entire national top-level web domain, but require technical expertise and infrastructure that put web archiving out of the reach of many smaller institutions and individuals. Sufficiently motivated individuals often had to rely on smaller-scale crawling tools instead, such as HTTrack, WGET, or cURL. Such tools can effectively capture many websites with varying degrees of support for standardized web archiving formats like WARC, but often struggle with database-backed web applications and JavaScript-heavy Single Page Applications (SPAs) that have become increasingly common on the web, since they do not rely on an actual web browser and therefore may not execute the scripts required to download all page resources.

In recent years, the availability of hosted services made it possible for smaller institutions, ad hoc groups, and even individuals to participate in archiving the web. Hosted services such as the Internet Archive's Archive-It and Rhizome's Conifer (originally webrecorder.io) made it possible for such groups and individuals to have access to advanced web archiving technologies without having to develop in-house technical expertise. Perma.cc, a service run by the Library Innovation Lab at Harvard Law School, similarly "helps scholars, journals, courts, and others create permanent records of the web sources they cite," addressing the persistent issue of link rot in

scholarly and legal citations (Perma.cc). Conifer and Perma.cc both use free and open source software tools developed and maintained by Webrecorder as part of their services. These tools resulted in better quality archives as they employed actual web browsers as part of the archiving process for the first time.

Webrecorder's free and open source web archiving tools have made web archiving easier and more accessible to a wide audience, while also introducing the possibility of capturing complex and interactive web content that was previously difficult or even impossible to archive. This high-fidelity archiving is made possible by driving and recording data loaded by web browsers, alongside complex rewriting of archived content as it is viewed on replay to ensure that JavaScript-heavy modern websites work as expected and links point to resources within the web archive rather than the live web.

The most easily accessible of these tools is ArchiveWeb.page, a Chrome browser extension which allows users to archive the contents of web pages in their browser as they navigate the web. HTML pages, media such as images and video, and a site's other resources such as JavaScript scripts and CSS stylesheets are all captured, allowing for later replay of the visited pages in any order that a user desires from the resulting archive in Webrecorder's client-side ReplayWeb.page web archive viewer. The resulting archives can be downloaded in the ISO-standard WARC file format or Web Archive Collection Zipped (WACZ), a format developed by Webrecorder which packages WARC files among indices, page lists, and other accompanying files in a portable zipped format that allows for faster loading of archived resources (Webrecorder, 2021). These files can then be replayed off of local storage devices directly in a user's browser with ReplayWeb.page, without needing to upload data to any server.

Webrecorder's Browsertrix Crawler enables browser-based high-fidelity crawling of websites at a larger scale than is possible in an individual's browser session. Browsertrix Crawler is a simplified browser-based high-fidelity crawling system, designed to run a complex, customizable browser-based crawl in a single Docker container. This crawler is integrated in the cloud-native Browsertrix application, which provides a user-friendly interface and additional features for web archiving, and is described in detail below.

Quality assurance

Archiving websites at large scales introduces a number of challenges to web archiving practitioners. Identifying whether pages and their many resources have been captured correctly is a challenging endeavor when crawls comprise many thousands of pages or more.

Quality assurance (QA) workflows and tools have long been recognized by practitioners as an important and underdeveloped area of practice. Brenda Reyes Ayala, Mark E. Phillips, and Lauren Ko of the Web Archiving Team at the University of North Texas conducted a survey of 54 institutions engaged in web archiving in 2014 to understand the current state of QA practices in the field, building on a similar but smaller survey undertaken by the National Library of the Netherlands (KB) in 2010 (Reyes, 2014). Respondents worked in national libraries, universities, museums, art libraries, and some smaller organizations, and used both in-house technologies as well as hosted services such as Archive-It and the since-discontinued California Digital Library's

Web Archiving Service. Just over 90% of the respondents indicated that they had some sort of QA process for archived websites, with around 64% of respondents describing this as an entirely manual process requiring close human review. Those who indicated they used a combination of manual and more automated processes largely relied on inspecting crawl logs and crawl reports such as summaries of response codes and MIME types as generated by Archive-It. The most popular way of checking the quality of a particular web page after crawling was manually inspecting the page in Internet Archive's Wayback Machine. The authors conclude that "manual QA is labor-intensive, complicated, and requires staff to receive special training," largely due to a lack of tools to automate the QA process and to make manual QA processes more efficient (ibid). The authors conclude with a case study of an automated QA process co-developed by the Library of Congress and the Internet Archive, which they note was outside of the reach of less well-resourced organizations and better at detecting crawl issues than replay problems.

In subsequent years, presentations on web archiving QA have been commonplace at conferences. Nicholas Taylor, then Web Archiving Service Manager at Stanford University Libraries, presented on rethinking web archiving QA processes for sustainability at the Society of American Archivist's 2016 annual conference. In the presentation slides, he notes that the 2015 National Digital Stewardship Association (NDSA) Web Archiving Survey found that quality assurance was a highly desired skill with low perceived programmatic progress at a majority of web archiving institutions, and describes efforts at Stanford to refine local QA processes to maximize their impact and efficiency. Presentations at the 2023 IIPC Web Archiving Conference by web archiving practitioners from Library of Congress and National Archives (UK), focused on the development of institutional QA workflows and attempts at automating parts of the QA process, respectively (Bicho, et al, 2023; Feissali & Bickford, 2023). A similar conversation was facilitated by the Digital Preservation Coalition that same year in their online event "Web Archiving: How to Achieve Effective Quality Assurance." Taken together, these presentations and discussions demonstrate a great deal of interest and perceived need for tools to assist with performing QA on web archives in more efficient ways than have been previously available to practitioners.

Browsertrix

Webrecorder's most recent project, Browsertrix (formerly Browsertrix Cloud), is a cloud-native application which wraps Browsertrix Crawler, ReplayWeb.page, and other tools into a single, intuitive web archiving platform. As free and open source software, Browsertrix is run as a hosted service available to individuals and organizations by Webrecorder but also available for any interested individual or institution to install and run on their own hardware. Browsertrix is in active development, supported in part by funding from the Filecoin Foundation and the IIPC, with a roadmap of planned features that will continue to improve institutions' and individuals' abilities to archive important content on the web regardless of its format or the underlying technologies used on a particular website.

Within an organization in Browsertrix, multiple users are able to collaboratively create seeded crawl workflows that start from a root ("seed") page and then continue through a website through link extraction or parsing of a sitemap, as well as URL List crawl workflows that target a specific list of URLs. These crawl workflows can be configured with a number of settings, including scope

types for seeded crawls, page limits, timeouts and delays, whether or not the crawler should "hop out" to external links, and exclusion settings that use regular expressions to include or exclude URLs from the crawl. Crawls can be started manually from their workflows at any time as well as scheduled to run at regular intervals such as weekly or monthly.

While a crawl workflow is running, users get a live view of Browsertrix Crawler's windows through screencasts, and have the ability to inspect and manage the queue of URLs to be crawled by editing the workflow's exclusion settings in real time. The visibility and control provided into the crawling process helps users understand exactly what the crawler is doing, adjust the scope of their crawl as it runs by adding exclusions, and avoid crawler traps such as pages that link to an infinite number of other pages with identical content but different URLs.

Once crawling is complete, the archived items produced by the workflow can be replayed directly in the Browsertrix interface, and WACZ files can be easily downloaded from the UI. WACZ files created by Browsertrix are digitally signed on creation in accordance with the WACZ Signing and Verification specification, to aid in establishing their provenance and authenticity.

Crawl workflows can also take advantage of browser profiles, saved browser user profiles containing browser settings and cookies such as those used to store user authentication sessions on websites. New browser profiles can be created in an interactive browser embedded within the Browsertrix user interface and then saved and applied to crawl workflows. This feature is most commonly used to crawl content hidden behind logins, accept cookie popups, and capture algorithmically-generated content such as social media feeds that are unique to a particular user or even a particular browser session. Browser profiles can also be used to configure and save browser settings such as ad blocking, cookie blocking, and other privacy settings in Brave Browser, the browser used by Browsertrix Crawler. The ability to edit browser profiles in the user interface allows users to ensure that logins have not expired prior to running a crawl with a browser profile, as well as to make any adjustments that may be necessary.

WACZ files created in other applications such as ArchiveWeb.page can also be uploaded into Browsertrix through the user interface or through an integration within the ArchiveWeb.page browser extension. Uploaded archived items can then be replayed in and downloaded from Browsertrix like crawls created from within the application, and are backed up to any configured replica storage locations.

Curation and sharing features in Browsertrix are facilitated through collections, which allow users to combine any number of crawls and uploaded archived items so that they are capable of being replayed together and even downloaded as a single WACZ file. Users can make collections sharable, which allows anyone on the web with the sharing link to view and interact with the collection. The sharing dialog also provides embedding snippets allowing website owners to easily embed the collection into their sites using ReplayWeb.page. Additional development work is planned for the future to enhance discoverability and presentation for publicly shared collections, making it easier than ever for organizations and individuals to provide access to their archived items.

Developers and technically advanced users also have the option to use Browsertrix without interfacing with the front end user interface (UI) at all through use of the same REST application programming interface (API) that the Browsertrix UI uses to interact with the backend of the application. Webhooks can be configured through the REST API to notify external applications when specific events occur, such as a crawl starting or stopping. This gives other applications the ability to submit a request to crawl a particular website or page from Browsertrix without requiring users to manually create the crawl in the Browsertrix UI, as well as receive notifications with the link to download the resulting WACZ files when the crawl completes. Documentation for the REST API is automatically generated and available within the application at the "/api/redoc" URL.

Archiving new(s) media sites with Webrecorder tools

Browsertrix and other Webrecorder tools such as ReplayWeb.page, ArchiveWeb.page, and Browsertrix Crawler, have been used by a number of groups to archive and replay online news sites, social media, digital publications, digital humanities projects, and other historically difficult to preserve forms of online media. By taking advantage of the affordances of Webrecorder tools, our collaborators have been able to manage content that they previously had difficulties crawling, embed archived social media posts in news articles with an archival receipt that demonstrates the provenance of the post, ensure that digital publications are archivable, and continue providing access to digital news projects and digital humanities projects long after their active development has concluded.

In February 2024, Vice Media laid off most of its journalists explaining that their online publishing to vice.com was no longer sustainable. Rumors quickly started appearing online from current and past employees that their website would be deleted in its entirety (@janus.bsky.social, 2024; @thelincoln.bsky.social, 2024; @theloniusly, 2024). Rapid archiving as websites or organizations fall apart is not a new practice for web archivists — ArchiveTeam, a collective of hobbyists, has long maintained a list of sites that are dying or at risk of deletion — but we had never used Browsertrix to capture a "large website" in full before and this seemed like a good test case (ArchiveTeam; Jackson, 2023). To begin we ran some tightly scoped tests of contributor pages and started a crawl of the full site before realizing that Vice's website was too large for our current infrastructure and the upper per-crawl limit of 50,000 pages on our hosted Browsertrix service. Browsertrix Crawler also spent too long on each page, as Vice's articles allow users to continue scrolling past the article to load a new one — a behavior that (despite changing the URL for the user as the page scrolls) the crawler did not see as a unique URL. After updating the crawler to stop scrolling when pages update the URL, increasing the page limit of our development instance, and scaling it up to 32 browser windows, we were able to crawl the entire English language site, successfully capturing 1.2 million pages and over 2.7 TB of data over the course of 7.5 days. We aren't quite ready to release these capabilities as part of our hosted service, but as we continue to improve the reliability of our infrastructure, workflow scoping settings, and performance of ReplayWeb.page, we also expect to increase these limits accordingly in the future.

While an instructive test case, Webrecorder isn't the only one crawling news websites with Browsertrix. The National Library and Archives of Quebec (BAnQ), a client of Webrecorder, uses Browsertrix through our hosted service to regularly archive the websites of a dozen major Quebec

newspapers, utilizing browser profiles to crawl paywalled content that would otherwise be inaccessible for archiving. They cite the "flexibility of Browsertrix and the functions to create [b]rowser [p]rofiles," as well as the "ease of viewing and replaying the crawls" from within the Browsertrix interface as significant advantages to their work. While initial results have been positive, certain websites with bot detection systems in place will sometimes block the crawler from continuing and while Browsertrix's crawl workflows include a user agent setting, allowing automated activity according to user agents requires cooperation from website administrators.

Another difficulty with archiving large news sites is picking up new content on subsequent crawls without duplicating content that has already been archived in previous crawls. To this end, we plan to add an option to crawl workflows in a future release of Browsertrix that will enable subsequent crawls within a workflow to only capture new pages as found on user-specified "index" pages. Once implemented, this feature will be useful for the BAnQ and other institutions that crawl news sites on regular schedules to ensure that only new content is archived during crawling while keeping crawling costs and storage footprint of the resulting archived items as small as possible. The groundwork for this feature was already laid as part of the assistive QA features in Browsertrix 1.10 with the addition of the pages list within each archived item.

Outside of archiving news content for historical preservation, Browsertrix and ReplayWeb.page have also been used to give journalists the ability to protect journalistically relevant web content from link rot. In December 2022, The Associated Press (AP), in collaboration with the Starling Lab for Data Integrity, covered a story regarding the use of facial recognition and other machine learning-based surveillance technologies by police forces that were a cause for community concern in Hyderabad, India. They used open source investigation methods to identify and authenticate social media posts and other media posted by police agencies. Reporters at AP used Browsertrix to capture embedded posts from X (formerly Twitter) and display them within their test case article as self-contained web archives to preserve the source should the original post be removed as a result of their reporting. Because embedded social content is usually not available at a URL that can be loaded by a crawler or played back in a web archive viewer, Starling used Webrecorder's oembed.link service — a simple, freely available web application that generates a web page at a unique URL with the requested embed available within the page — to create a web page out of the embed for archiving. Archived captures of the Tweet embedded within the oembed.link pages were then digitally signed by Browsertrix upon archive creation to create a provable chain of custody. These were displayed within the Associated Press story using ReplayWeb.page's archival receipt embed mode which provides end users with the file, and signing information proving its creation by a trusted source.

Starling's most recent collaboration with Black Voice News, through their story Combating Racism as a Public Health Crisis, was launched as an effort by Black Voice News to better understand, report on, and visualize data for Black Californians. The initiative uses web archives to equip the Black community with a data dashboard to track and address regional and local systemic inequities. Racism is a driving force of the social determinants of health (such as housing, education and employment) and is a barrier to health equity. By capturing and aggregating data from public officials and jurisdictions — posted on websites vulnerable to link rot — this project was able to give evergreen access to accurate data.

This project goes a step further by recording the hashes of their archived content — provably unique, strings of text determined by a file's contents — to the Numbers, Avalanche, and Likecoin blockchains. Copies of these hashed and signed archives were also provided through IPFS: a decentralized, peer-to-peer data sharing network, allowing users to download, inspect, and verify the authenticity of the documents themselves.

Combating Racism as a Public Health Crisis also used Browsertrix for capture and ReplayWeb.page to display the primary source documents they reference. Instead of using ReplayWeb.page's archive receipts directly to display provenance data, Starling built their own wrapper for ReplayWeb.page including the blockchain transaction IDs in addition to the signing metadata mentioned above. This method of blockchain-based timestamping can cryptographically prove the existence of their web archives at a given point in time without revealing the contents of the archive until the story goes live, adding another layer of provenance to the existing timestamp provided by the signed WACZ.

Publishers have also taken advantage of Browsertrix and ReplayWeb.page to archive and simplify ongoing hosting and maintenance of digital publication and digital humanities projects. Stanford University Press makes archived versions of their digital projects available online on their own website alongside the live projects through an embedded instance of ReplayWeb.page. One such example is *Black Quotidian: Everyday History in African-American Newspapers*, a 2019 Garfinkel Prize in Digital Humanities award-winning digital humanities project by Matthew F. Delmont which "guides readers through a wealth of primary resources that reveal how the Black press popularized African-American history and valued the lives of both famous and ordinary Black people" (Delmont). Visitors to the project site are able to navigate through an "Archive" link on the project landing page to an archived version of the project. This ensures that the project can continue to be served to users from a WACZ file with all of its original assets and interactivity preserved even if ongoing maintenance and hosting of the project becomes increasingly difficult with time as its technological dependencies become obsolete.

This model applies broadly to digital humanities projects, which are often developed in the context of project-based funding such as grants without funding earmarked for their long-term maintenance. By archiving these sites with Browsertrix, it becomes possible to not only preserve but also continue to host them at greatly reduced cost and complexity. Rather than needing to maintain web and database servers for a digital humanities project and continue to update dependencies such as JavaScript libraries which become obsolete quickly, it's only necessary to serve a static HTML page with ReplayWeb.page embedded as a web component and to point it at the WACZ file containing the archive of the project, which can be served from any file server or S3 bucket. A guide to embedding ReplayWeb.page is freely available as part of the ReplayWeb.page documentation site (Webrecorder). In addition to its use in digital humanities projects, embedded instances of ReplayWeb.page have been implemented in several digital archives and public access systems for digital collections, enabling web archives to be easily accessed and replayed alongside digital media in other formats. One such example of this is the Feminist Institute, which archives websites using tools such as Browsertrix Crawler and ArchiveWeb.page and subsequently makes them available to users in their digital archive (The Feminist Institute).

Webrecorder has additionally been involved with the Mellon-funded Embedding Preservability project, led by NYU Libraries. This project works with digital publishers to ensure that the publications they create are preservable despite complexities such as "dynamic features including embedded audio and video, map navigation, embedded software, and annotating" (NYU Libraries, 2021). By conducting testing of publications created on several digital publication platforms with Browsertrix, Webrecorder was able to provide feedback back to the project team and publishers, enabling them to reduce barriers to archiving directly at the source.

Quality assurance in Browsertrix

Browsertrix 1.10, released in May 2024, includes a state of the art machine-assisted QA process designed to help practitioners address some of the inefficiencies in current QA workflows and focus their attention and judgement where it is most valuable. Screenshots, extracted text, and detailed reports of the resources loaded per-web page are collected during crawling and later compared against the replay of each page during an *analysis run*, outputting per-page match percentages that describe how well the replayed archive compares to what the browser originally received from the live web during crawling. The results are then presented to users in a UI, sortable by screenshot and text comparison percentages.

Extracted text from the crawl and from replay of the page during the analysis run is compared on the basis of Levenshtein distance, the "minimum number of single-character edits (insertions, deletions or substitutions) required to change one word into the other" (Wikipedia). Screenshots are compared by measuring the perceived difference between color samples and by the intensity of difference between pixels using the open source Pixelmatch JavaScript library. Finally, crawl analysis counts the number of page resources that return "good" (2xx or 3xx) and "bad" (4xx or 5xx) HTTP status codes during the crawl and during replay, which may indicate that particular resources were not captured during the crawl or are unable to load as expected during replay.

The QA interface allows users to quickly identify pages based on their assigned scores, visually inspect the screenshots, extracted text, and resource counts for a given page, and quickly switch to the replay of that page for further inspection if needed. As these QA heuristics are meant to inform rather than replace human judgement, users are provided the opportunity to approve, reject, and comment on each page, as well as leave an overall review score for the crawl as a whole.

The initial version of QA features in Browsertrix aims to help users answer questions such as:

- How well did a page get archived?
- Which pages warrant human attention?
- Which pages were the most / least successful while crawling?
- When there are issues, are they crawl issues or replay issues?

With this information, it's possible for users to know whether to change the configuration of their crawl workflow to, for example, give each page additional time to load before capture, or whether a crawl was largely successful but only a few individual pages should be patched. When content has been captured during the crawl but does not replay as expected, the information gathered during QA can additionally be used to diagnose and fix replay issues in future releases of ReplayWeb.page.

Patching crawls that the review process identifies as missing content is currently possible by using the ArchiveWeb.page browser extension to archive a given web page manually, then uploading the resulting WACZ file to Browsertrix and combining it with the crawl in a collection. This enables users to replay, download, and share the original crawl and patched content together, making the patching process invisible to end users such as researchers who will interact with the archived content.

This feature greatly enhances users' experience and confidence while crawling dynamic sites, websites behind paywalls, and large websites, allowing them to quickly identify whether the resulting archived item has issues or is ready to be moved to preservation or access systems. For instance, in the case of crawling a news site behind a paywall, the screenshots and extracted text comparisons in the QA user interface allow users to at a glance ensure that the crawl browser was indeed signed in and not hitting paywall pages, that page contents were appropriately captured, and that images, videos, and other embedded resources render as expected during replay, without needing to manually inspect each individual page within a crawl. This promises to reduce the amount of repetitive labor involved in quality assurance workflows for web archiving practitioners, responding to the needs identified in current web archiving literature.

Conclusion

As news media and primary sources increasingly are found online, behind paywalls, or as interactive JavaScript-heavy web applications, it will become increasingly crucial for librarians, archivists, publishers, and others invested in preserving news media to have access to tools that are suited to the task. As evidenced by the collaborations and case studies presented in this paper, Webrecorder seeks to ensure that practitioners have the means to conduct the important work of preserving and providing access to news media regardless of its format. We hope that Browsertrix and its features, including browser profiles and the quality assurance (QA) features discussed in this paper, will continue to lower barriers to entry for web archiving, making it easier than ever for those invested in the work of maintaining access to culturally important websites and pages to succeed in their pursuits even as the web continues to develop and evolve.

References

ArchiveTeam. n.d. “Deathwatch”. Accessed May 17, 2024. https://perma.cc/F6Y2-HA5Q.

Bicho, Grace, Meghan Lyon, and Amanda Lehman. 2023. “The Human in the Machine: Sustaining a Quality Assurance Lifecycle at the Library of Congress.”, Accessed September 21, 2023. https://digital.library.unt.edu/ark:/67531/metadc2143888/.

Delmont, Matthew F. 2019. “Black Quotidian: Everyday History in African-American Newspapers.” Stanford University Press. Accessed May 17, 2024. https://sup.org/books/title/?id=29420.

Dowd, Trone L. (@theloniusly). 2024. “Word on the street is they're deleting the @vice website. This means that as of later this week, all three publications where I accrued thousands of bylines since the very start of my career will no longer have an online presence.” X, February 22, 2024. https://perma.cc/ZUF4-WM9T.

Feissali, Kourosh and Jake Bickford, Jake. 2023. “Open Auto QA at UK Government Web Archive.” Accessed September 21, 2024. https://digital.library.unt.edu/ark:/67531/metadc2143893/.

Grotke, Abbie. 2011. “Web Archiving at the Library of Congress. *Computers in Libraries* 31, no. 10 (December): 16-19. https://web.archive.org/web/20131215201723/http://www.infotoday.com/cilmag/dec11/Grotke.shtml.

International Internet Preservation Coalition. n.d. “About Us.” Accessed May 17, 2024. https://perma.cc/94J5-UWQJ.

Internet Archive. n.d. “Wayback Machine.” Accessed May 17, 2024. https://web.archive.org/.

Jackson, Andrew. 2023. “What makes a large website large?” Accessed May 17, 2024. https://perma.cc/ZX7J-4SSZ.

Michel, Lincoln (@thelincoln.bsky.social). 2024. “Can I just... republish my Vice pieces on my newsletter if they actually delete the site?” Bluesky, February 22, 2024. https://perma.cc/RA58-7LRS.

New York University Libraries. 2021. “The Andrew W. Mellon Foundation Awards NYU $502,400 For Libraries Project to Expand Capabilities For Preserving Digital Scholarship.” Accessed May 17, 2024. https://perma.cc/QA3X-TJGJ.

Perma.cc. n.d. Accessed May 17, 2024. https://perma.cc/.

Reyes Ayala, Brenda, Mark E. Phillips, and Lauren Ko. 2014. “Current Quality Assurance Practices in Web Archiving.” Accessed September 21, 2023. https://digital.library.unt.edu/ark:/67531/metadc333026/.

Rose, Janus (@janus.bsky.social). 2024. “rumors flying that VICE is going to be deleting our entire website, which hosts 14 years of my professional work. senior leadership was given multiple chances to dispel this rumor today and has not done so. i’m also now disabled from downloading my emails WHEEEE I LOVE WORKING IN MEDIA.” Bluesky, February 22, 2024. https://perma.cc/8FPE-UFBF.

Taylor, Nicholas. 2016. “Rethinking Web Archiving Quality Assurance for Impact, Scalability, and Sustainability.” Accessed September 21, 2023. https://perma.cc/XZZ8-4ZVM.

The Feminist Institute. nd. “Archived Websites.” Accessed May 17, 2024. https://www.thefeministinstitute.org/digital-archive/archived-websites.

Webster, Peter. 2017, “Users, technologies, organisations: Towards a cultural history of world web archiving.” In *Web 25. Histories from 25 Years of the World Wide Web*, edited by Niels Brügger, 179-190. https://hcommons.org/deposits/item/hc:26187.

Webrecorder. 2021. “Web Archive Collection Zipped (WACZ)”. Accessed May 17, 2024. https://perma.cc/73WQ-LNF2.

Webrecorder. n.d. “Browsertrix.” Accessed May 17, 2024. https://browsertrix.com.

Webrecorder. n.d. “ReplayWeb.page documentation.” Accessed May 17, 2024. https://replayweb.page/docs/.

Wikipedia. n.d. “Levenshtein Distance.” Accessed May 17, 2024. https://perma.cc/9W52-56GW.