Creating, storing, and sharing your own web archives with open source Webrecorder tools

Mulliken, Jasmine Tiffany

jasmine.mulliken@stanford.edu
Stanford University, United States of America

Kreymer, Ilya

ikreymer@gmail.com
Webrecorder

This half-day tutorial will provide DH researchers with an introduction to web archiving tools, which they can use on their own or via a free and open-source toolset provided by Webrecorder.

Web archiving has recently garnered mainstream attention for its indispensability in preserving online content in emergency situations, as with the Saving Ukranian Cultural Heritage Online (SUCHO) project. Launched in February 2022, two days after Russia's invasion of Ukraine, SUCHO has worked to capture web-based Ukrainian artifacts, and Webrecorder tools have been an essential part of this workflow (see especially Verma 2022; Serrano 2022). But web archiving isn't only for emergency situations. Key to the crowdsourcing success of efforts like SUCHO has been the user-friendly tools that amateur, as well as advanced, archivists can easily acclimate to. It's this kind of accessibility that makes Webrecorder's tools especially useful for digital humanists pursuing individual or small-group-led projects that require direct handling of web-based content and data. Whether it's to archive your own website, web publication, or student web projects; to capture web archives of important and frequently updated web content from news sites or social media for research; or to curate archived content for others to engage with, this workshop will provide the foundation for getting started with hands-on (and automated!) web archiving.

The tutorial will start with an introduction to "high-fidelity web archiving," covering the basics of what is meant by these terms, and a demonstration of a specific use case and workflow for archiving digital humanities and digital scholarly works published by Stanford University Press. As part of its Mellon-funded digital publishing initiative (Harvey 2019), SUP has partnered with Webrecorder to produce web archived versions of all eleven of its digital publications (Mulliken 2020). We will discuss the value of scholarly web archiving as an intervention to the perceivable ephemerality of digital scholarship and as a way to add value, via longevity, to non-traditional publication formats. Thus, this part of the tutorial will cover the rationale behind, ethical concerns of, and process of creating the web archives for complex digital publications using Webrecorder tools (Browsertrix Cloud, ArchiveWeb.page), storing them in an existing institutional repository, and making them available to the general public as a statically hosted site with minimal maintenance.

In the next part of the tutorial, we will cover how users can create high-fidelity web archives of their own using the Browsertrix Cloud crawling system. Users will be presented with an interactive user interface to create crawls, allowing them to configure crawling options. We will describe the crawling configuration options and how these affect the result of the crawl. Users will have the opportunity to run a crawl of their own for at least 30 minutes and to see the results. We will then discuss and reflect on the results.

In the following section, we will discuss how the web archives can be shared with others, using the ReplayWeb.page viewer. Participants will be able to download the contents of their crawls (as WACZ files) and load them on their own machines. We will also present options for sharing the outputs with others, by uploading to an easy-to-use hosting option such as Glitch or our custom WACZ Uploader. EIther method will produce a URL which participants can then share with others, in and outside the workshop, to show the results of their crawl. We will discuss how, once complete, the web crawl is no longer dependent on the crawler infrastructure, but can be treated like any other static file, and, as such, can be added to existing digital preservation repositories.

In the final section, we will demonstrate how to augment the automated crawling with manual patching of the crawl using the ArchiveWeb.page extension. Participants will be able to use the extension to patch any content that was missed by the automated crawler, or to capture additional content that was not part of the original crawl. We will discuss how these tools can be used in combination to create both automated and manually-created web archives, and discuss the trade-offs between the two. Participants will then be able to upload their patched web archives and share with others, giving them another opportunity to practice a learned skill. In the end, we will discuss any issues encountered. We'll also discuss current and on-going challenges of web archiving, including web-archiving ethics and what can be done to make these tools even easier, giving special attention to how we can make web archiving more accessible to digital humanities practitioners.

The tutorial will cover: basic discussion of web archiving practices, introduction to web archiving terms, discussion of formats used, including WARC and WACZ, an overview of a specific use case at Stanford University Press, followed by hands-on web archiving experience (both automated and manual), sharing/publishing of web archives, and discussion of ethics, challenges, and things learned during the course of the session.

Participants will leave the workshop with actual, web archive data which they can take with them and which they will have optionally published on the web for others to use.

Outline

Introduction to web archiving	10 min
Use Case: Web archiving workflow at Stanford University Press; other useful scenarios, ethics	25 min
Break	5 min
Hands-On: Automated web archiving with Browsertrix Cloud	35 min
Break	5 min
Hands-On: Viewing web archives and sharing with others	20 min
Discussion: Q&A, issues so far	15 min
Break	5 min
Hands-on: Manual archiving with ArchiveWeb.page extension	20 min
Discussion: Automated vs manual, challenges encountered	20 min
Wrap-Up	10 min

Workshop Leaders/Presenters

Ilya Kreymer (ikreymer@gmail.com) is the founder and lead developer of the Webrecorder project. Ilya has worked in the field of web archiving for 11 years, and on the Webrecorder project since 2014. From 2011-2014, he worked at Internet Archive, where he contributed to their Wayback Machine. From 2016-2019, he was working on Webrecorder at Rhizome.org, an arts non-profit focused on preserving online digital culture. Since 2020, Webrecorder project has been an independent entity focusing on expanding the open-source web archiving ecosystem.

Jasmine Mulliken (jasmine.mulliken@stanford.edu) is a digital humanist working at the intersection of publishing, project development, and digital archiving. Currently serving as Production and Preservation Manager, Digital Projects at Stanford University Press, a role she's been in since 2016, her responsibilities include evaluating the technology powering the digital projects SUP publishes to ascertain their sustainability and amenability to current and developing web hosting, archiving, and preservation methods. She works with Stanford University Libraries and external preservation initiatives and agencies to identify, establish, and execute archiving solutions for the unique needs and formats of digital publications. She has been using Webrecorder's tools for SUP's digital projects since 2017 and has coordinated SUP's partnership with Webrecorder since 2019. She holds a Ph.D. (2011) in English with an emphasis in Digital Literacies and Literatures from the University of Texas at Austin.

About Webrecorder

Webrecorder project builds and maintains an ecosystem of open source tools to support decentralized web archiving. From the beginning, Webrecorder has focused on empowering others to make their own archives and store them wherever is most convenient, from cloud services like AWS or Digital Ocean to integrating with existing digital repository systems, like Archipelago. Our motto is "web archiving for all" and we place emphasis on quality over quantity, allowing users to create archives as accurately as possible. Webrecorder is currently supported by a grant from the Filecoin Foundation and has previously been funded by Mellon and various open source development and support contracts.

Target audience

- Librarians who collect, curate, or share web-based digital content
- Researchers working with social media or frequently updated web-based news sources
- Creators of web-based content like digital projects, course websites, web-based scholarship
- Students and faculty with web-based digital portfolios
- Historians of web content

Technical Requirements

Presenters will need screen and projector and wifi internet access for their laptops. Participants should bring their own laptops. Charging outlets for participants would be useful, as would be tables to accommodate everyone working on laptops.

Resources

Participants will be provided with online and print resources at the time of the workshop. Before the workshop, participants will be asked to submit basic info to generate an account in the user interface to be used in the workshop.

Bibliography

Harvey, Alan (2019): “The Stanford University Press digital publishing initiative receives $1.15 million to implement phase 2 of the program”, in: *supDigital Blog* https://blog.supdigital.org/press-release/.

Mulliken, Jasmine (2020): “Archival success!”, in: *supDigital Blog* https://blog.supdigital.org/sup-webrecorder-partnership/.

Serrano, Jody (2022): “How to Stop Ukrainian Websites From Vanishing During War”, in: *Gizmodo* , https://gizmodo.com/how-sucho-stops-ukrainian-websites-vanishing-in-russias-1848737441.

Verma, Pranshu (2022): "Meet the 1,300 librarians racing to back up Ukraine’s digital archives", in: *Washington Post* https://www.washingtonpost.com/technology/2022/04/08/ukraine-digital-history/.