

SERVER-SIDE WEB ARCHIVING USING REPROZIP- WEB

Katherine Boss

*New York University
USA
katherine.boss@nyu.edu
0000-0003-2148-8386*

Ilya Kreymer

*Webrecorder Software
USA
ilya@webrecorder.net*

Vicky Rampin

*New York University
USA
vicky.rampin@nyu.edu
0000-0003-4298-168X*

Rémi Rampin

*New York University
USA
remi.rampin@nyu.edu
0000-0002-0524-2282*

Abstract - Current client-side, or “static,” web archiving crawlers have been tremendously successful in capturing and archiving millions of pages of the internet. Unfortunately, over the decades the web has evolved beyond the reach of many of these crawlers, and today’s static crawlers fail to capture the look, feel, and functionality of a significant amount of interactive web content, including maps, visualizations, database-reliant projects and social media feeds. Archiving these dynamic websites requires a different approach, including a server-side web archiving option.

ReproZip-Web is an open source, grant-funded [1] web-archiving tool that can address this need. It builds on the high-fidelity crawling tools of Webrecorder by also encapsulating a dynamic web server software and its dependencies. The output is a self-contained, isolated, and preservation-ready bundle, an .rpz file, with all the information needed to replay a website, including the source code, the computational environment (e.g., the operating system, software libraries) and the files used by the app (e.g. data, static files). Its lightweight nature makes it ideal for distribution and preservation.

This interactive workshop will be particularly useful for web archivists, digital archivists, digital humanities scholars and others seeking to archive and preserve

complex web projects. Attendees, who should be familiar with the command line interface, will practice packing and tracing a web application and recording the front-end of the site using ReproZip-Web. They will then be able to test replaying the site from the newly created and preservable .rpz file.

Keywords - **Dynamic web archiving, Server-side web archiving, ReproZip-Web, Webrecorder**

Conference Topics - **We’re All in This Together; Immersive Information**

1. REFERENCES

- [1] Institute of Museum and Library Services, “Preserving the Dynamic Web: Building a Production-level Tool to Save Data Journalism and Interactive Scholarship,” *NLG-L Recipient, LG-250049-OLS-21*, August, 2021. <http://www.imls.gov/grants/awarded/lg-250049-ols-21> (accessed Mar. 10, 2023).