Preserving the Dynamic Web: Building a Production-level Tool to Save Data Journalism and Interactive Scholarship

The NYU Division Libraries, in collaboration with the NYU Visualization and Data Analytics Research Center and Webrecorder, an independent organization developing open-source web archiving tools, requests a two-year project grant of $249,999.09 to bring our IMLS-funded web archiving prototype, ReproZip-Web, into production. The project is in the National Digital Infrastructures and Initiatives category and would be in the scaling phase of maturity. This work will be conducted with the participation of our external partners, ProPublica and the Data Desk team at the *Los Angeles Times*. Further development of our prototype will enable the capture of a widerange of dynamic websites, at scale, for long-term archiving and preservation.

Statement of National Need

Data journalism stories are among the most innovative and original works being produced by newsrooms today. Iconic examples include "Old Oil Wells" by the *Los Angeles Times*, "Where Harvey's effects were felt the most in Texas," by *The Texas Tribune*, and "Are Hospitals Near Me Ready for Coronavirus?" by ProPublica. Projects like these, created by news organizations in dozens of countries, are custom-built websites that display content dynamically in the browser. The focus of these stories, in keeping with the larger missions of journalism, is to serve underserved communities, expose inequalities in society, and give voice to the concerns of people from all backgrounds. During the COVID-19 pandemic they have taken on a compelling gravity, as they help us understand the spread, severity, and impact of the coronavirus. Yet because of their technological complexity, many of these websites cannot be fully or systematically captured by current web archiving tools. As a result, they are disappearing.

Journalists aren't the only ones documenting, analyzing, and recording history. Digital humanities (DH) projects, often geared towards illuminating how the past impacts our future, have much in common with dynamic news stories, and face the same archiving issues. Both fields produce unique and dynamic websites that incorporate maps and other visualizations, allowing readers to explore and interrogate underlying data. In doing so, they enable the public to create meaning from large volumes of information, personalize stories, or find new insights in historical, literary, and artistic works. Impactful DH works have changed the ways we interact with current events, history, art, literature, and more. Examples of high-profile projects include "Slave Voyages, "Mapping the Scottish Survey of Witchcraft Database," and "Mapping Police Violence." Researchers and academics struggle to maintain and host these projects for the long term, and face few options for depositing them in university repositories. Consequently, DH projects are also being lost.

The most widely used web archiving technologies, like the Heritrix crawler used by the Internet Archive, have been successful in capturing static snapshots of web content for decades. Unfortunately, these crawlers fail to capture the look, feel, and functionality of a significant amount of dynamic content, including social media feeds, interactive maps, visualizations, and database-reliant websites.[1] This is because the evolution of dynamic web design has outpaced the ability of web archiving crawlers to capture it. This is an issue for nearly all media, but especially digital media, from e-books to videos, in which rapid technological innovations precede the ability to capture or

[1] Boss and Broussard, "Challenges of Archiving and Preserving Born-Digital News Applications."

archive new formats.[2] Web crawlers built to capture the static Internet are missing an increasingly large number of sites. They are able to capture snapshots of the "front end" of a website, the part users interact with through a browser, but not the browser itself (which is increasingly important), or any "back end" content and dynamic elements that render in the browser.

In addition to issues of incomplete capture, there is the question of future access and replay of thesites. In the future, the ways we access the web may change so drastically that those technologies cannot display web archive files. Traditionally, archivists have approached this type of problem by migrating collections from obsolete formats to more modern, accessible ones. Libraries convert films from VHS tapes to DVDs, then to streaming MP4s; they move print newspapers to microform reproductions and then to digitized PDF/A files. This approach, known in the archiving community as migration, changes the digital object in order to prepare it for access and rendering in future environments. This strategy has its shortcomings, but has been largely successful for preserving physical materials and static digital objects like text, images, or sound.[3] However, 50 to 100 years from now, if the internet and computers have changed so much that today's web archives are difficult to access and view for researchers, migration of these collections would be challenging. An emulation-based archiving strategy offers another way. It does not change the digital object, but instead recreates the environment in which the object was previously rendered via an emulator.[4]

This project would advance an emulation-based approach to web archiving by bringing together two technologies to capture the dynamic web: ReproZip, a tool for computational reproducibility, and Webrecorder, a suite of open source high-fidelity web archiving tools. Webrecorder tools can archive highly complex web pages and web applications which load in the browser. However, Webrecorder does not have the capability to archive or encapsulate the web server, which may also serve dynamic content. ReproZip is designed to create a reproducible bundle of all dependencies of an app (such as sourcecode, data, and workflow steps), and is well suited to archive running servers, database operations, and any other dynamic operations. While Webrecorder tools create an encapsulated archive of the content loaded in a user's browser, ReproZip can create an encapsulation of a dynamic web server software and its dependencies. By leveraging the capabilities of both tools, we hope to create a unified preservation system capable of encapsulating and preserving dynamic web server environments and dynamic web sites as accurately as possible.

Beyond the technical challenges of capturing and archiving dynamic websites, there are organizational barriers to advancing a solution, especially in newsrooms. In the early 2000s, as newspapers became less profitable and budgets were cut, hundreds of in-house newsroom libraries were shuttered.[5] Today, digital archivists at news organizations are rare, and few, if any, newsroom staff wake upeach morning thinking about how to save their digital content. Consequently, web archiving is often an afterthought. In 2002 only seven percent of newsrooms with libraries (already a minority) were conducting any sort of web archiving.[6] Though outside organizations such as the Internet

[2] Hansen and Paul, *Future-Proofing the News*.

[3] von Suchodoletz and van der Hoeven, "Emulation"; Broussard and Boss, "Saving Data Journalism."

[4] von Suchodoletz and van der Hoeven, "Emulation," 147.

[5] Paul and Hansen, "Reclaiming News Libraries"; Hansen and Paul, *Future-Proofing the News*.

[6] Paul and Hansen, "Reclaiming News Libraries."

Archive have stepped in to save millions of pages of articles, a tremendous amount of online news has been lost, more so now that an increasing amount of dynamic news cannot be captured in the first place.

These losses are eroding the collective cultural record.[7] News stories are known as the "first draft of history," and archiving this history is vital for our cultural memory and any future research based on journalism. Libraries have the mandate, expertise, and infrastructure to save these works for the long term. In a 2018-2019 "Saving Data Journalism" planning grant generously funded by the IMLS (LG-87-18-0062-18), NYU Libraries built the first-ever emulation-based web archiving prototype capable of capturing the look, feel, and functionality of a database-reliant news app. Our prototype, ReproZip-Web, is an open-source web archiving tool that creates a self-contained, isolated, and preservation-ready bundle with all the information needed to reproduce a news application or DH project.[8] This bundle, an .rpz file, contains all of the dependencies needed to replay dynamic web apps, including the precise information and source code for the computational environment (e.g., the operating system, software libraries) as well as all the files used by the app (e.g. data, static files). Its lightweight nature makes it ideal for distribution and preservation. However, further work is needed to make it operational in newsrooms, where the archiving process must begin.

Project Design

The goal of *Preserving the Dynamic Web* is to provide a pathway to preserve these dynamic web apps in service of cultural memory and our collective historical record. We previously demonstrated that our prototype can archive and reproduce dynamic news apps, and in this project we will build on that success to refine and operationalize ReproZip-Web. We aim to extend its capabilities, improve its usability, and build an ecosystem of tools for the preservation of and access to dynamic web applications. To do this, we will undertake two main streams of work: 1) technical improvements to and expansion of ReproZip-Web and Webrecorder and 2) user experience (UX) testing with digital archivists, data journalists, and computational humanists (our target users at this time) to optimize adoption and usability. The first stream of work will be led by two uniquely qualified software engineers: Rémi Rampin, lead developer on ReproZip, and Ilya Kreymer, lead developer of pywb and Webrecorder. The second stream of work will be led by Katherine Boss and Vicky Rampin, two of the faculty investigators on the *Saving Data Journalism* (SDJ) planning grant, both of whom have experience with UX testing and software-development project management. The team's committed project partners include the ProPublica data desk team, the data and graphics department at the *L.A. Times*, and the NYU Libraries Digital Scholarship Services department. These partners have committed to attending two UX sessions over the course of two years, and to working with our team to provide at least one of their news applications for testing. We plan to expand this list of project partners in the first stage of the grant to include several more newsrooms, digital humanities scholars, and digital archivists.

An outline of the project plan in months is as follows:

PM=Project Manager **PDs**=Project Directors **Dev GA**=Development Graduate Assistant
Dev Leads=Development team leads **UX GA**=User Experience Graduate Assistant

[7] Broussard and Boss, "Saving Data Journalism," 1208.
[8] Boss, Steeves, Rampin, Chirigati, and Hoffman, "Saving Data Journalism."

Sept - Oct '21	(PM): Secure participation from additional project partners for UX work. (PDs): Advertise Dev GA position, organize an all-hands Nov. planning session, begin soliciting apps for testing.	**Sept - Oct '22**	(PDs/US GA): Analyze and share UX feedback, schedule the second round of UX sessions. (Dev Leads/Dev GA): Begin user/technical documentation. Incorporate feedback, iterate and develop.
Nov - Dec '21	(PM/PDs): Advertise UX GA position. (Full team): Hire and onboard Dev GA, hold dev planning session. (Dev Leads/Dev GA): Secure news apps for testing. Begin work on existing challenges, e.g., crawling Mapbox tiles, graphical user interface (GUI) for ReproZip-Web.	**Nov - Dec '22**	(PDs/UX GA): Hold second UX sessions — revised task and scenario testing. (Dev GA): Continue work on user/technical documentation, finishes work in Dec. (Dev Leads): Incorporate feedback, iterate and develop.
Jan - Feb '22	(PDs/UX GA): Hire and onboard UX GA, begin planning first round of UX testing. (Dev GA): Finish soliciting news apps and begin testing. (Dev Leads): Existing issues — Mapbox tiles, Amazon S3 content, GUI.	**Jan - Feb '23**	(PDs/UX GA): Hold second UX sessions. (PDs): advertise for freelance web designer. (UX GA): Finishes work in Jan. (Dev Leads): Second dev planning meeting in Feb to incorporate feedback, iterate and develop; updates to ReproServer.
Mar - Apr '22	(PDs/UX GA): Schedule and conduct first UX working sessions. (Dev Leads/Dev GA): Internal testing and existing issues — Mapbox tiles, Amazon S3 content, GUI (ongoing).	**Mar - Apr '23**	(Full team): Attend NICAR conference. (PDs): Analyze and share UX feedback, hire freelance web designer. (Dev Leads): Responsive development based on feedback, finalize GUI.
May - June '22	(PDs/UX GA): Conduct first UX working sessions and analyze UX feedback. (Dev Leads/Dev GA): Internal testing and existing issues. Note: some team members will be out on vacation during this time.	**May - June '23**	(Full team): Dissemination and promotion. (PDs): Begin reporting; work with web designer on a website and branding. (Dev Leads): Development on replay and access of preserved apps.
July - Aug '22	(PDs): Continue to analyze UX feedback and synthesize it for the dev team. Note: some team members will be out on vacation during this time.	**July - Aug '23**	(PDs): Report results to the IMLS; finalize work with web designer on a website and branding.

The PDs will spend the first two months of the project frontloading administrative work and getting ready to undertake testing-and-development cycles to ensure the work of the following months can be done with efficacy. During these months we will secure additional project partners for the UX work, solicit apps from those partners for testing, establish testing and documentation protocols, advertise for a development graduate assistant, create onboarding materials for both project partners and graduate assistants, and schedule a development planning session for November.

In November 2021 the Dev Leads will begin their work, and we will onboard the Dev GA, who will continue soliciting apps from our additional partners for testing. Testing with a wider array of apps will be important for surfacing both immediately fixable issues and wider discussion points to bring up with partners later on during the iterative UX testing phase. The Dev Leads will also work on existing challenges and fixes in November and December that can be addressed immediately. This initial round of development will be based on our planning session and known areas of need for improvement. For instance, in *SDJ* we found that dynamic applications sometimes need access to external APIs and data to work instead of all being served from the same server (e.g., sites that dynamically upload and download data from Amazon Simple Storage Service [S3] stores, and sites that use existing services' map tiles). The Dev Leads will implement solutions to these noted challenges, such as automated crawling of Mapbox projects using Webrecorder's browser extension, Browsertrix Crawler, a highly customizable, browser-based crawling tool that allows users to archive dynamic web pages in a self-directed way.[9] This tool extends Webrecorder's high-fidelity archiving capabilities in a more automated way and includes a system to run site-specific behavior on each page. For example, on social media sites Browsertrix can automatically navigate a feed, scroll, expand comments, load images, play videos and wait for them to finish loading in order to create a more complete archive. Another need that the Dev team can begin to address immediately is developing a Graphical User Interface (GUI) for packing apps with ReproZip-Web to be more user-friendly to those not familiar with the command line. At this time the PDs will also advertise for a UX GA.

In January and February 2022 the PDs will hire and onboard the UX GA, while the development team will test apps that were received from project partners. Using the updated version of ReproZip-Web, the Dev GA will test functionality, user-experience, and archival coverage (how much of the web app was captured), as well as document testing gaps to inform the next development cycle.

March 2022 will be devoted to planning the first round of user testing with our project partners. Users who will test ReproZip-Web for this project will fall into one of three groups: data journalists, who are often responsible for the longevity of their work (as newsrooms may or may not employ a dedicated archivist); computational humanists, who seek to share their scholarly work in a cite-able and reproducible way; and digital archivists, whose mission is preservation. We will work closely with these potential users, gathering their feedback on ReproZip-Web's usability, the time commitment required to use it in their daily workflow, and the types of software they would be willing and able to install on the production servers where ReproZip-Web would be deployed. This work will also include an investigation of packing multiple apps at one time. Feedback will be instrumental in helping us design an effective tool that fits the needs and workflows of newsrooms. We plan to thoughtfully create feedback loops from the users directly to developers.

In April, May, and June 2022 the PDs and the UX GA will conduct the first UX sessions with our project partners to test the current manifestation of ReproZip-Web and create an archival bundle of one of their appsof choice (different from one that we tested with, to ensure coverage of a wide range of apps). Originally,we were

[9] Kreymer, "Webrecorder | Introducing Browsertrix Crawler."

going to bring together partners in person to NYU to do this testing. However, because of COVID-19, we have pivoted to virtual events. During these virtual, recorded UX working sessions, we will do a deeper dive into each user's workflows. We will create a scenario-and-task-based guide and ask that each participant perform some tasks using ReproZip-Web, narrating as they go to identify gaps in the workflow and potential improvements. After each session, we will use the detailed data, input, and user feedback we have gathered to inform the technical development work, which will be highly iterative as we encounter further edge cases and requirements. We will also provide feedback mechanisms outside of these UX sessions via open feedback forms, a mailing list, and/or other communication channels as desired by the participants (and suitable for our purposes, such as a chatroom). While the UX team is conducting these sessions, the Dev team will continue work on known issues and fixes. In July and August 2022, the PDs will analyze and summarize the UX feedback for the Dev team.

In September 2022 the Dev team will focus on addressing feedback from the first UX sessions. Development priorities will be adjusted as necessary to accommodate new feedback. The Dev GA will take the lead on updating user-facing documentation as development continues, with the Dev Leads informing and overseeing the work. The Dev team will be responsible for internal testing during this time, through automated software testing as well as internal user testing. In October 2022 the UX GA will resume work and begin planning the second round of UX testing, to begin the following month. Around this time, the PDs will also submit a presentation proposal to the annual meeting of the National Institute for Computer-Assisted Reporting (NICAR), the primary conference of data journalists in the U.S., in order to unveil the results of our work in the spring. We will also submit proposals to relevant conferences for our other user communities, such as in the computational humanities (e.g. Digital Humanities) and digital preservation (e.g. iPres).

November 2022 through January 2023 will be devoted to the second round of UX testing with project partners, which will follow the same protocol as our first round of testing. We will create a revised scenario-and-task-based guide for new features, functionality, or user interfaces to ReproZip-Web, and ask our participants to archive an app of their choice (again, a different one from the previous two calls for apps in order to test on a wide range). The PDs and the UX GA will compile and analyze this data in preparation for a second development planning meeting in February. As the additional UX data is gathered, the Dev Leads will continue updating user-facing and development documentation and begin work on the replay of the preserved apps through Replayweb.page and ReproServer. The Dev GA will finish work in December 2022, and the UX GA will finish work in January 2023.

In February 2023 the core team of PDs and Dev Leads will hold a second development planning meeting to re-prioritize our development roadmap based on feedback from the second UX sessions. At this time the PDs will also advertise for a freelance web designer. In March, if our proposal was accepted, the full team will also attend the NICAR journalism conference to present on our work.

The remaining four months (March — June 2023) will be dedicated to responsive development based on the February planning session as well as on replay and future access to apps. While the majority of the project work will be devoted to the capture side of archiving, we want to carve out time to address the access side of preservation. For instance, ReproServer exists to provide a way to replay ReproZip bundles in-browser, which is beneficial for those users who are limited in what they can install on their computer but who still want to replay different work. Development work is needed to ensure that dynamic apps captured with ReproZip-Web can be unpacked in-browser. This could open opportunities to interoperate with other web replay technology, such as

Replayweb.page, which could load apps from the web archive first, and direct a certain domain to the running ReproServer (there are already some hooks in Replayweb.page to enable this, which supports this goal). During this time, we will also be augmenting technical and user documentation to ensure accuracy with the newest development work.

While the remaining dev work is unfolding, the PDs will hire a freelance web designer in April, and devote May and June 2023 to working with the web designer on branding the tool and building a user-friendly, accessible website for ReproZip-Web.

The project team values openness and transparency, and we will publish our work openly throughout this project. ReproZip-Web, ReproServer, and Webrecorder are open source. We currently distribute the software via our website and GitHub, and will archive new releases of the tool with Zenodo. We will make ReproZip-Web as easy to use as possible for users and developers alike, and will create a variety of documentation in service of this goal. We will have full documentation of the software for usage and development purposes. Developer-based documentation will detail the architecture and deployment of the software itself. For computational humanists and data journalists, we anticipate creating short, more focused documents such as workflow diagrams for how users might leverage ReproZip-Web in a variety of situations (e.g. packing locally vs packing on a server). For digital archivists, we are particularly interested in replicating the model of the Data Curation Network primers, which are two-page documents aimed at digital preservation practitioners to guide their curation of specific types of works,from Jupyter notebooks to shapefiles (GIS data).

In addition to the software and related documentation, we will have a number of scholarly products from this project. We plan to publish the documentation around conducting these UX sessions, including task-based scenario and interview guides, emails to participants, and a blank copy of our informed consent forms. The RPZ bundles resulting from the tests may or may not be made public via a repository (such as our institutional repository at NYU), according to the terms outlined with partners. For all our work, we will publish openly when we can, but will keep it closed when necessary (for instance, if an industry partner can't share the source code of their apps publicly because of internal policies). Other scholarly products from the project will include peer-reviewed papers and conference presentations and posters.

We plan to evaluate the extent of the projects' successes through a variety of metrics. To ensure that we are staying on task and reaching our goals in a timely manner we will be using a dedicated project management platform such as Monday.com. For our scholarly materials, we'll use more quantitative metrics as indicators of success, such as the number of citations/downloads (the goal is to have a lot of these!). For evaluating the software tooling, we will be benchmarking the quality of the capture, such as the percent of app materials successfully vs. not successfully packed with ReproZip-Web. Throughout the project (and ongoing after the project ends), we will be evaluating usage statistics of software, documentation, and related products. We will also use repetitive qualitative evaluation throughout the project to ensure that ReproZip-Web will be useful to our user groups (through the two UX sessions with project partners). We will be directly interviewing users about their opinion of the tool (in service of usability), asking specific questions to partners about the gap between desired workflow and presented workflows. By collecting this feedback throughout the project, we will be able to tweak our project design and development goals as needed to meet the needs of our users. There will be other measures that will be observed during the UX sessions that provide useful benchmarks to evaluate against, such as the number of clicks to desired results (less is more!).

National Impact

By further developing our prototype we will advance the technological capability to save data journalism and DH projects at scale. Newsrooms like ProPublica have a decade of websites that they want and need to begin archiving in a thoughtful and responsible way, and this project will make that possible. DH scholars worried about the ephemerality of their work will have a solution to keep it accessible in the long term. With a production-ready version of ReproZip-Web, data journalists in the newsroom and DH scholars will be able to pack up each project into a single archivable, distributable, and preservable .rpz file that can be transferred to libraries for long-term preservation and access. And time is of the essence. With each year that passes, the technologies used to build the projects become more and more out of date, making the work increasingly difficult to save. When unveiling a video demo of our successful ReproZip-Web prototype at the 2019 NICAR conference in Newport Beach,[10] our team received a standing ovation from the data journalists in attendance. These communities have felt the loss of their works for more than a decade, and are eager for an archiving solution and partnerships that could be replicated nationally. Bringing our prototype into production so that ProPublica's journalists can save their collection of award-winning data journalism applications will be, in itself, an important success, and is one focus of this project.

A broader impact of this project is the opportunity for newsrooms and libraries to begin collaborating directly on archiving data journalism. DH scholars typically have affiliations with institutional libraries and archives, and often work with them on their projects, while newsrooms usually do not. Libraries and archives have the expertise, the infrastructure in the long-term funding and support of our institutions, and the historical mission to continue this work. With respect to this project, the bundles created by ReproZip-Web are preservation-ready, portable (e.g. these files are often under 1 GB), and self-contained, in that they contain everything necessary to rerun the packed work, from software to data to workflow.[11] As we collaborate with our partners on this project, we will consult with them on how to archive their work and continue conversations about establishing NYU as the long-term home for their archives.

Further development on the Webrecorder Browsertrix Crawler tool to automate the capture of Mapbox visualizations will also have a tremendous impact on the ability of journalists, archivists, librarians and the public to archive these visualizations. Mapbox, an open-source alternative to Google Maps, has been a staple of data journalism projects for almost a decade. The inability of static web archiving technologies to capture these works has left a gaping hole in the cultural record. If this project is successful, hundreds of these sites, like the *L. A. Times'* L.A. County COVID tracker and their California wildfires map, could be saved.

This tool would also enable archiving of dynamic DH project websites that are more complex than what can be handled by a web archive alone. To better define this criteria, we refer to the encapsulation complexity scale[12] which categorizes the difficulty of encapsulating and preserving web objects. Web archives can fully represent objects that have a countable number of resources (URLs), which can be enumerated and stored in a web archive. Such objects

[10] Boss, Steeves, Rampin, Chirigati, and Broussard, "Saving Data Journalism."
[11] Steeves, Rampin, and Chirigati, "Using ReproZip for Reproducibility and Library Services."
[12] Kreymer, "Webrecorder | Web Object Encapsulation Complexity (Part I)."

can be described as having Level 1 encapsulation complexity. However, many complex projects may involve dynamically generated URLs or dynamic search queries sent to a server. For example, a website might perform a search based on user input by loading data from a database on-demand. Such projects can be categorized as having a Level 2 web object encapsulation complexity and are not fully preservable via web archives alone. Thanks to ReproZip's ability to archive and encapsulate a running web server, ReproZip-Web is able to encapsulate and preserve such Level 2 objects and reproduce/replay them, thus expanding the possibilities of web objects that can be encapsulated and preserved. This would allow researchers, librarians, and archivists to begin retiring and archiving complex projects that are not being actively updated, freeing up valuable hosting capacity and allowing the works to be preserved for the long term in university archives.

The operations, incentive structures, and publishing workflows of our different project partners are likely to vary widely, which is why we are including partners of different types (scholarly, commercial and non-profit, legacy and startup) in this grant. In our conversations and user testing with these partners we hope to accomplish a number of goals. First, we hope to gather more information on their various publishing workflows: how their files are organized, what types of security keys or firewalls protect the data, who has access to the server, and other important "unknown unknowns" that we anticipate will arise in these conversations. This information will be collected in the first round of UX interviews and sessions, so that we can use it to inform later phases of the work. This qualitative data will answer many questions related to organizational and operational challenges and opportunities in addressing dynamic web archiving at scale, and allow us to build a solution that will work for a wider range of potential users. However, it is important to note that the scope of this grant will not cover a wider campaign to build specific communities of practice around archiving for our partners, or engage in a national effort to to build workflows around the capture and preservation this tool will enable. That crucial next step will require additional funding, which our team plans to pursue.

To bring more awareness to this issue and promote ReproZip-Web and our extension to Webrecorder capture and replay tools we plan to disseminate this work widely in our scholarship, on social media, and at computational journalism, digital humanities, digital archiving, and preservation conferences. We will also promulgate our results to the wider intersecting communities impacted by the work of this grant (e.g. digital preservation, computational humanities, data journalism). This will include writing blog posts, publishing scholarly manuscripts, giving conference presentations and posters, and doing targeted outreach to user groups. We will also look for opportunities to present workshops to our user communities to train a wider array of people on how to use Reprozip-Web for their own work. ProPublica and the *L.A. Times* have been pioneering partners on this work, and their shared advocacy has significant impact in those communities. We're excited to include new digital humanities partners as well as a key user community to serve through this work.

The NYU Department of Public Affairs is responsible for, among other things, communication with the news media and the external community. Its self-described work is "to tell the University's story." The Public Affairs officer assigned to the Libraries has deep contacts throughout the library and education press as well as the commercial press. They will push the announcement of a completed, grant-funded project to all these outlets, and endeavor to interest reporters in covering ReproZip-Web in depth. The Libraries will announce the new tool on its website news section and on social media.

Looking forward, further work and funding will be needed for important questions related to the description, discovery, and public access to these files. All of those aspects are crucial in being able to transfer the works from newsrooms or researchers' personal cloud storage to libraries. Once libraries are able to host and preserve the works, there are additional questions related to making the projects discoverable and providing access to the archived sites. However, this work is all dependent on establishing a technical tool to capture the works in the first place. This grant advances the next step in the process.

ReproZip-Web, an open-source tool that can easily pack and replay dynamic digital objects, is the first tool of its kind, an archiving software with the potential to save innumerable complex, digitally-born works. Its interoperability with Webrecorder, and the established history of developmentand support for ReproZip, make it a sustainable solution for preserving these valuable projects.

New York University Libraries

SCHEDULE OF COMPLETION*: Preserving the Dynamic Web: Building a Production-level Tool to Save Data Journalism and Interactive Scholarship*

Major Tasks	Date In-Progress												Assigned Project Staff					
ADMINISTRATION + PREPARATION																		
Hire & onboard UX grad student			x	x									n	n				
Hire & onboard Dev grad student	x	x											n	n	n	n		
Development planning	x	x	x										n	n	n	n		
COMMUNICATION																		
Secure participation from additional project partners	x	x											n					
Solicit apps from partners for testing		x	x	x	x								n	n			n	n
Synthesize UX feedback for development team								x	x	x	x		n	n				n
Write user documentation											x	x		n			n	n
Write developer documentation											x	x			n	n	n	
TESTING																		
Conduct first UX session with project parnters								x	x	x			n	n				n
Intra-team testing of coverage					x	x	x								n		n	
ANALYSIS + DEVELOPMENT																		
Analyze results of UX sessions									x	x	x	x	n	n				n
Development of known issues, fixes, desires, and bugs			x	x	x	x	x	x	x	x					n	n	n	
Development of replay and re-access mechanisms to packed work								x	x	x					n	n	n	
Development based on UX session feedback										x					n	n	n	
Project Title: Preserving the Dynamic Web: Building a Production-level Tool to Save Data Journalism and Interactive Scholarship **Principal Investigators/Project Directors:** Katherine Boss and Vicky Rampin	Sept-21	Oct-21	Nov-21	Dec-21	Jan-22	Feb-22	Mar-22	Apr-22	May-22	Jun-22	Jul-22	Aug-22	**Katherine Boss**	**Vicky Rampin**	**Remi Rampin**	**Ilya Kreymer**	**Dev Graduate Student**	**UX Graduate Student**

New York University Libraries

SCHEDULE OF COMPLETION*: Preserving the Dynamic Web: Building a Production-level Tool to Save Data Journalism and Interactive Scholarship*

Major Tasks	Date In-Progress												Assigned Project Staff						
ADMINISTRATION + PLANNING																			
Development planning						x							n	n	n	n			
Hire and onboard freelance web designer						x	x	x					n	n					
Report results to the IMLS											x	x	n	n					
COMMUNICATION																			
Write user documentation	x	x	x	x	x	x	x	x	x	x					n	n	n		
Write developer documentation	x	x	x	x	x	x	x	x	x	x					n	n	n		
Website and branding								x	x	x	x	x	n	n					n
Dissemination and promotion							x	x	x	x	x	x	n	n	n	n			n
TESTING																			
Conduct second UX session with project parnters			x	x	x								n	n				n	
ANALYSIS + DEVELOPMENT																			
Analyze results of UX sessions	x	x				x	x						n	n				n	
Development based on UX session feedback	x	x	x	x	x	x	x	x	x	x					n	n	n		
Development of replay and re-access mechanisms to packed work			x	x	x				x	x					n	n	n		
Project Title: Preserving the Dynamic Web: Building a Production-level Tool to Save Data Journalism and Interactive Scholarship **Principal Investigators/Project Directors:** Katherine Boss and Vicky Rampin	Sept-22	Oct-22	Nov-22	Dec-22	Jan-23	Feb-23	Mar-23	Apr-23	May-23	Jun-23	Jul-23	Aug-23	**Katherine Boss**	**Vicky Rampin**	**Remi Rampin**	**Ilya Kreymer**	**Dev Graduate Student**	**UX Graduate Student**	**Freelance Web Designer**

Year 2

INSTITUTE of Museum and Library SERVICES

DIGITAL PRODUCT FORM

INTRODUCTION

The Institute of Museum and Library Services (IMLS) is committed to expanding public access to digital products that are created using federal funds. This includes (1) digitized and born-digital content, resources, or assets; (2) software; and (3) research data (see below for more specific examples). Excluded are preliminary analyses, drafts of papers, plans for future research, peer-review assessments, and communications with colleagues.

The digital products you create with IMLS funding require effective stewardship to protect and enhance their value, and they should be freely and readily available for use and reuse by libraries, archives, museums, and the public. Because technology is dynamic and because we do not want to inhibit innovation, we do not want to prescribe set standards and practices that could become quickly outdated. Instead, we ask that you answer questions that address specific aspects of creating and managing digital products. Like all components of your IMLS application, your answers will be used by IMLS staff and by expert peer reviewers to evaluate your application, and they will be important in determining whether your project will be funded.

INSTRUCTIONS

If you propose to create digital products in the course of your IMLS-funded project, you must first provide answers to the questions in **SECTION I: INTELLECTUAL PROPERTY RIGHTS AND PERMISSIONS.** Then consider which of the following types of digital products you will create in your project, and complete each section of the form that is applicable.

SECTION II: DIGITAL CONTENT, RESOURCES, OR ASSETS
Complete this section if your project will create digital content, resources, or assets. These include both digitized and born-digital products created by individuals, project teams, or through community gatherings during your project. Examples include, but are not limited to, still images, audio files, moving images, microfilm, object inventories, object catalogs, artworks, books, posters, curricula, field books, maps, notebooks, scientific labels, metadata schema, charts, tables, drawings, workflows, and teacher toolkits. Your project may involve making these materials available through public or access-controlled websites, kiosks, or live or recorded programs.

SECTION III: SOFTWARE
Complete this section if your project will create software, including any source code, algorithms, applications, and digital tools plus the accompanying documentation created by you during your project.

SECTION IV: RESEARCH DATA
Complete this section if your project will create research data, including recorded factual information and supporting documentation, commonly accepted as relevant to validating research findings and to supporting scholarly publications.

SECTION I: INTELLECTUAL PROPERTY RIGHTS AND PERMISSIONS

A.1 We expect applicants seeking federal funds for developing or creating digital products to release these files under open-source licenses to maximize access and promote reuse. What will be the intellectual property status of the digital products (i.e., digital content, resources, or assets; software; research data) you intend to create? What ownership rights will your organization assert over the files you intend to create, and what conditions will you impose on their access and use? Who will hold the copyright(s)? Explain and justify your licensing selections. Identify and explain the license under which you will release the files (e.g., a non-restrictive license such as BSD, GNU, MIT, Creative Commons licenses; RightsStatements.org statements). Explain and justify any prohibitive terms or conditions of use or access, and detail how you will notify potential users about relevant terms and conditions.

The source code of Reprozip-Web is fully open-source and public online via GitHub, licensed under a BSD 3-Clause license. Copyright is held by New York University. All subsequent development on ReproZip-Web will be licensed

A.2 What ownership rights will your organization assert over the new digital products and what conditions will you impose on access and use? Explain and justify any terms of access and conditions of use and detail how you will notify potential users about relevant terms or conditions.

Full license details for ReproZip-Web are available online at https://github.com/reprozip-news-apps/reprozip-web/blob/master/LICENSE. Users are notified automatically when any changes are made to the license or the GitHub repository.

A.3 If you will create any products that may involve privacy concerns, require obtaining permissions or rights, or raise any cultural sensitivities, describe the issues and how you plan to address them.

N/A

SECTION II: DIGITAL CONTENT, RESOURCES, OR ASSETS

A.1 Describe the digital content, resources, or assets you will create or collect, the quantities of each type, and the format(s) you will use.

N/A

A.2 List the equipment, software, and supplies that you will use to create the digital content, resources, or assets, or the name of the service provider that will perform the work.

N/A

A.3 List all the digital file formats (e.g., XML, TIFF, MPEG, OBJ, DOC, PDF) you plan to use. If digitizing content, describe the quality standards (e.g., resolution, sampling rate, pixel dimensions) you will use for the files you will create.

N/A

Workflow and Asset Maintenance/Preservation

B.1 Describe your quality control plan. How will you monitor and evaluate your workflow and products?

N/A

B.2 Describe your plan for preserving and maintaining digital assets during and after the award period. Your plan should address storage systems, shared repositories, technical documentation, migration planning, and commitment of organizational funding for these purposes. Please note: You may charge the federal award before closeout for the costs of publication or sharing of research results if the costs are not incurred during the period of performance of the federal award (see 2 C.F.R. § 200.461).

N/A

Metadata

C.1 Describe how you will produce any and all technical, descriptive, administrative, or preservation metadata or linked data. Specify which standards or data models you will use for the metadata structure (e.g., RDF, BIBFRAME, Dublin Core, Encoded Archival Description, PBCore, PREMIS) and metadata content (e.g., thesauri).

N/A

C.2 Explain your strategy for preserving and maintaining metadata created or collected during and after the award period of performance.

N/A

C.3 Explain what metadata sharing and/or other strategies you will use to facilitate widespread discovery and use of the digital content, resources, or assets created during your project (e.g., an API [Application Programming Interface], contributions to a digital platform, or other ways you might enable batch queries and retrieval of metadata).

N/A

Access and Use

D.1 Describe how you will make the digital content, resources, or assets available to the public. Include details such as the delivery strategy (e.g., openly available online, available to specified audiences) and underlying hardware/software platforms and infrastructure (e.g., specific digital repository software or leased services, accessibility via standard web browsers, requirements for special software tools in order to use the content, delivery enabled by IIIF specifications).

N/A

D.2. Provide the name(s) and URL(s) (Universal Resource Locator), DOI (Digital Object Identifier), or other persistent identifier for any examples of previous digital content, resources, or assets your organization has created.

N/A

SECTION III: SOFTWARE

General Information

A.1 Describe the software you intend to create, including a summary of the major functions it will perform and the intended primary audience(s) it will serve.

Remi Rampin at the NYU School of Engineering developed a tool, ReproZip, that captures and preserves the code, data, and computational environment associated with a research workflow (such as, a script that does data processing). Our previous work (which was funded by an IMLS project grant, and included Rampin) resulted in ReproZip-Web, a prototype that builds on ReproZip and pywb to extend ReproZip's capabilities to pack interactive news apps. In this proposal, we'll take the prototype and develop it so it's ready to be deployed in a production environment. ReproZip-Web is intended to serve librarians, archivists, museum professionals, data journalists, and computational humanists.

A.2 List other existing software that wholly or partially performs the same or similar functions, and explain how the software you intend to create is different, and justify why those differences are significant and necessary.

There are more conventional web archiving methods include crawlers such at Archive-it or the Heritrix crawler (both used by Internet Archive), and the automated archiving feeds of companies like Lexis-Nexis. These methods are not sufficient to capture projects that are interactive, and/or involve databases, streaming data, or interactive graphics. ReproZip-Web will fill this gap.

Technical Information

B.1 List the programming languages, platforms, frameworks, software, or other applications you will use to create your software and explain why you chose them.

ReproZip is a thoroughly-tested tool that has gained significant traction in the reproducible research community. It is primarily written in Python (with parts written in C), packs on Linux (which most web servers run, which makes it attractive), and can unpack on any operating system by using an unpacker plugin (or by simply unzipping the .rpz bundle!). ReproZip-Web has the same major components and development will continue in Python and C. Given we are using an existing software and do not plan a major refactor to change languages and frameworks, we will keep with the existing stack that ReproZip is built on.

B.2 Describe how the software you intend to create will extend or interoperate with relevant existing software.

ReproZip is widely used as a part of digital preservation, workflow, and reproducibility platforms and tools, such as:

+ WholeTale, an open source reproducibility web app (an overview is in this presentation: https://reusableresearch.com/slides/bludaescher.pdf)
+ Emulation as a Service Infrastructure (EaaSI), an open source emulation system (see PI's thread: https://twitter.com/euanc/status/1143966909421019136)
+ Cloud of Reproducible Records (CoRR), an open reproducible records store from NIST (see their repository: https://github.com/usnistgov/corr-reprozip)
+ Model Insertion Checker (MIC), a part of DARPA's World Modelers program (see their docs: https://mic-cli.readthedocs.io/en/latest/model_configuration/01-overview/#step-2-trace-your-model-execution)
+ Spot, a tool to reconstruct pipeline graphs without modifying them (see their paper: https://pubmed.ncbi.nlm.nih.gov/33269388/)

Further, ReproZip bundles (.rpz files) are small enough that they are easily stored in institutional or generalist repositories for widespread distribution. There are more platforms that leverage ReproZip that cannot be listed in full here due to space limitations. ReproZip-Web, the software we'll be productionizing in this grant, produces an RPZ bundle that can be made interoperable with these systems with minimal changes. ReproZip and ReproZip-Web are language-agnostic in terms of what they capture, so they can be used to preserve all manner of work. The work of this grant will be testing and expanding the capabilities of ReproZip-Web to make it more and more interoperable with tools in scholars', archivists', and journalists' workflows. ReproZip-Web itself is the culmination of building interoperability between ReproZip and pywdb, a core web archiving technology.

B.3 Describe any underlying additional software or system dependencies necessary to run the software you intend to create.

ReproZip-Web works by tracing the app's systems calls to automatically identify which files should be included. A user can review and edit this list and the metadata before creating the final package file. Packages can be reproduced in different ways, including chroot environments, Vagrant-built virtual machines, and Docker containers; more can be added through plugins (we have a Singularity unpacker in the works right now, for instance -- we can add/remove virtual machine and container for unpacking at will so we don't rely on a particular one).

B.4 Describe the processes you will use for development, documentation, and for maintaining and updating documentation for users of the software.

We will use Git for version control and GitHub for documentation and communication. ReproZip's documentation thoroughly describes how the tool operates; ReproZip-Web's documentation will describe how to use the product and possible uses and configuration. We have already developed a variety of use cases, and we intend to develop practical guides and case studies for library users, data journalists, and computational humanists.

A regularly monitored email address, users@reprozip.org, is currently used for feedback, questions, concerns, and issues. This address is also used to share use cases with the developers, as well as to report on best practices and lessons learned for reproducibility. Bugs and feature plans are tracked via GitHub issues.

We will also have direct UX sessions with participants (who are part of our intended audience for the tool) and who will surface types of documentation that will be most relevant for them.

B.5 Provide the name(s), URL(s), and/or code repository locations for examples of any previous software your organization has created.

Code: https://github.com/reprozip-news-apps/reprozip-web
Docs: https://reprozip-web.readthedocs.io/en/latest/#

Access and Use

C.1 Describe how you will make the software and source code available to the public and/or its intended users.

ReproZip-Web is fully open-source under BSD-3 license. It is publicly available on GitHub. We will make installers for users to be able to install the software with ease.

C.2 Identify where you will deposit the source code for the software you intend to develop:

Name of publicly accessible source code repository:

ReproZip-Web

URL:

https://github.com/reprozip-news-apps/reprozip-web

SECTION IV: RESEARCH DATA

As part of the federal government's commitment to increase access to federally funded research data, Section IV represents the Data Management Plan (DMP) for research proposals and should reflect data management, dissemination, and preservation best practices in the applicant's area of research appropriate to the data that the project will generate.

A.1 Identify the type(s) of data you plan to collect or generate, and the purpose or intended use(s) to which you expect them to be put. Describe the method(s) you will use, the proposed scope and scale, and the approximate dates or intervals at which you will collect or generate data.

We will be generating qualitative data as a part of our UX working sessions with our project partners. This will not be used in a research setting, but rather for direct improvements to the tools. We will not be using it for research purposes.

A.2 Does the proposed data collection or research activity require approval by any internal review panel or institutional review board (IRB)? If so, has the proposed research activity been approved? If not, what is your plan for securing approval?

N/A

A.3 Will you collect any sensitive information? This may include personally identifiable information (PII), confidential information (e.g., trade secrets), or proprietary information. If so, detail the specific steps you will take to protect the information while you prepare it for public release (e.g., anonymizing individual identifiers, data aggregation). If the data will not be released publicly, explain why the data cannot be shared due to the protection of privacy, confidentiality, security, intellectual property, and other rights or requirements.

N/A

A.4 What technical (hardware and/or software) requirements or dependencies would be necessary for understanding retrieving, displaying, processing, or otherwise reusing the data?

N/A

A.5 What documentation (e.g., consent agreements, data documentation, codebooks, metadata, and analytical and procedural information) will you capture or create along with the data? Where will the documentation be stored and in what format(s)? How will you permanently associate and manage the documentation with the data it describes to enable future reuse?

N/A

A.6 What is your plan for managing, disseminating, and preserving data after the completion of the award-funded project?

N/A

A.7 Identify where you will deposit the data:

Name of repository:

N/A

URL:

N/A

A.8 When and how frequently will you review this data management plan? How will the implementation be monitored?

N/A