NYU | LIBRARIES

Elmer Holmes Bobst Library
70 Washington Square South
New York, NY 10012-1091

Saving Data Journalism: A prototype to enable systematic collection & preservation of dynamic websites

From interactive visualizations to easy-to-use databases, storytelling with data has revolutionized modern reporting. Some famous examples of this genre include the Panama Papers, "The Color of Debt" by ProPublica, and "Gun Deaths in Your District" by *The Guardian*. These paradigm-changing investigative projects made issues accessible to the public, brought down corrupt government officials, and prompted international changes in transparency, security, and human rights.

New York University Libraries, in collaboration with the NYU Arthur L. Carter Journalism Institute, the NYU Center for Data Science, and ProPublica, request $48,766 for a planning grant to develop a prototype for a software tool to capture and preserve data journalism projects like these in a scholarly archive.

The prototype will add to an existing open source computational reproducibility tool, ReproZip.[1] The proposed extension will be designed to quickly and easily pack and unpack dynamic websites, enabling the first large-scale collection, preservation, and discovery of complex interactive websites.

We plan to focus initially on a single kind of data journalism artifact, a database-reliant journalism project known as a news application or "news app," which cannot be captured by any current web archiving tools.[2] Each news app has multiple components: a database, the graphical interface that appears in the browser, text, images, and other multimedia components.[3] For future scholars to see the same look, feel, and functionality, all of these components must be saved and reassembled in the same way. Web crawlers only capture snapshots of the "front end" of a news app, the graphically appealing part that is seen through a browser; ReproZip is a "back end" tool that only preserves the database and server side of a news app. We propose to create a prototype that brings together open source front end crawling with back end reproduction in a single open source software package.

National Impact

The production and popularity of news apps has exploded in the past five years, both through start-ups devoted to data journalism, such as FiveThirtyEight and Vox.com, and through legacy news publications like *The New York Times*, *The Washington Post*, and *The Wall Street Journal*. The reach of these projects speaks to their role in shaping our culture. Unfortunately, many terabytes of born-digital content will inevitably be lost to the black hole of technological obsolescence.[4] Our approach to solving this problem will allow a significant shift in libraries' strategies for archiving. Historically, the technique libraries use in archiving is migration: we migrate print newspapers to microform, and then microform newspapers to digitized PDF/A files; or 16mm film reels to VHS tapes, DVDs, and streaming MP4 files. This strategy has its limitations, but has been largely successful for physical materials and static digital objects like text or images. For dynamic digital objects, the amount of time, work, and organization involved in migrating or updating every software dependency is unfeasible. Digital archivists believe that to save these sites for the long term, we must emulate, not migrate, the object in its computational environment.[5]

The tool we plan to build will be an emulator for dynamic digital websites like news applications. It will produce a file type, a .rpz, which will be suitable for long-term preservation,[6] and years from now could be opened and unpacked to emulate a website produced today. The .rpz file that ReproZip produces will contain all of the files, programs, and data needed to recreate the website, as well as the precise version of the operating system, software libraries, and other software dependencies needed for the site to display. It will also create the necessary metadata to describe these objects in a discovery tool so that the public can find and access them in library catalogs, repositories, and Internet archives.

The impact of this project will extend well beyond news applications. Database-reliant journalism stories are but one example of the sophisticated digital projects that news organizations produce. Data visualizations need to be preserved too, as do newer forms like virtual reality storytelling projects. An emulation tool that could quickly and simply pack and unpack these projects, irrespective of the operating system or other dependencies, would be the first of its kind, and could enable the systematic collection of hundreds of thousands of dynamic digital objects internationally. In keeping with the focus and goals of the "Curating Collections" project category, the success of this project would enable new processes and workflows for the collection of this content on a mass scale.

Project Design

In the fall of 2016, we tested ReproZip as a capture and archiving tool in a pilot experiment with ProPublica's news app "Dollars for Docs." While ReproZip successfully captured all of the app's back end files, it was unable to capture remote front end files, and the archiving was incomplete. We plan to fund a software developer to add open-source web crawling functionality to ReproZip in order to capture and archive the front end of Dollars for Docs and other data journalism projects. A graduate assistant will work with ProPublica and our ReproZip developers to test the prototype, add cataloguing metadata, and plan for additional refinements.

Outcomes of this project will include an open-source prototype for an extension of ReproZip, scholarly and popular articles in which we share our findings, and presentations at three or more journalism, computational reproducibility, and Internet archiving conferences.

A rough outline of the proposed schedule, to begin May 1, 2018, is as follows:

Weeks 1-4: Evaluate web crawlers and proxy options

Weeks 5-16: Integrate crawler with ReproZip, add server component to ReproUnzip

Weeks 17-24: Test, iterate, validate with ProPublica; write up findings

Weeks 25-28: Release code and documentation; publish findings; create a plan to capture and archive more apps

Budget

Budget summary: The requested funds of $48,766 would cover 3.5 months of a developer's salary ($10,000/month for $35,000) to create a prototype, and 100 hours of work for a graduate assistant (10 hours/week, for 10 weeks @ $20/hour, for $2,000) to test and update the prototype. As part of his responsibilities as Associate Dean, Principal Investigator Michael Stoller will devote a portion of his time for administrative oversight. Indirect costs will total $11,766 at a negotiated rate of 31.8%.

[1] Fernando Chirigati et al., "ReproZip: Computational Reproducibility With Ease", SIGMOD, San Francisco, USA, 2016, 2085–88.

[2] Katherine Boss and Meredith Broussard, "Challenges of Archiving and Preserving Born-Digital News Applications," *IFLA Journal* 43, no. 2 (June 2017): 150–57.

[3] Meredith Broussard, "Preserving News Apps Present Huge Challenges," *Newspaper Research Journal* 36, no. 3 (September 1, 2015).

[4] Kathleen A. Hansen and Nora Paul, *Future-Proofing the News* (Lanham: Rowman & Littlefield, 2017).

[5] Dirk von Suchodoletz and Jeffrey van der Hoeven, "Emulation: From Digital Artefact to Remotely Rendered Environments," *International Journal of Digital Curation* 4, no. 3 (July 12, 2009): 146–55; Stewart Granger, "Emulation as a Digital Preservation Strategy," *D-Lib Magazine*, October 2000.

[6] Vicky Steeves, Rémi Rampin, and Fernando Chirigati, "Using ReproZip for Reproducibility and Library Services," *LIS Scholarship Archive*, 2017.