Social Media Preservation in Times of Change: A Workflow and Policy Analysis Case-Study at Towson University's Archives

EMIEL BROWN and JASMINE S. MALONE

Towson University

ABSTRACT

The archival profession is witnessing a concerted effort to preserve the web in the current political climate. The retirement of the Albert S. Cook Library's X account and move to Bluesky exposed an opportunity for Towson University's Special Collections and University Archives (SCUA) to reassess our social media preservation procedures. SCUA currently utilizes Archive-It as its web-crawling tool; however, it has been inconsistent with social media capture. We are investigating other tools for crawling, direct content download, or consistent capture as a preservation practice. Our case study seeks to evaluate how our archives can best gauge its capacity for social media preservation and how we determine "good" sustainable practices amidst platform evolution, retirement, and migration.

The TU Digital Archivist and Associate for Digital Initiatives will present their work-in-progress reassessment of the social media preservation workflow and policy development. Their presentation will cover 1) the political context surrounding web archives and our library's response 2) our process for demoing web capture products 3) proposal for a social media preservation policy that accounts for the accessibility implications of such media and its relevance for a diverse group of researchers.

INTRODUCTION

The Towson University Special Collection's and University Archives (SCUA) adopted Archive-It as its web capture tool in 2017. There was a determined need to archive the web as TU's web presence continued to expand and diversify. In the first subscription year (2017 – 2018), SCUA captured 427.1 gigabytes (GB) of webpage data. To date, SCUA has captured 2.6 TB of data across twenty-two (22) collections.

Our collections scope for Archive-It aligns with our general collecting policy and currently encompasses collections falling within the categories of TU webpages, student publications and organizations, TU alumni and retired faculty or staff, and TU-affiliated projects. The webpages included in the case study scope are the university's official website and webpages, TU WordPress sites created within the TU WordPress environment, Campus Labs Engage for student activities, Sidearm Sports for Athletics, and various social media sites.

CASE STUDY BACKGROUND

Albert S. Cook Library's Director of Library Communications & Engagement, who manages the library's social media pages, announced during the March Cook Library staff meeting that they posted a retirement message and were leaving the Twitter (X) platform, exploring Bluesky as an alternative. At the April staff meeting it was announced that the library opened a Bluesky account. The reason the library gives for leaving X is that their audience is no longer based on the platform, particularly Black students. The move to Bluesky is a means to reach this audience, and the political context around X leadership is a minor role.

Figure 1: Screenshot of X post: x.com/CookLibraryOfTU

Moving to Bluesky is related to what has been dubbed online as "the great X-odus," wherein folks are leaving the X platform for similar sites since X came under new leadership in 2022[1][2]. Additionally, the shift aligns with a major uptick in Bluesky membership since the November 2024 U.S. presidential election[3][4]. Reasons why users, including Black users, have cited switching platforms include:

[1] Tom Ambrose, "What Is Bluesky and Why Are so Many People Suddenly Leaving X for the Platform?" The Guardian, November 16, 2024. https://www.theguardian.com/technology/2024/nov/16/what-is-bluesky-and-why-are-so-many-people-suddenly-leaving-x-for-the-platform-elon-musk.
[2] Dani Di Placido, "The X (Twitter) Exodus to Bluesky, Explained." Forbes, November 19, 2024. https://www.forbes.com/sites/danidiplacido/2024/11/19/the-x-twitter-exodus-to-bluesky-explained/.
[3] Ambrose, "What Is Bluesky".
[4] Di Placido, "The X (Twitter) Exodus".

- New ownership and major policy changes including the subscription model, changes in blocking functions, and automatic opt-in to training the X artificial intelligence model[5]
- High quantities of bots, misinformation, disinformation, and bigotry[6][7]
- Choosing other platforms such as Bluesky for its federated server style[8]
- Mastodon, Instagram Threads, and Black-owned site Spill[9][10]

This case study exists within this context, and that context is not neutral. It is nuanced, political, and resource dependent. The Library Associate for Digital Initiatives and the Archivist for Digital Collections and Research saw an opportunity to preserve the Library's X presence while testing alternatives to the current workflow. The case study sought to evaluate how SCUA can best gauge its capacity for social media preservation and how "good" sustainable practices are determined amidst platform evolution, retirement, and migration.

METHODS

SCUA captures online content generated both by the university and by a variety of on- and off-campus groups. This includes official web pages as well as social media accounts on various platforms, some of which are no longer in use. The ideal tool for web capture is one that regularly and automatically crawls a substantial list of both static web pages and dynamic social media feeds, including Facebook, X, Instagram, TikTok, and Bluesky, and is open to developing ways to crawl new social media platforms as they rise in popularity. The goal is to find something able to faithfully capture, store, and facilitate access to these pages, particularly those that are only accessible to site users with login credentials, but does not require staff to have expert knowledge in the details of the capture process.

The following criterion were used to evaluate selected tools for the case study:

- Crawling / capture process – How does it work? What can it do?
- Expertise needed – What knowledge is necessary? What support is available?
- Type and capacity of storage – Where do captured files go? How are they stored?
- Interaction with login-protected sites – Can it capture things like X and Instagram?
- Replay of captured sites – How can researchers use these resources in the future?
- Cost – What do we pay for? How much? How often?

[5] Char Adams and J.J. McCorvey, "Black Twitter Helped Define the Internet. Where Will the Exodus from Musk-Led X Lead?" NBC News, November 17, 2024. https://www.nbcnews.com/news/nbcblk/black-twitter-x-elon-musk-exodus-bluesky-rcna180147.
[6] Adams and McCorvey, "Black Twitter".
[7] Janice G. Asare "Black Exodus: Why X Users Are Deactivating and Building New Digital Communities." Forbes, November 23, 2024. https://www.forbes.com/sites/janicegassam/2024/11/23/black-exodus-why-x-users-are-deactivating-and-building-new-digital-communities/.
[8] Di Placido, "The X (Twitter) Exodus".
[9] Adams and McCorvey, "Black Twitter".
[10] Di Placido, "The X (Twitter) Exodus".

FINDINGS

The case study team compiled a list of 15 social media archiving tools from sources referencing their use. They then narrowed that list to tools that were cost effective, moderately easy to use, and not discontinued. The tools fully reviewed for this project were Archive-It (the archives current tool), Browsertrix, Archive Webpage, CivicPlus, and Pagefreezer. The table below provides a high-level comparison of their features and functionalities related to preserving social media pages.

Table 1: Tool Comparison

Tool Name	**Capture Process**	**Expertise Needed**	**Storage Specs.**	**Account Access and APIs**	**Replay**	**Price**
Archive-It	Manual & Scheduled web crawls	Moderate (for advanced settings)	Wayback Machine; WARC files	Social media account login per platform required	Wayback Machine; Client Interface, downloaded WARCs	Tiered by data volume
Browsertrix	Manual & Scheduled web crawls	Minimal	Tiered cloud storage; self-hosted storage; WACZ files	Social media account login per platform required	Replay Webpage; can be embedded	Tiered by data volume and crawl time
Archive Webpage	Manual web crawls – requires page to be scrolled	Minimal	Local storage; WACZ and/or WARC files	Social media account login per platform required	Replay Webpage; can be embedded	Free (open source)
CivicPlus	Automatic crawls (includes retroactive)	Minimal	AWS cloud storage, XML, HTML, PDF exports	Connects to APIs as available; Social media account login account per platform required	Client Interface	Tiered by monthly data volume
Pagefreezer	Automatic crawls (includes retroactive)	Moderate	AWS cloud storage	Connects to APIs as available; Social media account login account per platform required	Client Interface	Priced by # of social media accounts

The tools are compared in more detail below.

Effective Capture and Replay:

Archive-it has worked well in the past, but frequent changes in social media mean that it must continually update and frequently has replay issues with dynamic pages. Browsertrix follows a similar capture format; however, the capture is faster, and replay is nearly immediate. Though its free version, Archive Webpage, is able to accurately capture long and dynamic social media feeds and plays back accurately with the Replay Webpage tool; it sometimes crashes when interacting with extremely long profiles. CivicPlus and Pagefreezer provide the most detailed capture, crawling in near-real time and allowing viewers to access and search multiple versions of deleted or edited content.

Scheduled Crawling Capability:

Archive Webpage is unable to scale to meet more extensive needs in terms of capturing large numbers of pages or those that must be captured regularly. Conversely, the near-continuous updating provided by CivicPlus and Pagefreezer may go beyond many users' needs. Both Archive-It and Browsertrix provide a variety of frequency options for scheduling repeat crawls.

Cost:

CivicPlus and Pagefreezer are priced by volume of records and number of social media accounts, therefore they are not cost-effective for those planning to preserve a large number of accounts. For a small-scale project, Archive Webpage is a useful free solution. For large or ongoing projects, both Archive-It and Browsertrix offer a spectrum of subscription pricing based on the scale of what an institution plans to capture.

Interaction with Login-Protected Sites:

The Archive Webpage browser extension is able to effectively capture pages protected behind both logins and two-factor authentication by allowing someone with site user credentials to log in fully before starting the capture (starting it before logging in will result in the autopilot function being interrupted). While the browser extension can be activated after logging in, the desktop app is unable to effectively bypass login popups.

Archive-it and Browsertrix both allow users to save login information to be used with selected crawls – Archive-It associates seed URLs with relevant credentials, and Browsertrix allows the creation of "browser profiles" that can be applied to specific crawls.

Rather than using login credentials to access social media pages from a "viewer" perspective, CivicPlus and Pagefreezer capture data from "inside" individual social media accounts. This allows greater detail to be captured but requires individual login information for each account to access them.

Direct download of data from X:

This option requires both login credentials and access to an associated e-mail account for confirmation. An institution may be able to download data for their departmental account but would likely have to collaborate closely with many other account holders to use this method for general capture. For accounts that are abandoned or no longer regularly in use, this information may be completely inaccessible. It also does not provide an easy way for researchers to interact with downloaded data as it would originally have appeared online. However, this might still be a useful method for institutions preserving the contents of their own X accounts.

RECOMMENDATIONS

Preserving individual pages

Archive Webpage allows logged-in, one-time capture of social media pages via free browser extension and local storage of data. It works well with dynamic and "endless scroll" pages, and can be used to capture pages that are behind a login barrier, provided the user activates the browser extension after fully logging in. For institutions that are interested in capturing a small number of pages and storing them locally or in their own cloud storage, Archive Webpage is a good tool for small-scale capture that doesn't need to be repeated regularly.

Managing institutional social media accounts

CivicPlus and Pagefreezer allow near-real-time management of accounts with individual login information. These tools are well suited for institutions with a strong social media presence and an interest in capturing viewer and commenter interactions as well as detailed changes to these pages. However, both products require user login credentials to capture these materials, making them impractical for those who need to capture content from a variety of different creators, especially if that content includes abandoned or defunct accounts.

Ongoing dynamic crawls

Archive-It and Browsertrix allow individual or scheduled crawls of sites and store user-captured data on a subscription basis. Each allows the user to choose how often these crawls occur. Archive-It prevents duplication of data that has already been captured, a feature that is not currently possible through Browsertrix. Pagefreezer and CivicPlus allow nearly continuous capture – this is excessive in most cases but could be helpful for institutions where social media is in a constant state of rapid change.

Policy

A web archives collection management policy is needed to address accession, appraisal, or reappraisal and deaccession, and to ensure that the sites being preserved align with an institution's collecting areas, are of enduring value, and, if needed, can be removed from the collection to free up digital space for something deemed more appropriate.

For an institution with a well-established and efficient web capture workflow, the initial process of confirming that a page or site is a suitable addition to a collection is often the most complex aspect of this workflow. By developing a detailed and thoughtful policy for the accession and potential deaccession of web materials, this process can be simplified, reducing the time spent on case-by-case evaluation of potential captures. A broad goal is to develop a policy that reflects the pathway to the collection an institution would like to build and provides a series of guidelines to aid staff in maintaining that trajectory.

Procedures

Similarly, documenting and regularly revising workflows ensures consistency between captures and can help communicate errors if a workflow is not consistently successful. Establishing and documenting a consistent workflow will allow staff to be efficient in testing and integrating new tools and processes as they become available and will illuminate what training and professional development are necessary for efficient use of this workflow.

Professional Development

Offering continued professional development opportunities for archives staff related to web archives expands the web capture potential of the archives, allowing staff to broaden their experience with available tools and deepen their knowledge of the principles of web capture.

The priorities in this area are to 1) familiarize staff with new tools, provide opportunities to answer questions about current or potential resources, familiarize staff with collection policies and plans, and allow staff to interact with other groups and institutions to discuss developments around web capture and digital preservation.

CONCLUSION

Following the evaluation of available tools, an institution must choose an ideal tool, integrate that tool within workflows, update policies to include web archives, and, most importantly, anticipate future changes and remain flexible in capturing the ever-changing landscape of the web.

SCUA's next steps are to use this data to select a tool the fits within the scope and financial confines of the archives, then integrate the workflows while learning a new tool if a new tool is selected. Policies must be developed to standardize the collection of social media sites affiliated with TU which must address how to reevaluate the issues with migration and obsolescence of the web and social media, particularly.

Acknowledgements

Allison Fischbach – Digital Archivist at Alan Mason Chesney Medical Archives, Johns Hopkins University and Medicine

Ashley Todd-Diaz – Director of Distinctive Collections and Digital Scholarship at Towson University

Joyce Garczynski – Director of Library Communications & Engagement at Towson University

SAA Research Forum 2025

Bibliography

1. Adams, Char, and J.J McCorvey. "Black Twitter Helped Define the Internet. Where Will the Exodus from Musk-Led X Lead?" NBC News, November 17, 2024. https://www.nbcnews.com/news/nbcblk/black-twitter-x-elon-musk-exodus-bluesky-rcna180147.
2. Ambrose, Tom. "What Is Bluesky and Why Are so Many People Suddenly Leaving X for the Platform?" The Guardian. The Guardian, November 16, 2024. https://www.theguardian.com/technology/2024/nov/16/what-is-bluesky-and-why-are-so-many-people-suddenly-leaving-x-for-the-platform-elon-musk.
3. Asare, Janice G. "Black Exodus: Why X Users Are Deactivating and Building New Digital Communities." *Forbes*, November 23, 2024. https://www.forbes.com/sites/janicegassam/2024/11/23/black-exodus-why-x-users-are-deactivating-and-building-new-digital-communities/.
4. Di Placido, Dani. "The X (Twitter) Exodus to Bluesky, Explained." *Forbes*, November 19, 2024. https://www.forbes.com/sites/danidiplacido/2024/11/19/the-x-twitter-exodus-to-bluesky-explained/.
5. GitHub. "What Is an API?" GitHub, October 15, 2025. https://github.com/resources/articles/what-is-an-api.