Web Archiving Local Election and Government Websites

Julia Ezzo, Ed Busch, Elisa Landaverde, Jessica Martin, Lydia Tang

Local election and political websites are highly ephemeral due to their nature, especially for losing candidates. Thus, they are highly vulnerable to loss from the historical record. A survey during Spring 2019 of previously captured web archives on the Archive.org website showed a scarcity of captured websites for local governmental and political elections in Michigan. The University of Michigan's Bentley Historical Library currently captures websites related to the Governor, Michigan senators, and some congress members. However, less high-profile candidates were not being captured. As such, many websites from the 2018 midterm elections are vulnerable to loss. Furthermore, 2020 held a presidential election as well as many local elections, and with these campaigns, political websites hosting valuable candidate information were put up on the web for a limited time. Preserving content from these websites could be of great value for future researchers.

This project, funded by a microgrant provided by the Michigan State University (MSU) Libraries, explored the capture and curation of local political and election websites using an existing Archive-It (https://archive-it.org/) subscription. For this project, we had a commitment to both internal and external collaboration. We collaborated in the early site identification with the Library of Michigan staff to coordinate collection development. Internally, we also hosted a focus group consisting of faculty, graduate students, and local government administrators to help define our target areas.

The Internet Archive defines web archiving as "a series of steps that work together for an end goal: to interact with a website as it looked on the day that it was archived."[1] The process of web archiving can be a time-consuming endeavor depending on the complexity of the website, such as the inclusion of embedded content like PDFs and videos, as well as other factors such as if the site was built using website development services such as Wix, which uses customized themes and captures are often incomplete or embedded content does not replay. Typically, web archiving is conducted by librarians and archivists. However, as the nature of web archiving is time-consuming, in order to alleviate some of this strain on professional staff we explored the use of a non-librarian/archivist to conduct and verify the technical aspects of web capturing under staff supervision.

Focus Group Session

An important aspect of our project design was collaboration, both outside the MSU Libraries through our partnership with the Library of Michigan, and internally through collaborative collection development and a facilitated focus group. Rather than solely choosing the websites to be archived ourselves, we desired to bring together stakeholders from various communities to help guide our collection development strategy. Stakeholder participation helped to confirm that the sites identified by members of the microgrant team had research value, though some sites were excluded due to a variety of technical issues discussed later in this article. This proved to be a fruitful method, which generated a productive conversation and useful feedback for both this pilot phase and future web archiving initiatives.

The focus group was held on September 12, 2019 at the MSU Main Library. Email invitations were sent to 29 stakeholders, including faculty and graduate students from Political Science and History, staff from the MSU Libraries and Library of Michigan, and stakeholders from local government. Box lunches were provided to the participants as an incentive for participating in the focus group.

In addition to the five members of the microgrant team, nine stakeholders were able to attend the focus group, including five faculty members (three from Political Science and two from

History), one participant from local government, two representatives from the Library of Michigan, and one stakeholder from MSU Libraries (not inclusive of the microgrant team). Every member of this group was actively engaged in conversation and the activity, described below.

After a brief introduction to the project and web archiving, the group engaged in a conversation facilitated by Jessica Martin, addressing the following questions:

1. What are your current research and/or future research plans?
2. How might you use political websites in your research?
3. How might you use political websites in your teaching?
4. What kind of content matters now/in the future? Campaign promises? Policy platforms? Propositions?
5. What is most important to you: what politicians are saying or what politics are like in 2019? Does format matter?
6. What are the value of social media crawls to your research/teaching?
7. Are there other information streams you'd like us to capture?
8. Is it important to capture commentary about the politicians in addition to their campaign programs? (ethics? copyright?)
9. What would be the easiest way for you to access this material?

From this conversation, we took notes on key points raised by the group. Some salient points include:

- Information on ballot initiatives and millages is important and difficult to find at the local level.
- There is interest in the websites of both the candidates (demographics, personal history, policies) and policy advocacy groups.
- Social media plays an important role in political communication today, and capturing that information should play a role in collection development.
- Commentary also matters, in social media and in local media, as this is where citizens communicate their views as well.
- Digging deeply into websites is important, as meeting minutes and other documentation are often found as attached PDFs.
- There are courses at MSU that would directly benefit from access to this content, including the Master's in Public Policy capstone class, which is particularly interested in local data from the Detroit area.
- Participants were familiar with and happy to use an interface like the Wayback Machine, with additional requests made for value added services like text mining, mapping, etc.

Following the group conversation, lasting about 45 minutes, the group participated in an activity to brainstorm collection development priorities for the project. We began by placing sticky notes of the sites we already planned to crawl on a board, giving sticky notes and markers to the group to add notes related to specific sites, or types of sites, they believed we should consider. From this, we gathered well over one hundred suggestions, which we then organized into categories. These categories included Advocacy Organizations, General Election Information, Government Information and Sites, Ingham County, Media, Meetings, Proposals, and State. The collated information was then captured through photographs and added to a shared folder for later analysis.

Site Identification

Prior to the focus group, grant team member and subject matter expert Julia Ezzo developed a list of possible websites by utilizing 2018 and 2019 ballot information from the Ingham County clerk. Candidate and proposition names were searched on Google and Facebook to determine if there was a web presence for the campaigns. These URLs were presented as capturing options at the focus group, and participants were able to rank their importance, as well as suggest other potential sites. Participant-recommended sites were searched and if a URL was available, these sites were added to a spreadsheet for future capturing. Sites were then searched within the Internet Archive's Wayback Machine to determine if and how frequently these websites may have been archived to allow the grant team to prioritize those sites that were either never captured, were captured only a few times, or were missing content due to the limitations of the web archiving tool, such as difficulty preserving embedded content such as PDFs on webpages.

Additionally, based on focus group feedback, the scope of this project was adjusted from narrowly focusing on election candidates in Ingham County to also include the websites for the mayors of Detroit and Flint, as well as local government meeting sites (such as city councils and board of trustees) within Ingham County.

Crawling

Crawling is a process of activating web crawlers or robots (software that identifies and captures web content) and telling these crawlers what site to crawl (capture) and how frequently to perform this task.[2] Within the Archive-It subscription for MSU Libraries (MSUL), a Michigan Politics Web Crawling Project collection was created. A total of 110 sites were entered into the collection, but only 96 were considered active at the time this microgrant ended. Crawls were executed primarily between October and December 2019. Refined crawls were executed in January, February, and March 2020 to fill in missing content from the original crawls due to technical scoping issues. A total of 65.8 GB were captured during this initial stage of the microgrant.

The identified sites were added to this collection and assigned to eight Archive-It groups. These were:

- County Commissioners
- City Council
- Advocacy
- Judges
- Mayors
- School Board
- Propositions
- Meetings

Archive-It provides several frequency options for scheduling crawls. For this project, we selected a "One-Time Crawl" frequency in order to evaluate the quality of the crawl before deciding if further adjustments in the crawl's scope were necessary or if it could be saved after one attempt. Archive-It has two types crawls, One-Time crawls and Test crawls. The major difference between the two is that Test crawls can be deleted if deemed unsatisfactory. All crawls were run as type Test Crawl and used the Brozzler experimental Crawling Technology, which provides enhanced media capture capabilities and mimics how a human user interacts and experiences web browsing.[3] Data limits and Time limits were set in accordance with Archive-It's recommendations.

Metadata

Metadata for the collection was added at two levels, the collection level and the seed level. Seeds are URLs "with a unique identifier in the Archive-It backend."[4] Seeds can be the URL to the full website, the URL to a specific part of a website, or the URL to a specific document on the website. The Archive-It application supports Dublin Core (https://dublincore.org/) as a schema, providing 15 standard elements as well as custom fields. The application contains no required fields, however, we devised a set of recommended fields based on metadata practices previously established for collections archived by the MSU University Archives and Special Collections. MSUL Special Collections had previously developed an internal manual for creating metadata that included recommended fields as well as recommendations for the use of Library of Congress Subject Headings and other formatting specifications. At the collection level, the elements selected included title, subject, description, rights, collector, language, and genre. These elements reflected an overall introduction to the collection, along with broad subject headings.

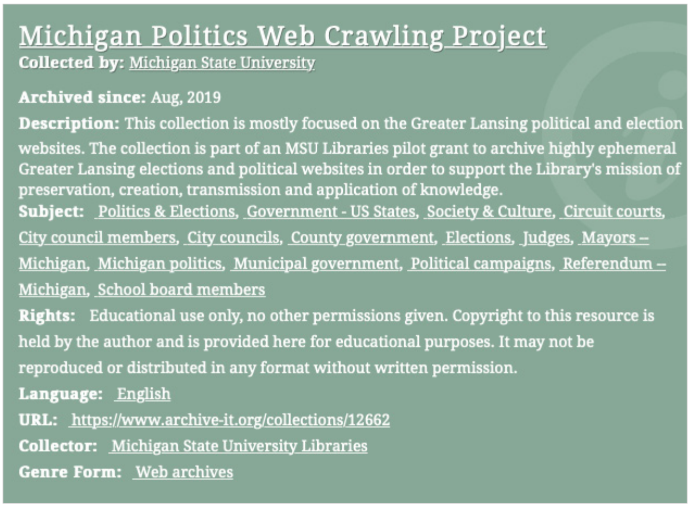
At the seed level, elements included title, subject, description, language, coverage, format, rights, collector, contributor, URL, and when available, creator. URLs were added as a custom field in order to preserve the original URL used to archive the website. Titles were often taken verbatim from websites, but on occasion they were adapted for consistency. An example of one of the seeds is provided in figure 2. We used Library of Congress Subject Headings, shown in figure 3, and made the geographical coverage in figure 4 as specific as possible to maximize the use of the facet feature shaped by the captured metadata.

Training

The graduate student assistant (Daniel Fandino) for this microgrant was provided with a login for MSU's Archive-It account, the MSU Archive-It and metadata manuals, as well as information on how to access online training materials provided by Archive-It. After a review of these materials, we did a brief walkthrough with the graduate assistant in order to make him more comfortable with the tool.

The team used a spreadsheet crawl log for manually documenting work accomplished, information regarding crawls, and what types of information to record. During the crawling phase of the microgrant, we modified the log to suit our needs. The final columns headers for the log were:

- Category
- Number of sites
- Crawl ID#
- Crawled?
- Type of crawl
- Sites crawled
- Facebook sites?
- Data Limit
- Time Limit
- Technology
- Date
- Notes
- Facebook added
- FB ID#



Michigan Politics Web Crawling Project
Collected by: Michigan State University

Archived since: Aug, 2019

Description: This collection is mostly focused on the Greater Lansing political and election websites. The collection is part of an MSU Libraries pilot grant to archive highly ephemeral Greater Lansing elections and political websites in order to support the Library's mission of preservation, creation, transmission and application of knowledge.

Subject: Politics & Elections, Government - US States, Society & Culture, Circuit courts, City council members, City councils, County government, Elections, Judges, Mayors -- Michigan, Michigan politics, Municipal government, Political campaigns, Referendum -- Michigan, School board members

Rights: Educational use only, no other permissions given. Copyright to this resource is held by the author and is provided here for educational purposes. It may not be reproduced or distributed in any format without written permission.

Language: English

URL: https://www.archive-it.org/collections/12662

Collector: Michigan State University Libraries

Genre Form: Web archives

Figure 1. Collection Example

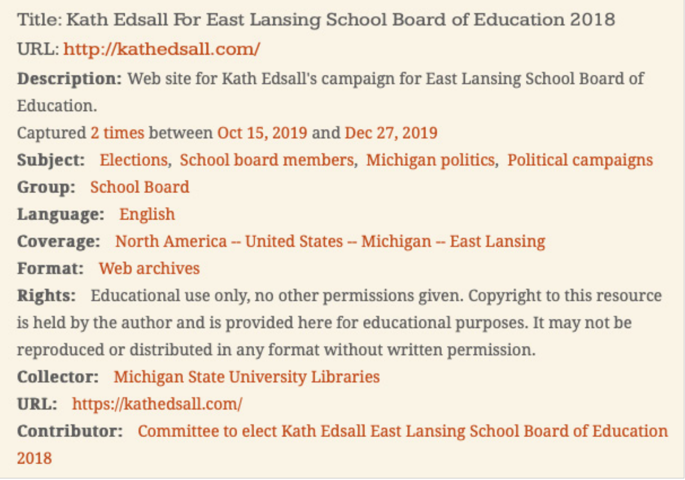
- FB sites
- Scope
- Total sites crawled
- QA
- Complete
- Date complete
- Status
- Notes

Once the seed list was finalized, the graduate assistant was oriented to the basics of entering metadata for collections and seeds and how to execute crawls. As crawls completed their runs, the graduate assistant was trained on reviewing crawls and how to troubleshoot challenging scenarios such as missing images and videos, or page formatting.

Archive-It requires a significant amount of training time, advising, and experience to acquire a good working basis for the full scope of work (seed selection, metadata, crawling, and quality assurance). As part of a continued project, it may be worth the time to provide advanced training opportunities in the use of "regular expressions" to help refine scoping. Regular Expression (also known as reggex) is a sequence of characters that specifies a search pattern.

Quality Assurance and Scoping

Quality Assurance (QA) for web crawling is not a simple task and requires significant time and manual effort. Decisions need to be made on how extensive (follow every link) or what percentage of a website to check. The types of seeds crawled for this microgrant were, for the most part, not overly complex and most site links could be visually verified with just a few clicks.



Title: Kath Edsall For East Lansing School Board of Education 2018

URL: http://kathedsall.com/

Description: Web site for Kath Edsall's campaign for East Lansing School Board of Education.

Captured 2 times between Oct 15, 2019 and Dec 27, 2019

Subject: Elections, School board members, Michigan politics, Political campaigns

Group: School Board

Language: English

Coverage: North America -- United States -- Michigan -- East Lansing

Format: Web archives

Rights: Educational use only, no other permissions given. Copyright to this resource is held by the author and is provided here for educational purposes. It may not be reproduced or distributed in any format without written permission.

Collector: Michigan State University Libraries

URL: https://kathedsall.com/

Contributor: Committee to elect Kath Edsall East Lansing School Board of Education 2018

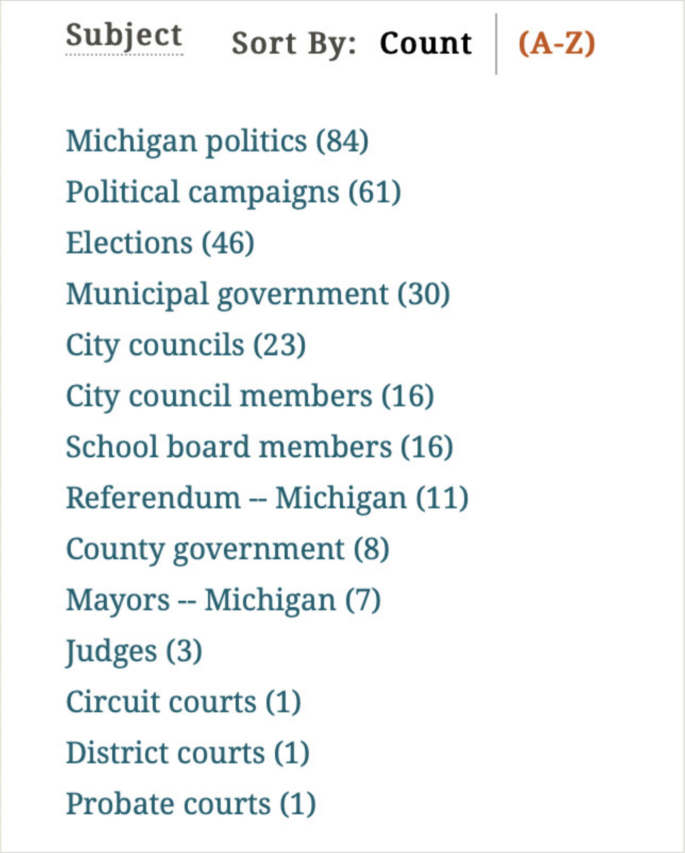
Figure 2. Example of what a seed looks like, including descriptive elements related to its capture

Archive-It provides some high-level help on the QA process (via their Help Center) and a Wayback QA Tool. Once a crawl is saved, the tool scans your currently viewed web crawled page and identifies links not captured due to blocks or scope. This information can then be used to run a "patch crawl" but is also useful in identifying scoping issues such as missing images. More information about the QA process we followed can be found at the University of Michigan's website at https://sites.google.com/a/umich.edu/bhl-archival-curation/web-archiving/05-quality-assurance.

After an initial crawl, several seeds, such as mcrgo.org, progressmichigan.org, aclumich.org, mcfn.org, were noted as having good crawls already completed by the Internet Archive (https://archive.org/web/) and were removed from this microgrant work by marking them as inactive in the Archive-It administrative user interface. These seeds can be activated later for future crawls if necessary.

During QA playback of captured crawls, some web pages would come up blank. Playback is a tool that allows to see if a captured URL (seed) is displaying correctly or as close to the original page as possible; it's links are traversed (played back) by following them to see if captured.) A review of the page's source code often revealed this to be due to its web host and technology. A number of these sites were created with the Wix cloud-based website builder. Unfortunately, this builder notoriously creates websites with "crawler traps" (endless invalid URL links) and playback issues. Some of these issues can be alleviated with various scoping techniques such as:

- entering explicit seeds for a site's menu pages or a needed external URL,



Subject Sort By: Count | (A-Z)

Michigan politics (84)
Political campaigns (61)
Elections (46)
Municipal government (30)
City councils (23)
City council members (16)
School board members (16)
Referendum -- Michigan (11)
County government (8)
Mayors -- Michigan (7)
Judges (3)
Circuit courts (1)
District courts (1)
Probate courts (1)

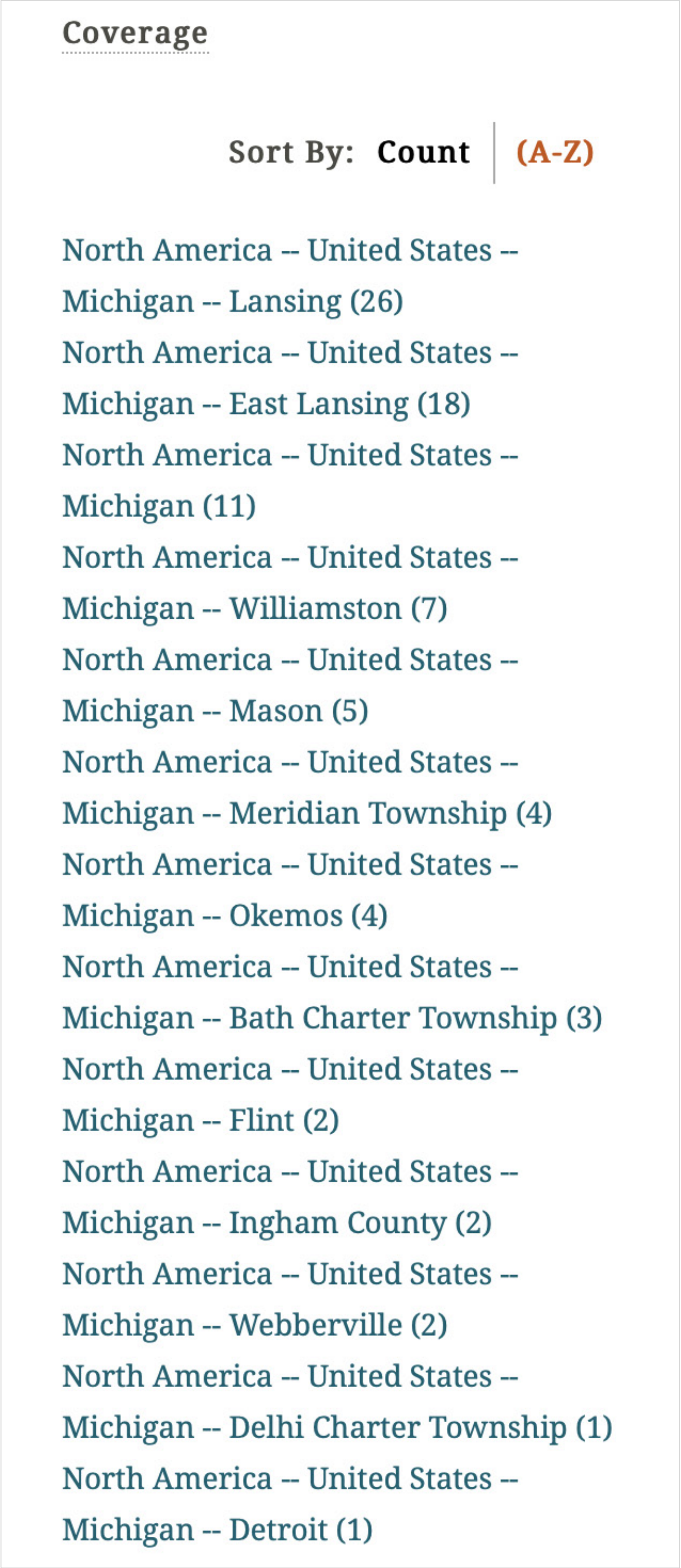
Figure 3. Library of Congress Subject Headings used for this project

- using the Brozzler crawler,
- testing crawls using different browsers during playback,
- or simply reloading a page in the browser.

Examples of added seeds include the URL for the About page or Publication's page of a website.

Another consideration is the fact that social media sites change with great frequency and the quality of captured Facebook sites can vary considerably. Archive-It provides some default scoping rules for platforms like Facebook but, in some cases, the sites just don't playback well even with added scoping.

An area that was not addressed during the active crawling period was a review of the public collection page (https://archive-it.org/collections/12662) and correcting issues here such as multiple seeds for the same page and bad links. A final review indicated that some crawl scoping was done at the collection level and might be better accomplished at the seed level. This will need to be reviewed when the project goes forward.



Coverage

Sort By: Count | (A-Z)

North America -- United States -- Michigan -- Lansing (26)
North America -- United States -- Michigan -- East Lansing (18)
North America -- United States -- Michigan (11)
North America -- United States -- Michigan -- Williamston (7)
North America -- United States -- Michigan -- Mason (5)
North America -- United States -- Michigan -- Meridian Township (4)
North America -- United States -- Michigan -- Okemos (4)
North America -- United States -- Michigan -- Bath Charter Township (3)
North America -- United States -- Michigan -- Flint (2)
North America -- United States -- Michigan -- Ingham County (2)
North America -- United States -- Michigan -- Webberville (2)
North America -- United States -- Michigan -- Delhi Charter Township (1)
North America -- United States -- Michigan -- Detroit (1)

Figure 4. Geographic headings used to allow for more precise faceting by location

Access

While access to the archived websites is available through Archive-It (https://archive-it.org/collections/12662), the organization of the content and ease of understanding the scope of the collection could be challenging to some users. To create an easier to use access point a LibGuide (https://libguides.lib.msu

.edu/MI-Politics-Archive) was created. This LibGuide provides a brief overview of the project, along with tabs for City Council, County Commissioner, Judges, Mayors, Meetings, Proposals, and School Board. Each tab then includes category boxes, such as election year and geographic delineation, with each box containing the name of the candidate or topic with a direct link to the Archive-It landing page to the web captures. Since some websites required multiple captures, taking users to an intermediate page to see all crawl dates available made the most sense.

In April 2020, links to the LibGuide and to the Archive-It page were shared via email with the focus group participants and they were asked to provide feedback. The responses we received were positive and supportive, with no suggestions for improvement provided. The LibGuide was set up to include contact information to allow any researcher to submit questions, concerns, or other comments to one of the team members.

2020 Election

The goals stated in our grant proposal were to: explore the capture and curation of local political and election websites, commit to both internal and external collaboration, host a focus group meeting, and create a LibGuide to access the archived sites. Having met the goals of the initial pilot, the team decided to expand web archiving from the grant-funded scope of the 2018/2019 election cycle and work to capture the sites for the 2020 election. However, unlike the pilot project, the team did not have funding to hire a graduate assistant due in part to the COVID-19 pandemic. The subject librarian identified possible websites; crawls and QA was done by the project librarians and the graduate student and metadata was added by one of the archivists. Additional categories were added which include Register of Deeds, County Clerk, Drain Commissioner, Prosecuting Attorney, Sheriff, Williamstown Supervisor, Treasurer, and Trustee. In all, 97 websites were crawled at least once for the 2020 election cycle.

Challenges

During the pilot aspect of this project, the graduate student assistant was not hired until October 2019 and some sites had already gone offline, further underscoring the transitory nature of these websites. Due to other commitments, the graduate student assistant's time to devote to this project was limited, which made it challenging to balance capturing as many sites as possible and ensuring the captured sites and their content were successfully archived—which proved to be a particularly time-consuming task for sites with a lot of media items and complex sites with a lot of subpages and embedded content.

City government sites, in particular, proved to be the most complicated to crawl and required more in-depth QA since they tend to contain a large variety of content types, such as video, PDF documents, images, and other materials, and were thus more time-intensive to capture than the election-related websites. Due to the complex nature of these sites, they also used more of the Archive-It data allotment.

While candidate election websites were generally more straightforward in terms of crawling and QA, certain platforms added additional complications. Sites built using Wix.com, for example could be crawled successfully, but the content would not play back successfully. Crawling Facebook pages also proved to be problematic, as it would often take multiple crawl attempts, with each attempt producing different results. The web browser used when crawling and doing QA does make a difference, with Google Chrome being the better choice for these tasks.

One area outside of our control, but a deficit worth noting, is that in these smaller, local elections, not all candidates have a web presence for their campaign. If a researcher uses the archived websites that we created to try to get a full picture of the campaign, there are many candidates who are not represented and for whom we were unable to gather digital information regarding their platforms, policies, and even their candidacy. For full candidacy information, researchers still need to reference the election information archived through the county clerk.

Recommendations

Although we found that it is possible to conduct this sort of web archiving project using available staff, an ideal scenario would be to hire a digital archivist, who would be willing to assume primary responsibility for web archiving activities. However, understanding budget constraints an alternative solution is to dedicate staff members for web crawling. This could include creating one or more web crawling assistantship positions or possibly centralizing web crawling into a staff or librarian position(s). Using dedicated workers would free up staff for other duties and also ensure that we fully utilize our annual allotted storage that we pay for in Archive-It subscription fees. This would allow archivists and librarians to focus on establishing thoughtful and ambitious web archiving goals and further collection development and seed selection. Consideration should be given to creating a committee comprised of archivists, subject librarians, and other stakeholders to recommend future web archiving initiatives and select the seeds to crawl. A committee could establish clear collection development guidelines for future local election archiving, including assessing the frequency of crawling websites to capture changes over time.

Periodic review and reconciliation of existing Subject terms and other metadata used by other collecting institutions on Archive-it Home (the central Archive-It public interface portal) to enhance discoverability of collections should be considered as part of a web archiving workflow. For example, we used the subject "Elections," while other institutions have used "Local elections." The State Elections Web Archive (https://archive-it.org/collections/10793) created by the Ivy Plus Libraries Confederation could be a good model to follow.

Since crawling Facebook was especially problematic, exploring niche tools for archiving social media such as ArchiveSocial, Social Feed Manager, and other tools should be investigated. There are many technical and ethical implications for capturing social media, and these tools are able to preserve social media better than Archive-It. Archiving social media accounts is especially important with local elections, because many candidates might only create a Facebook page as opposed to an actual website, but one should consider whether a candidate's Facebook account is dedicated to their campaign or if they are using their own personal Facebook page and be considerate of the person's privacy if the latter is used.

Further thought should be put into the use of an Archive-It subscription for capturing a website's extensive video versus using an offline capture tool (such as youtube-dlG) and preserving using other in-house workflows. Using the latter approach would extend the subscription allotment for Archive-It to crawl websites, with the drawback that researchers would need to access downloaded video using multiple tools and portals.

Conclusion

Through this web archiving project, 187 domains were crawled and preserved on Archive-It. Following the 2020 election, team members intended to do an end of election crawl after the results of November 3, however it was discovered that many of the sites had already been taken down, thus emphasizing the transient nature of this content. While this type of web archiving project will not be able to completely fill in the historical record since not all candidates choose to have a web presence, it does fill the gap and will allow future researchers to analyze aspects of local political platforms and discourse between the public and candidates on Facebook. Local politics may not be as scintillating as state and national politics, but it is impactful nonetheless and worth considering capturing for the future.

Julia Ezzo (julia@msu.edu), Government Information, Packaging, and Political Science Librarian, Michigan State University; **Ed Busch** (buschedw@msu.edu), Electronic Records Archivist, Michigan State University; **Elisa Landaverde** (elandav@msu.edu), Special Collections LGBTQ+ Librarian, Michigan State University; **Jessica Martin** (achberg2@msu.edu), African Studies Librarian and Adjunct Assistant Professor History, Michigan State University; and **Lydia Tang** (ltang5@msu.edu), Special Collections Archivist, Michigan State University.

References

1. Jillian Lohndorf, "What is Web Archiving?," https://support.archive-it.org/hc/en-us/articles/360041674111-What-is-web-archiving-.
2. Jillian Lohndorf, "Archive-It Crawling Technology," https://support.archive-it.org/hc/en-us/articles/115001081186-Archive-It-Crawling-Technology.
3. Sylvie Rollason-Cass, "What is Brozzler?," https://support.archive-it.org/hc/en-us/articles/360000343186-What-is-Brozzler-.
4. Jillian Lohndorf, "Select Seed URLs," https://support.archive-it.org/hc/en-us/articles/208331753-Select-Seed-URLs.