

Web Archiving Local Election and Government Websites Microgrant

Final Report

17 April 2020

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

1 Introduction	2
2. Focus Group Session	2
3. Seed Identification	4
4. Crawling	4
4.1 Metadata	4
4.2 Training	6
4.3 Quality Assurance and Scoping	7
5. Access	7
6. Budget and Expenses	8
7. Issues	8
8. Recommendations	9
Appendices	10
Appendix A: Archive-It Training Manual	10
Appendix B: Seed List	10
Appendix C: Metadata Manual	10
Appendix D: Seed Crawl Logs	10
Appendix E: Presentation Slides	10

1 Introduction

Local election and political websites are highly ephemeral due to their nature, especially for losing candidates. Thus, they are highly vulnerable to loss to the historical record.

A survey during Spring 2019 of current captured web crawls on the Archive-it.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 elections were not being captured. Many such websites from the 2018 midterm elections are vulnerable to loss. Furthermore, in 2020 a presidential election will occur as well as many local elections, and with these campaigns, political websites hosting valuable candidate information will be put up on the web for a limited time and it is important that these sites are captured for future researchers.

This microgrant explored the capture and curation of local political and election websites using an existing [Archive-It](#) subscription. For this project, we had a commitment to both internal and external collaboration. We collaborated in the early seed defining period with the [Library of Michigan](#) staff to coordinate collection development. Internally, we also hosted a focus group consisting of faculty, graduate students, and local governmental administrators to help us define our target areas.

Web crawling can be a very time consuming process. In this microgrant, we also explored the use of a non-staff member for conducting and verifying web crawls under staff supervision.

2. Focus Group Session

An important aspect of our project design was collaboration, both outside the 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 crawled and archived through this pilot ourselves, we desired to bring together stakeholders from various communities to help guide our collection development strategy. This proved to be a fruitful method, which generated an interesting conversation and useful feedback for both this pilot phase and future web archiving initiatives.

The focus group was held on Thursday, September 12th, from 12:00 to 1:30pm in the DSL Flex Space. 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.

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 Facebook/Twitter 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 is often found as attached PDFs.
- There are courses at MSU that would directly benefit from access to this content, including the Masters 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. The collated information was then captured through photographs and added to a shared folder for later analysis.

3. Seed Identification

Prior to the focus group, grant team member Julia Ezzo developed a list of possible seeds 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 crawl options at the focus group and participants were able to rank their importance, as well as suggest other potential seeds. Participant-recommended seeds were searched and if a URL was available, these sites were added to the spreadsheet of seed crawls. The Internet Archive's Wayback Machine was then searched to determine if and how frequently these sites may have been crawled to allow the grant team to prioritize those sites that were either never crawled or crawled only a few times.

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.

4. Crawling

Within the Archive-It subscription for MSUL, a Michigan Politics Web Crawling Project collection was created. A total of 110 seeds 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 rawls were executed in January, February, and March 2020 to fill a perceived need in gaps of capture. A total of 65.8 GB were captured during this microgrant.

The identified seeds 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

The seeds were all created as “one time” frequency for this microgrant. All crawls were run as type Test Crawl and used the [Brozzler](#) experimental Crawling Technology. Data limits and Time limits were also set in most cases.

4.1 Metadata

Metadata for the collection was added at two different levels, at the collection level and the seed level. The Archive-It application supports Dublin Core 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 a 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. We used Library of Congress subject headings, and made the geographical coverage as specific as possible to maximize the use of the facet feature shaped by the captured metadata.

Collection example

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>

Seed example

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 3 times between Oct 15, 2019 and Dec 27, 2019

Facets:

Subject Headings

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)

Geographical coverage

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 --

4.2 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 (Appendix A), 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.

We created a spreadsheet crawl log (Appendix D) for 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#
- 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.

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.

4.3 Quality Assurance and Scoping

Quality Assurance (QA) for web crawling is not a simple task and requires a lot of 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 very deep so most site links could be visually verified.

Archive-It provides some high level help on the QA process (via their Help Center) and a Wayback QA Tool. 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. Another useful website on how to conduct QA can be found 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 by the Internet Archive and removed from this microgrant work by marking them as inactive in the user interface. These seeds can be later activated for future crawls.

During QA playback of captured crawls, some web pages would come up blank. 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 techniques, such as entering explicit seeds for a site’s menu pages, using the Brozzler crawler, using different browsers during playback, or simply reloading a page in the browser. Examples of added seeds are including the URL for the About page or Publications page of a website.

Social media sites are constantly changing and the quality of captured Facebook sites varies considerably. Archive-It provides some default scoping rules for 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. While writing the report, additional crawls were completed and some seeds were moved from public to private.

A final review indicated that some crawl scoping was done at the collection level and may be better accomplished at the seed level. This will need to be reviewed when the project goes forward.

5. Access

While access to the archived websites are available through [Archive-It](#), 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 to the web archive, a [libguide](#) 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 has category boxes, such as election year and geographic delineation, with each box containing the name of the person 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 libguide also contained contact information to allow researchers to submit their questions, concerns, or other comments to one of the team members.

6. Budget and Expenses

Activity	Budgeted	Actual
Graduate student (for Archive-It crawling)	120 hours/\$1800	27 hours/\$405 (Our student worker had other campus commitments that kept him from working as many hours as we had planned.)
Archive-It subscription augmentation	\$1500	0 (fit within existing annual Archive-It subscription)
Focus Group Event	\$180	\$311.28

7. Challenges

1. By the time we had hired our student assistant in October 2019, some sites had already gone offline, example <http://www.adamhussainfor3rdward.com/> and <https://www.hughesforlansing.com/>, further underscoring to us the transitory nature of these websites.
2. City government sites were fairly extensive with video, PDF documents, and other material, and were much more complex than most election sites. Such complex sites took up more of our Archive-It annual data allotment and also required more in-depth QA to ensure that all relevant materials were preserved.
3. Wix sites could be crawled but won't play back, for example <https://www.yaniceforlansing.com/>.
4. www.voteterchambers.com images didn't play back even though they were captured. This bug is logged for Archive-It improvement in ticket #10217.
5. Time availability of the student worker was challenging. We might have made other seed decisions if there had been more allocated time for this project, such as capturing more of the complex sites noted in #2 and more attempts at successful captures of Wix sites mentioned in #3.
6. We may wish to split this into two projects with different focuses: local election and local government. The two areas have different types of websites.

8. Recommendations

In general, our web archiving program can be dramatically expanded, but currently is essentially maintained by two staff juggling many other job duties. The ideal scenario would be to hire a digital archivist, who would be willing to assume primary responsibility for web archiving activities. Otherwise, we have the following recommendations:

1. Dedicated staff for web crawling. Creating one or more web crawling graduate assistantship positions or possibly centralizing web crawling into a staff or librarian position(s) to support University Archives, Special Collections and other MSUL units. Use of dedicated workers would ensure that we fully utilize our annual allotted storage that we pay for in Archive-It subscription fees and free up the labor of staff with other duties. This would allow other archivists and librarians to focus on establishing thoughtful and ambitious web archiving goals and further collection development and seed selection.
2. Create a committee comprised of archivists, subject librarians, and other stakeholders to recommend future web archiving initiatives and select the seeds to crawl.
3. Review and reconcile 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 our collection. For example, we used the subject "Elections", others

have used “Local elections”. The [State Elections Web Archive](#) created by the Ivy Plus Libraries Confederation could be a good example.

4. Explore niche tools for archiving social media such as ArchiveSocial, Social Feed Manager, and other tools. 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.
5. Further thought should be put into the use of a 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.
6. Establish a more systematic log for seeds for crawl, permission notes, and other valuable information.
7. Establish clear collection development guidelines for future local election archiving, including assessing the frequency of crawling websites to capture changes over time.

Appendices

[Appendix A: Archive-It Training Manual](#)

Appendix B: Seed List

[Appendix C: Metadata Manual](#)

Appendix D: Seed Crawl Logs

[Appendix E: Presentation Slides](#)