Journal of Contemporary Archival Studies

Volume 8 Article 12

2021

Web Archiving in North Carolina's Piedmont Triad During COVID-19

Jessica Dame
University of North Carolina at Greensboro, jessicadame86@gmail.com

WEB ARCHIVING IN NORTH CAROLINA'S PIEDMOND TRIAD DURING COVID-19

As part of its ongoing work to document the history of the University of North Carolina at Greensboro (UNCG), the Martha Blakeney Special Collections and University Archives (SCUA) has archived UNCG websites using Archive-It since 2015. The collections include UNCG websites, news feeds, social media, student organizations, and a selection of community content. The online information relevant to the communities SCUA serves is highly volatile and if not captured regularly may be lost. Utilizing Archive-It's COVID-19 Web Archiving Special Campaign in the spring of 2020, SCUA expanded web archiving efforts to include COVID-19-related web content from the local community. This work-in-progress paper discusses the creation and challenges of web archiving COVID-19 web content and explores the significance of the web content and importance of capturing it during a global pandemic.

Background

Web archiving is a field with many noted challenges. In addition to technical challenges, web content is constantly in a precarious state: it is ever-changing, its born-digital formats continue to grow, and there are benign and deliberate means in which content can disappear.[1] The normal challenges to capturing web content, combined with rapidly changing information about the COVID-19 pandemic, created a sense of urgency to archive quickly and was made more poignant by the historical significance of the times.

Shelter-in-place orders, remote work, and changes to social and professional life alike increased internet usage and the need for web-based content in 2020. According to the Pew Research Center, 53 percent of Americans said that the internet was essential to them during the pandemic.[2] The Global Web Index reported that 68 percent of consumers were seeking out pandemic updates online.[3] Data usage and online-based activities such as streaming, gaming, and conducting business increased significantly.[4] Individuals were more reliant on web content with offices shuttered, and governments, organizations, and businesses met this need by providing web-based information including statistics, travel restrictions, practical guidance on protection, and governmental response.[5]

Initiatives to collect COVID-19 web content have been underway from the start of the pandemic. Many institutions and groups have shared their collections in collaborative documents including Documenting the Now's "Documenting COVID-19" and the Society of American Archivists Web Archiving Section's "COVID-19 Web Archives" roundup.[6] These collaborative documents are an accessible way for professionals to share their collections with each other and provide a network of people focusing on the archiving of COVID-19 content. The School for Historical, Philosophical, and Religious Studies at Arizona State University invites the public to share personal stories

[1] Taylor, "Introduction to the Special Issue on Web Archiving," 3.
[2] Vogels et al., "53% of Americans Say the Internet Has Been Essential."
[3] Jones, "This Is How COVID-19 Has Changed Media Habits."
[4] Cohen, "Data Usage Has Increased 47 Percent during COVID-19 Quarantine."
[5] United Nations Division for Public Institutions and Digital Government, "UN/DESA Policy Brief No. 61."
[6] "Documenting COVID-19"; Greenhouse, "Web Archiving Round-Up."

and digital objects through crowdsourcing for "A Journal of the Plague Year."[7] The National Library of Medicine began web archiving COVID-19 resources in the "Global Health Events Web Archive" and is just one of many collections captured using web archiving services like Archive-It.[8]

Archive-It is a web archiving service built at the Internet Archive and used by a wide range of organizations for, among other uses, the collecting of cultural heritage from the web. The service captures and preserves web-based content; allows users to add, import, and export descriptive metadata; and provides for public access and full-text search.[9] The public-facing side (https://archive-it.org/) is freely accessible and connects to the Internet Archive's WayBack Machine, which allows users to select a crawled URL and the date it was crawled to view archived content. In response to an increase in institutions wanting to build web archive collections documenting the impact of the COVID-19 pandemic, Archive-It presented the COVID-19 Web Archiving Special Campaign in the spring of 2020.[10] The campaign was an opportunity for subscription holders to expand their web archiving data and for new subscribers to begin their introduction into web archiving through subsidized accounts for archiving their institutional and community COVID-19-related web content.[11]

Following UNCG's early announcements and monitoring of COVID-19 in March 2020, SCUA began archiving UNCG's COVID-19-related web content in the UNCG COVID-19 Collection. In April, Archive-It announced the data expansion discount to help partners capture important content related to the COVID-19 pandemic.[12]

The university archivist reached out to other universities in the region by email to learn if anyone planned to web archive COVID-19 content in the Piedmont Triad (Triad). This twelve-county region is located in the central part of North Carolina and has a population of 1.7 million.[13] After learning that other institutions were not planning to web archive COVID-19 content, the archivist quickly decided to apply for the expansion to collect web content across the Triad during a historical moment for future research and context that had the potential to be lost.

Additional factors contributed to this decision. In March 2020, UNCG moved operations online and closed campus for the remainder of the spring trimester. SCUA had limited access, and work that could be completed from home took center stage. In the same timeframe, SCUA added the temporary part-time position of archives and records technician to start the week the campus closed. The individual in the technician position came with intermediate experience using Archive-It. Starting in March and accomplished throughout 2020, the technician created documentation including the SCUA "Web Archiving Manual and Metadata Dictionary," assessed existing UNCG web archive collections for gaps, updated redirected crawls and errors, applied metadata across all web collections, and performed quality control on recurring crawls. Additionally, web archiving

[7] "A Journal of the Plague Year."
[8] National Library of Medicine, "Global Health Events Web Archive."
[9] "About Archive-It."
[10] "COVID-19 Web Archiving."
[11] "COVID-19 Web Archiving."
[12] "COVID-19 Web Archiving."
[13] "About the Region."

provided the opportunity to collect COVID-19 content without requiring the community to create and submit directly to SCUA during a global pandemic.

Method

SCUA received the Archive-It data expansion in April 2020, and the Triad COVID-19 Collection began. The university archivist led the project, providing guidance and regional knowledge to the technician, who initiated the call for content submissions, assessed those submissions, ran tests and recurring web crawls, created metadata manually for each new seed (URL), and performed quality control of recurring crawls.

The Triad COVD-19 Collection captures and preserves how the Triad community shared information during the global pandemic and how the pandemic affected daily life. The scope of the collection follows the topics suggested by Archive-It to help guide collecting including coronavirus origins, information about the spread of infection, regional and local containment efforts, social aspects, economic aspects, and political aspects.[14]

The university archivist and technician narrowed down the scope to websites, documents, and videos containing COVID-19 information created by county government, regional hospitals, K–12 schools, universities and technical schools, nonprofit organizations, community landmarks, and community initiatives. To keep the focus on COVID-19 content, crawls were set up to only capture web pages containing COVID-19 information. For example, Cone Health's homepage (https://www.conehealth.com/) would not be a candidate for crawling, but its COVID-19 information page (https://www.conehealth.com/covid-19-information/), providing information on how to stay safe, how to get a COVID-19 test, and the latest numbers of hospitalization and infection, would. The scope excluded social media, with the exception of a small number of YouTube videos, due to challenges Archive-It was experiencing crawling social media at the time the collection was created and the amount of data the sites would use if captured successfully.

The university archivist and technician quickly developed a plan for selecting and capturing web content. This early stage of selecting web content was a collaborative effort. The technician invited SCUA staff members to participate voluntarily to nominate content from one or more of the twelve counties in the Triad region. Six staff members participated, and each focused on an average of two counties.

Participating SCUA staff members submitted nominations using Google Documents. A Google Document was created for each county and was "due" by the end of May 2020. This due date was flexible but was selected in an effort to collect as much web content as quickly as possible with evaluations beginning as content was nominated. Information accompanying nominations included URLs, title, creator (if known), and a description or note when applicable. Due to the urgent nature of the collection, the technician immediately evaluated captured web content and began web archiving as early as mid-May. The technician assessed each nomination and made final selections for inclusion in the Triad COVID-19 Collection based on the item's scope. This nomination period was a starting point. As SCUA staff submitted new web content throughout the year, it was evaluated and added on a case-by-case basis.

[14] Bailey, "Archiving Information on the Novel Coronavirus (Covid-19)."

The technician immediately crawled approved content, and as a result reached out to site owners regarding permissions afterward. The university archivist and technician employed an "opt-out" approach, notifying site owners about the collection and web crawls and informing them of the option to decline to be included if they preferred. The technician created a message template based on the template used for opt-out letters in SCUA's "Web Archiving Manual."[15] The template could easily copy and paste into an email or online form. The technician made an attempt to reach out to each site owner with the exception of government websites, whose information is a matter of public record. This process was difficult as many websites had no contact information beyond a generic Contact Us form. Other times the website belonged to a large entity, such as a university or hospital, and it was not clear who was the appropriate contact. In these cases, the technician contacted multiple individuals in hopes of reaching the right person. Of the approximately seventy-seven site owners contacted, only eight responded. Those eight who replied, however, were excited to be part of the archived collection.

The longest portion of the project was metadata creation due to the volume of newly added content that required manual entry. The technician added metadata following successful test crawls. Metadata included the required fields defined by SCUA's "Web Archiving Manual," and descriptive metadata was created using Dublin Core terms. Required fields outlined in SCUA's manual include "Collector," "Creator," "Language," "Rights," "Title," "Type," and "URL (Resource Identifier)." Added optional fields for context include "Date," "Description," and "Subject(s)."[16] Metadata and vocabularies in the manual are guided by OCLC Research's Web Archiving Metadata Working Group (WAM).[17]

The system crawled most web content weekly in the early months of the project, while other content (such as PDFs) was scheduled to crawl one time for an initial capture. This schedule worked from May into September 2020. Each seed added into Archive-It was test crawled and assessed for quality to ensure the site was captured as accurately as possible. For recurring crawls, the technician performed quality control when crawl reports were available to catch flagged errors. It was important to address broken URLs, redirected URLs, and incomplete crawls as they came up to be sure COVID-19 information continued to be captured.

In October 2020, the technician crawled a series of recorded testimonials featuring students from high schools and universities across the Triad made available by the ad agency Vitalink on YouTube.[18] The videos feature students sharing their COVID-19 experiences and reasons for wearing masks. The technician shared with the university archivist their concern about reaching the annual data limit after adding the YouTube videos to the collection. The videos were the first and only social media content added to the collection due to their large data size. To accommodate recurring crawls of the Triad COVID-19 Collection alongside the ongoing UNCG web collections, the collection schedule was adjusted to archive monthly. Despite this change, the web archive quickly reached its annual data limit in November 2020. All web crawling ceased from November 2020 through January 2021, when SCUA's Archive-It contract was renewed. Web crawls began

[15] Dame, "Web Archiving Manual and Metadata Dictionary," 4–5.
[16] Dame, "Web Archiving Manual and Metadata Dictionary," 8.
[17] Dooley and Bowers, "Descriptive Metadata for Web Archiving," 16–34.
[18] Vitalink, "COVID-19."

again in January 2021, and as of March 2021 the Triad COVID-19 Collection includes 150 unique pieces of crawled web content.

Across all actively crawled web archive collections in 2020, SCUA used 1.2 terabytes. Of this yearly total, 856.9 gigabytes were used by the Triad COVID-19 Collection. To highlight the significance of the data size, the two largest UNCG collections, UNCG Social Media Collection and UNCG Website Collection, over a six-year period as of March 2021 crawled 751 gigabytes and 591.6 gigabytes respectively.

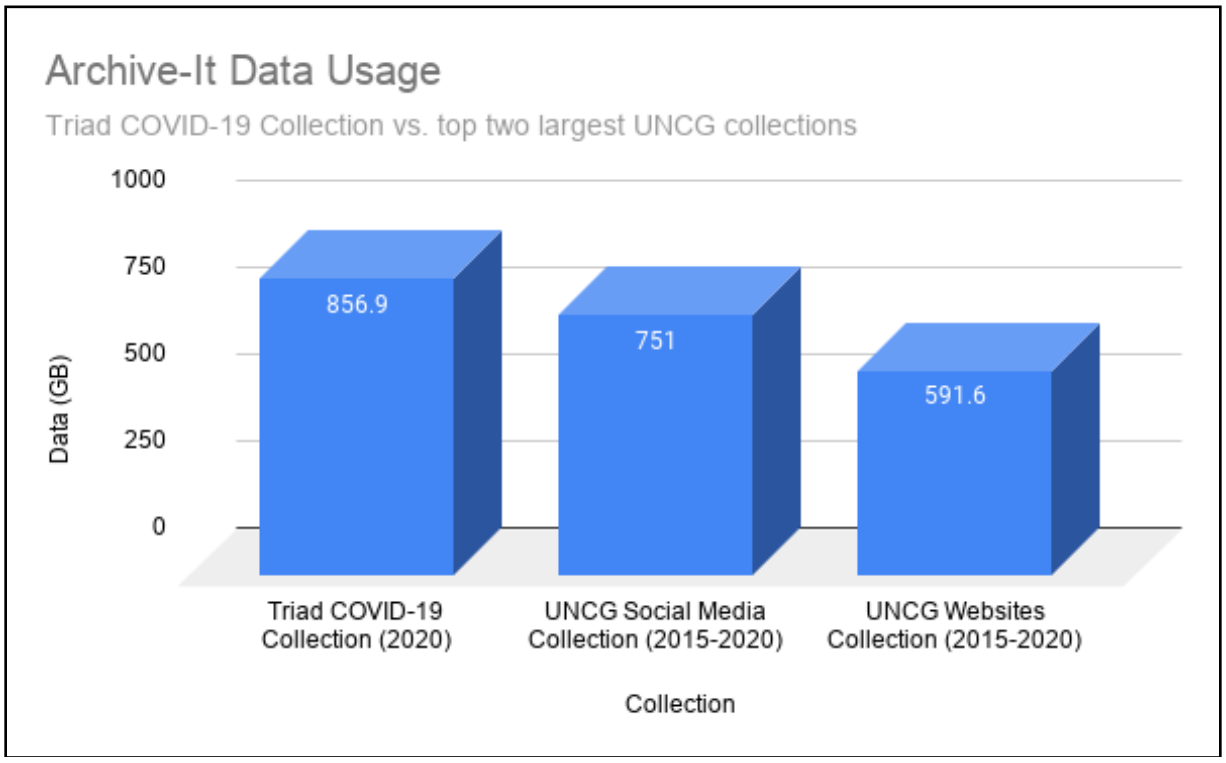

Figure 1. Archive-It data usage across SCUA's three largest web archive collections

Observations

Challenges emerged in the creation of the Triad COVID-19 Collection including reaching data limits quickly, limitations of capturing social media, dynamic content and quickly changing interfaces, web content disappearing, staffing time limitations, and email disruptions. The subscription level SCUA could purchase in 2020 and 2021 set the overall limit for how much content could be captured and the overall growth of the collection. Additionally, collecting content had the potential to be a challenge due to the rapid submission and collection process, but the larger issue came in capturing as much diverse content in scope within the data limit as possible. In creating the collection rapidly, the university archivist and technician presumed web content could be missed such as community initiatives in smaller counties or initiatives with information exclusively available on social media.

A highlight from the collection perfectly captures the urgency to collect content quickly. Project Mask WS is a mask sewing initiative in Winston-Salem created in response to the pandemic.[19] According to the Project Mask WS website in the WayBack Machine, they are a group of one thousand-plus volunteers who create masks for medical personnel and frontline workers who cannot obtain N95 masks.[20] The website features an introduction to the project, images of frontline workers wearing the masks, and examples of its impact on the community. The website expired in mid-July, and the group's online presence is now exclusively on Facebook.[21] While the initiative exists without its original website, the origins of this project have been potentially lost. Between May and December 2020, eleven websites and web pages have been taken down or expired in the Triad COVID-19 Collection.

Future Steps

SCUA renewed its Archive-It subscription with a data limit of 512 gigabytes in January 2021, half of the expanded 2020 limit. Although the pandemic is ongoing and information continues to be shared online, the lower limit impacts the frequency of crawls that can be accomplished alongside regularly scheduled UNCG web collections. The university archivist and technician agreed on crawling the Triad COVID-19 Collection quarterly throughout 2021. Quarterly crawls will include performing test crawls to catch report errors and to update redirected crawls as long as the new seeds still point to COVID-19 information.

The technician began outreach in December 2020 by publishing a short article on the Triad COVID-19 Collection in the *UNCG Special Collections and University Archives Newsletter*. The newsletter reaches various groups across the UNCG campus and groups offsite such as the friends of the library group and donors. In January 2021, the technician presented on a virtual library panel for UNCG's University Libraries Virtual Learning Community, an internal UNCG University Libraries series, to discuss resources for community connection and introduce the Triad COVID-19 Collection to the UNCG library community. Further outreach includes sharing the collection on SCUA's social media accounts, reaching out to *UNCGNews* (a weekly campus news email), and writing short articles announcing the collection for local professional organizations such as the Society of North Carolina Archivists.

The Triad COVID-19 Collection archives the information regional institutions and initiatives shared on the web during the global pandemic, how the pandemic affected daily life, and will inform future research and histories of the Piedmont Triad's response during a global crisis. Readers can access the Triad COVID-19 Collection at https://archive-it.org/collections/14142.

Bibliography

"About Archive-It." Archive-It. https://archive-it.org/blog/learn-more/.

"About the Region." Piedmont Triad Regional Council. https://www.ptrc.org/about/about-the-region.

[19] "Project Mask WS."
[20] "Project Mask WS."
[21] Project Mask WS, "About."

Bailey, Jefferson. "Archiving Information on the Novel Coronavirus (Covid-19)." Internet Archive Blogs, February 13, 2020. http://blog.archive.org/2020/02/13/archiving-information-on-the-novel-coronavirus-covid-19/.

Cohen, Jason. "Data Usage Has Increased 47 Percent during COVID-19 Quarantine." *PC Magazine*, June 5, 2020. https://www.pcmag.com/news/data-usage-has-increased-47-percent-during-covid-19-quarantine.

"COVID-19 Web Archiving Special Campaign." Archive-It, April 14, 2020. https://archive-it.org/blog/post/covid-campaign/.

Dame, Jessica. "Web Archiving Manual and Metadata Dictionary." Martha Blakeney Special Collections and University Archives, University of North Carolina at Greensboro, 2020.

"Documenting COVID-19." Documenting the Now (Google Doc), April 3, 2020. https://docs.google.com/document/d/1v5tso8spFq6SpW53h2OJULcdRoPEbyI6xpah31kW-H0/edit.

Dooley, Jackie, and Kate Bowers. "Descriptive Metadata for Web Archiving: Recommendations of the OCLC Research Library Partnership Web Archiving Metadata Working Group." *OCLC Research* (2018): 18–36. https://www.oclc.org/content/dam/research/publications/2018/oclcresearch-wam-recommendations.pdf.

Greenhouse, Nicole. "Web Archiving Round-Up: COVID-19 Edition." Society of American Archivists Web Archiving Section, April 13, 2020. https://webarchivingrt.wordpress.com/2020/04/13/web-archiving-round-up-covid-19-edition/.

Jones, Katie. "This Is How COVID-19 Has Changed Media Habits in Each Generation." World Economic Forum, April 9, 2020. https://www.weforum.org/agenda/2020/04/covid19-media-consumption-generation-pandemic-entertainment/.

"A Journal of the Plague Year." School for Historical, Philosophical, and Religious Studies, Arizona State University, Tempe. https://covid-19archive.org/s/archive/page/Share.

National Library of Medicine. "Global Health Events Web Archive." Archive-It. https://archive-it.org/collections/4887?fc=websiteGroup%3ACoronavirus+disease+%28COVID-19%29+outbreak.

Project Mask WS. "About." Facebook. https://www.facebook.com/groups/projectmaskws/.

"Project Mask WS." Project Mask WS. https://wayback.archive-it.org/14142/20200704101130/https://www.projectmaskws.org/.

Taylor, Nicholas. "Introduction to the Special Issue on Web Archiving." *Journal of Western Archives* 8, no. 2 (2017): 1–6. https://digitalcommons.usu.edu/westernarchives/vol8/iss2/1.

United Nations Division for Public Institutions and Digital Government. "UN/DESA Policy Brief No. 61: COVID-19: Embracing Digital Government during the Pandemic and Beyond." United Nations, April 14, 2020. https://www.un.org/development/desa/dpad/publication/un-desa-policy-brief-61-covid-19-embracing-digital-government-during-the-pandemic-and-beyond/.

Vitalink. "COVID-19." YouTube, October 22, 2020. https://www.youtube.com/playlist?-list=PL4SnDwq4DIZdwsg5D2nA5TWJi0D_u-fw6.

Vogels, Emily A., et al. "53% of Americans Say the Internet Has Been Essential during the COVID-19 Outbreak." Pew Research Center, April 30, 2020. https://www.pewresearch.org/-internet/2020/04/30/53-of-americans-say-the-internet-has-been-essential-during-the-covid-19-outbreak/.