Web Archiving Project: Urban Farming Collection

Group 1: Angela Hummel, Shelby Silvernell, Naomi Trujillo

INFO 284 Tools, Services, and Methodologies for Digital Curation

San José State University

December 6, 2019
Web Archiving Project: Urban Farming Collection

The purpose of this group project was to design, execute and critique a web crawl using Archive-It (specialized open-source software) to harvest and preserve a humanities web-based collection that will be permanently hosted at the Internet Archive. The Archive-It software harvests material from the Web through crawlers, which is then publicly accessible through Archive-It (with metadata) and the Wayback Machine at the Internet Archive (Internet Archive, 2018). Our group scoped the collection, created metadata, and performed troubleshooting on rendering issues, which allowed us to develop an understanding of website architecture, robot.txt files, copyright matters, and other web archiving issues and challenges. This document provides the description of and rationale for our web archive collection, the seed websites that comprise our collection, what data we captured and how we adjusted the scope of crawls, how we performed troubleshooting and solved website rendering issues, and our major takeaways from the project.

Collection Selection and Description

To select a topic for our web archive collection, our group brainstormed topics individually and compiled these into a list prior to our first meeting. On this list: seed libraries; tiny house blogs; museums: from brick and mortar to online websites; posters: art and propaganda; national parks; blogs about bacon; Women’s World Cup 2019; sustainable fashion; chocolatiers; tattoo artists; unions; LGBTQ+ organizations or archives; and diversity, equity, and inclusion movements in information organizations. We then met to vote on the themes that seemed the most interesting for a website collection.
After discussing each of the topics, we ended up excluding those that may not make for a compelling collection of websites or for which there may not be many websites. We ultimately narrowed our focus to the idea of urban farming. This subject seemed appealing given both the environmental and social impact these types of projects often have. Our group members already knew of exciting examples of these types of initiatives in our individual home cities, and each of these took a slightly different approach to urban farming: youth education, job training, and refugee and immigrant support. This provided some framework for our continued research to find seed sites, and we all began looking for examples of non-profits which would demonstrate a diverse range of the social impact of urban farm programs.

We created a spreadsheet to start compiling representative samples of non-profit urban farm websites, while keeping in mind the different areas of programming in which each specialized. After coming together for another meeting, we also decided that geographic distribution across the United States could help to structure our collection of websites. The non-profits were clustered into their geographic regions: West Coast, Midwest, South, and East Coast. The group then voted on their top two to three websites for each region to reach our final sampling of nine total seed sites.

Website selection based on a topic requires human judgement (Niu, 2012); while manual, value-based, representative-sample selection is time-consuming for information professionals, this process was suitable for our small-scale web archive. We did not limit our selection criteria by domain—while most relevant websites were in the .org domain, some were .com sites.

Collection Seeds, Metadata, and Significant Properties
The Urban Farming collection is made up of nine seed websites, as listed below:

<table>
<thead>
<tr>
<th>Organization</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>City Blossoms</td>
<td><a href="http://cityblossoms.org/">http://cityblossoms.org/</a></td>
</tr>
<tr>
<td>Farm LA</td>
<td><a href="http://farmla.org">http://farmla.org</a></td>
</tr>
<tr>
<td>Farms at NYCHA</td>
<td><a href="https://www.greencityforce.org/farmsatnycha/">https://www.greencityforce.org/farmsatnycha/</a></td>
</tr>
<tr>
<td>Growing Home</td>
<td><a href="http://growinghomeinc.org/">http://growinghomeinc.org/</a></td>
</tr>
<tr>
<td>Harlem Grown</td>
<td><a href="http://www.harlemgrown.org">http://www.harlemgrown.org</a></td>
</tr>
<tr>
<td>Living Well Kent</td>
<td><a href="http://livingwellkent.org/">http://livingwellkent.org/</a></td>
</tr>
<tr>
<td>Ohio City Farm</td>
<td><a href="https://www.ohiocityfarm.com/">https://www.ohiocityfarm.com/</a></td>
</tr>
<tr>
<td>Plant It Forward</td>
<td><a href="http://plant-it-forward.org/">http://plant-it-forward.org/</a></td>
</tr>
<tr>
<td>Rid-All Green Partnership / Green Ghetto</td>
<td><a href="https://www.greennghetto.org/">https://www.greennghetto.org/</a></td>
</tr>
</tbody>
</table>

Our collection started with 10 seed websites, which our three-member team divided into groups of four, three, and three, with each team member responsible for the archiving of their seed group. Each person made the initial archiving decisions for their seed group: initiating test crawls, reviewing post-crawl and Wayback results, troubleshooting rendering issues, researching and implementing scoping and filtering adjustments, and scope changes. During team calls, the group discussed any issues that were proving difficult to resolve in order to learn from each other, as well as issues that required a team decision.

Metadata was created using the Library of Congress subject headings (Library of Congress, n.d.-a) and place names from the Getty Thesaurus of Geographic Names® Online.
We consulted the MIME controlled vocabulary list provided in the assignment materials for the file formats, and we also reviewed one another’s metadata to verify that there was consistency and unity in content and tone. To reach consensus for the subject terms and format types collection-level metadata, we compiled the aggregate from our seed-level metadata and voted on terms that best represented the collection as a whole.

As we considered the scoping of our collection and what content we needed to capture, the issue of significant properties required careful consideration—in the context of web archiving, significant properties raise some interesting points about identity and sameness. The term significant property refers to the qualities or aspects of a digital (or any) object that make it uniquely that object (Yeo, 2010). These are the properties which must be preserved throughout the lifecycle of the digital object in an archive or, perhaps, a website, such that the meaning, usefulness, accessibility of that object remains intact. In the context of digital objects in an archive, the determination of significant properties may involve decisions about file formats, metadata, structure, appearance, and other technical aspects that are important to defining the authenticity of that object and must accompany that object through its lifecycle.

What constitutes the significant properties of a website and its contents, and how do these concepts apply to, and guide, the process of web archiving? Yeo (2010) makes several interesting points that explore the concept and application of significant properties. First, the idea that significance is relative to the user or audience and not necessarily based on something inherent in an object. Significance is something that we bestow upon an object based on the user audience context and need. The idea of significant properties is further challenged by the idea of authenticity in the digital realm. What is really original when considering digital objects, especially in situations where numerous instances of an object can be created, as in a website? In
the digital world, Levy (as cited by Yeo, 2010) suggests that “there are only copies” (p. 107). Ultimately, an exploration of what and for whom the object is being preserved, or the website is being archived, can inform decisions made around lifecycle processes such as selection and appraisal, data capture, etc.

In the context of this archive collection, which is clearly a copy, users defined our vision of the significant properties required for capture: the text, images, structure and functionality of the website seed that allow viewers to understand the organization’s mission, activities, and value in the world of urban farming in the United States.

**Collection Scoping and Filtering**

Initially, to better understand our seed site structure and file format makeup, we divided the websites and each ran test crawls with the default scoping rules. Our crawl limits only included time limits, and we opted for limits between three and five days for each of our respective groups of websites. These initial baseline crawls allowed us to determine what further adjustments needed to be made in order to better capture and represent the content of each website. While we each ran standard test crawls to begin the archiving process, several sites did not respond to scoping and filtering adjustments as expected. We found that those sites required a Brozzler (experimental) crawl to accurately capture site appearance, content, and functionality.

Due to issues with embedded links and media not appearing correctly, in addition to design and content problems, the collection-level rule “ignore robots.txt” was added for the following domains based on recommendations on the Archive-It Help Center: akamaihd.net, fbcdn.net, fonts.gstatic.com, gcs-vimeo.akamaized.net, googlevideo.com, images.squarespace-cdn.com, static1.squarespace.com, www.facebook.com, www.instagram.com, youtube.com,
ytimg.com, frog.wix.com, and static.parastorage.com (Archive-It, 2017b; Archive-It, 2019a; Archive-It, 2019b; Archive-It, 2019c; Archive-It, 2019d). We also had issues with portions of pages, entire pages, or entire websites not appearing, so the rule “ignore robots.txt” was also added to select individual seed rule scoping. Individual seed scoping rules included rules specific to the unique site construction of each site. For example, Wix sites require scope inclusion (accept URL if it…) of “siteassets.parastorage.com” and “static.wixstatic.com” to capture images and content, as well as the addition of a Regular Expression to avoid crawler traps (Archive-It, 2019c).

Seed Content and File Types Captured

Given the nature of these websites—informational sites for non-profit organizations—we opted to capture each seed at the root-level of the URL using the default standard (Archive-It, 2019e). The majority of the sites included pages that were linked from this homepage root URL, and these were successfully crawled given their direct relationship with the root URL (Archive-It, 2019f). Given the fact that the non-profit organizations were the focus of this collection, we were satisfied with Archive-It not crawling links to websites outside the root URL.

The group encountered a challenge in capturing embedded social media streams on some of the pages, as these dynamic feeds did not immediately successfully crawl. We attempted to add scoping rules that might enable this content to appear in our crawled sites, but ultimately, this was not successful. We decided not to crawl these social media pages individually in order to have this content show up on the pages, as this would eat into our storage space allotment. Crawling the social media pages also felt out-of-scope more generally speaking, as the core websites of the non-profits were our primary focus for this collection. In some ways, this collection feels like a pilot for what could be a larger, more comprehensive selection of websites.
on urban farming. If a second, more extensive phase of this collection were to evolve, we could imagine social media sites being included.

Our crawls produced a variety of file types for this collection of websites. The most common formats included: text/html, text/css, text/xml, image/jpeg, image/png, image/gif, application/javascript, application/json. A chart of the quantities of the primary file types follows:
Seed Rendering Issues and Troubleshooting

The group encountered a wide variety of rendering issues and conducted a significant amount of seed-level troubleshooting, as detailed below for each seed website. For all the websites, quality assurance was performed by checking the crawl reports, visually comparing the Wayback results with the live sites, and generally navigating around the Wayback results.
**CityBlossoms.org.** The first test crawl was conducted with no scope limits or rules in place; this was done with the intention of gathering initial information and to see how Archive-it handled the contents of the website. This crawl presented with an archive that had a variety of elements that displayed differently than on the live site; these included images not showing and fonts displaying incorrectly. For the second crawl and subsequent test crawls, various actions were taken such as adding the rule “ignore robots.txt,” using Brozzler, investigating what rules could be helpful in attempting to archive a Squarespace website (Archive-It, 2017), as well as noting rules that could allow for fonts to display correctly (Archive-It, 2016). By the last test crawl, much of the content was available for viewing though the whole website, but the fonts are still not displaying as the designer intended (as seen on the live website) in spite of managing the scope rules both at collection and at seed level. Viewing on other devices (for mobile and other than desktop) was preserved. There were a number of variables that appeared to impact the ability to accurately view the archived websites. These included which browser was used, how large the viewing window and viewing zoom percentage chosen, and the strength of the bandwidth accessed by the device.

**FarmLA.org.** This site is primarily structured as one long page, which visitors can either scroll down in order to access content, or they can use the navigation bar at the top to jump down to the relevant section. The first crawl for this website, which used the standard crawler and was limited to three days, resulted in one fairly obvious issue. The first and primary header image content automatically rotates through images, creating a slideshow; this content did not appear at all. The white, empty top of the page meant that the navigation bar at the top did not appear either. The first test crawl report did not show any hosts as having content that was blocked or queued. We tested two different approaches in order to fix the slideshow issue: trying the
Brozzler experimental crawler and adding the “ignore robots.txt” rule to the seed. Ultimately, adding the seed rule was enough to get this image content to render correctly. Comparing the Wayback test crawl with the live site, all content appeared correctly and navigation functioned properly otherwise. And fortunately, because the website was essentially one long page, verifying the results was more straightforward than navigating between pages in Wayback (we encountered serious lag when looking at our test crawls in Wayback). One navigation link that does take the visitor to another page is the Press page. This page’s structure crawled well, but some of the embedded YouTube videos did not—they resulted in black boxes. One final issue was mismatched font in the top navigation bar between the Wayback test crawled site and the live site. It was not apparent how to fix this given the fact that font is usually established in the CSS file. It may be worth getting feedback from the creator of the site, given the fact that font is a design choice, and could therefore be a significant property of the website.

Unfortunately, after saving one of the successful test crawls as a production crawl, the slideshow at the top of the archived page once again disappeared. After attempting to fix this via a patch crawl in Wayback QA mode, it came to our attention that the slideshow on the live site was no longer working. Therefore, patch crawls to fix this functionality were unsuccessful, as it would seem that these files have disappeared (HTTP 404 error). The test crawl for November 15, 2019 demonstrates that the slideshow was indeed functional on one previously crawled test version of this seed site.

GreenCityForce.org/FarmsatNYCHA/. The first test crawl was run with no constraints and primarily for information gathering with later crawls incorporating “ignore robots.txt” and Brozzler. The header banner appeared initially but showed up only intermittently in later crawls depending on the browser chosen as well as the power of the wireless network or other network
chosen to access the internet. In fact, there appear to be many variables that affect the viewing of the archived sites. Since this is a WordPress site, the appropriate considerations were put in place to overcome any limitations to archiving (Archive-It, 2017). Since this is a single page out of Green City Force’s website, the ability to gather the info for archival purposes was not as complex as some of our other seeds. On this page and really on the website as a whole, there is not a lot dynamic content. The page is not updated very often so the variables that might affect archiving intentions were less influential. In an effort to understand, though, why the banner header on this page did not appear consistent, we utilized the Mozilla Firefox Archive-it Proxy Mode extension (Archive-it, 2019g). This appeared to work initially though later changes in network access and browser view zoom issues again affected the banner header zoom percentage. What did work and display consistently were the fonts. These displayed accurately every time; this may be because the fonts were “SquareSpace fonts” and not those as indicated from a third party.

**GrowingHome.org.** The first test crawl revealed the capture of only page of the website, and a missing top logo, footer images, and incorrect fonts. We were able to fix these issues in a second test crawl by updating the seed URL and expanding the seed scope to include fonts (“fonts.googleapis.com” and “fonts.gstatic.com”). Dynamically-generated Google Maps and Google forms were not captured, which is a known issue (Archive-It, 2019g). Social media links (Twitter; Facebook; Instagram) and external links (YouTube) were not followed, which was intentional. A few graphics and PDF issues were identified as issues on the live site, which could not be corrected in a web archive. We were surprised that from Wayback we were able to download the compressed Press Kit folder and open all contents (even the image files).
**HarlemGrown.org.** This first test crawl was also conducted with no scope limits or rules in place. Even with no scope rules at either the collection or seed level, most of the website was accessible in this first crawl. The navigation was mostly intact; noted was that the top level navigation links (“Who We Are,” “Partners,” etc.) did not work, however. Later rules stated for the specific links for these top level pages to be included in the crawl. Embedded videos played back well. All images appeared to be in place. The fundraising site (third party) had lost its formatting. Fonts once again were not consistent with the live website. Later test crawls incorporated rules such as “ignore robot.txt,” and other rules related to fonts, etc., (at seed and collection levels). While the issue with fonts displaying incorrectly is still occurring even with several rules placed in hopes of influencing this situation and in spite of the fact that this is a SquareSpace site as was cityblossoms.org (Archive-It, 2017), the website was otherwise archived fairly successfully.

**LivingWellKent.org.** The first test crawl, using a standard crawler with a time limit set to three days, did not yield any content for the seed Living Well Kent. Only 15 documents (50.9KB) were crawled, and the page appeared completely blank in Wayback. There were three documents that were blocked, but these all related to Facebook. Otherwise, there were no clear indicators in the crawl report as to why the Wayback page was blank. Adding the rule “ignore robots.txt” to this seed and crawling it again resulted in the homepage and all the linked pages appearing correctly. One issue remained with this seed: empty embedded photo galleries on the Photo Gallery page. We tried using Brozzler experimental crawler to see if this would remedy the photo galleries issue, but it did not yield any noticeable helpful result. In looking at the source code, this particular page appears to be relying on Javascript to help this content render;
unfortunately, this dynamic media is a known challenge to Archive-It (Archive-It, 2019g). No solution was found to help this content display.

OhioCityFarm.com. The first test crawl revealed the capture of only the background color squares and limited text. We identified this site as a Wix site, for which Archive-It applies automated Wix site rules and recommends running Wix sites as a Brozzler crawl if there are issues (Archive-It, 2019c). This made all text appear and the menus work, but images still did not appear. After more research, we added a Regular Expression rule to avoid a crawler trap and keep the crawler from capturing unnecessary URLs generated by some Wix sites (Archive-It, 2019c). To address the missing images, we reviewed hosts and found that no hosts were listed as “out of scope” except the external links, goo.gl (Google Forms), and events.constantcontact.com; we ensured image hosts “siteassets.parastorage.com” and “static.wixstatic.com” were in scope, and couldn’t find any other image “<src>” in the source code that may be an issue or anything in the ArchiveReady report that could point to the issue. Ultimately, we added the image URLs for the images on the home page as a test to determine if adding specific image URLs to seed scope would work, which did not. We found that the images briefly flash in Wayback when viewed in Mozilla Firefox, but do not remain, even in Archive-It Proxy Mode.

After six test crawls we felt we had refined the site archiving as far as our knowledge and the Help Center could take us, and our earlier attempt to contact Support for a different seed (see OurSchoolatBlairGrocery.org below) was unsuccessful so we ruled out that option. We collectively decided to archive the site as is: the site is a valuable member of our collection and the text does paint part of the picture of the organization’s place in the community. While the photos would help paint the full picture, we have accepted that this seed site ideally illustrates
that web archiving is imperfect—we archive what we can when we can, rather than wait for perfect.

Using the site as evidence of the imperfection of web archiving proved prophetic. After saving the test crawl, we discovered that the previously showing text had disappeared entirely. We reviewed the Wayback QA report and ran a patch crawl, with no improvement. We then ran a last one-time Brozzler crawl for the site in an attempt to re-capture what was somehow lost between the test crawl and actual archiving of the site. The result was mixed—the text reappeared as did some previously missing images. Another Wayback QA review and patch crawl resulted in no change.

**OurSchoolatBlairGrocery.org & SchoolatBlairGrocery.blogspot.com.** Our group was interested in including either version of this urban farm’s website in its collection, and we originally attempted to crawl both. Unfortunately, after using the standard crawler and setting the time limit to three days, both seeds resulted in blank pages in Wayback. The crawl report did not flag any notable blocked or queued hosts. However, the .org site only included 7 documents in its crawl (4.5KB), and the majority of the content for the blogspot.com site was listed as “out of scope” in the report. The rule “ignore robots.txt” was automatically added to the blogspot site, and we also added it to the .org site; additionally, we used the Brozzler crawler. None of these interventions yielded better results in our subsequent crawls. We also attempted to reach out for guidance in early November to the Archive-It Help Desk, but we did not receive any response to our query. Because the sites crawled as blank pages, they would not be meaningful additions to our collection on Archive-It. As such, we decided not to proceed with either of these sites and dropped this nonprofit from our collection as a result of these unresolved problems.
**PlantItForward.org.** The seed site Plant It Forward crawled fairly successfully during the first test run using the standard crawler and opting for a three-day time limit. All of the pages linked from the navigation at the top of the site rendered correctly when the Wayback version of the site was compared with the live version. The crawl report did not show any hosts as having content that was blocked or queued. The biggest discrepancies came in the form of dynamic content on the pages: Google maps, a proprietary third-party volunteer calendar, and embedded social media sites (Instagram, Facebook, and YouTube). The scope for Plant It Forward was expanded to include the farmshare newsletter archive. This newsletter hub is hosted on mailchimp rather than the host seed’s primary URL, so it was beyond the threshold standard scope for the initial test crawls. After adjusting the scope rule for this seed to “expand [the] scope to include URL if” it contains the text “https://us11.campaign-archive.com” and adding the “ignore robots.txt” rule for “us11.campaign-archive.com,” the primary newsletter link worked properly. Otherwise, links to external pages, sponsor sites, the online store, and CSA login were not crawled, as was expected and intended. It is worth noting that two test crawl variations of the site were saved as production crawls: a mid-November version without a functioning newsletter page and a late-November version with a different homepage cover image and a functioning newsletter page. We opted to keep both to demonstrate this visual difference, especially given that this change occurred on the homepage of the site.

**Rid-All/Greennghetto.org.** The first test crawl was not accessible in Wayback, so we initiated a second crawl using Brozzler. Subsequent Brozzler crawls had limited accessibility in Wayback and missing page content, but the final test crawl with an expanded seed scope to include the specific page URLs for top-level menu pages worked. The search function does not work in the archived version. This site continues to have accessibility issues when viewed in
Google Chrome, but works well when viewed using the Mozilla Firefox browser with the Archive-It Proxy Mode extension (Archive-It, 2017a).

Project Takeaways

While this project required activities throughout the workflow of web archiving—appraisal and selection, acquisition, organization and storage, description and access (Niu, 2012)—the acquisition phase of the workflow made it clear to us that the Internet Archive has it right: “The Web Is a Mess” (Library of Congress, 2010).

There are many reasons this mess exists. One of the great qualities of the web is its democratic nature in which virtually (no pun intended) everyone can publish content in many different forms (Costa, Gomes, & Silva, 2017). We found that this was true for our seed websites - they are deceptively complicated and complex interactive resources. Even those sites that seemed simple were in fact composed of a variety of files of varying format. Unfortunately, this can lead to large quantities of information being lost every day. One deleted or misplaced file can break pages, and conflicts due to file format, web publishing platform, or browser updates can result in unanticipated changes to websites. As of August 2019, there were 1,706,119,317 websites online (Scott, 2019). However, “80% of web pages are not available in their original form after 1 year; 13% of web references in scholarly articles disappear after 27 months; 11% of social media resources, such as the ones posted on Twitter, are lost after 1 year” (Costa, Gomes, & Silva, 2017). There is a sense of urgency in the recognition of the need to face the challenging task of web archiving and harvesting web content.

Crawling as an acquisition method to harvest content from websites is inherently messy. The software used to build websites can impact archivability, with some platforms proving
particularly messy in their design—at least from a source code and crawler standpoint. Crawlers have limitations based on site design, resulting in significantly inconsistent acquisition of websites. This experience was illuminating in how something as simple as a robots.txt file could render an entire website uncrawlable. In this way, crawlers can be blocked so a website can’t be archived at all, or portions of websites may be inaccessible to crawlers by intention or simple lack of awareness by web designers.

Website builders incorporate robots.txt into their websites for a number of reasons. While initially the use of robots.txt might seem somewhat adversarial to well-meaning archivists, robots.txt files can serve helpful and good purposes. Robots.txt work to keep crawlers from: content that is not complete or ready; private or sensitive content; dynamic content that is often changing or will be obsolete in a short period of time; and duplicate content (Moline, 2017). All of these can negatively impact a website’s SEO and permit content not meant for search engine databases to be exposed (Moline, 2017).

The challenges crawlers face make it virtually impossible for an archivist to initiate a crawl that results in a flawless website archive. The archiving process requires a great deal of human intervention at the seed level, from adjusting the initial settings for each URL to reviewing the crawl reports and making adjustments—adjustments that require research, experimentation, and multiple test crawls.

Websites are ephemeral, at-risk born-digital content that must be preserved and archived in order to available to present and future researchers (Library of Congress, n.d.-b). However, archiving and making websites accessible, even simply as a single snapshot in time, can require a prohibitive amount of technological and human resources. While our group project was less
automated than major institutional archiving programs, the experience did clearly illuminate many of these challenges.

**Please see Urban Farming Collection Video Presentation at:**

https://drive.google.com/open?id=1jt7QWtABIahhcwRVrx4uLszSk5KI-vCk

Video will also be uploaded along with this paper.

Thanks for everything, Alyce. It’s been a pleasure!
References


