Journal of Western Archives

Volume 8
Issue 2 *Web Archiving Special Issue*

Article 2

2017

# Case Study: Washington and Lee's First Year Using Archive-It

Alston B. Cobourn
*Texas A&M-Corpus Christi*, alston.cobourn@tamucc.edu

# Case Study: Washington and Lee's First Year Using Archive-It

Alston B. Cobourn

**ABSTRACT**

This case study reviews the planning, methodology, and lessons learned during the first year of Washington and Lee University Library's Web archiving program.

## Introduction

On July 1, 2015 the Washington and Lee University Library in Lexington, Virginia, initiated a one-year subscription to the Archive-It Web crawling service offered by the Internet Archive. Some classes of undergraduate students on our campus had been creating website content previously, however the launch of a Digital Humanities initiative approximately three years ago spurred increasingly complicated projects that included interactive data visualizations, such as mapping, timelines, and 3-D models. The library recognized a need to collect and preserve these interactive Web-based scholarships created by our students as well as Web content that documented W&L's history, such as its policies and campus special events, as a part of our Special Collections and Archives. Therefore, we decided to see if Archive-It could help us achieve this goal.

Project staff consisted of me and one support staff member, each of us adding this project to our existing workload. No specific percentage of our time was designated to the project at its outset, so it was our task to juggle priorities accordingly. Having no prior experience in this arena, both we and the library administration were unsure what would be needed.

Prior to signing our subscription agreement, we identified and met with parties around campus that we believed might see the value of such a service and have a desire to archive specific content. Their input also helped shape our capture priorities. Under the terms of our subscription, we would be able to save approximately a quarter terabyte (TB) of data.

## Methodology

### Identifying Sites and Sending Opt-Out Notifications

Most course websites set up through the university are created on a single WordPress server and have the URL base of academic.wlu.edu. This server is managed by the University Library, but the administration of user accounts and course sites is handled by Information Technology Services (ITS), a separate entity serving the entire campus community. I began by setting up an arrangement with ITS for them to send the Digital Scholarship Librarian (me) an initial list of all the website URLs existing on this W&L server and to send a subsequent list of newly created sites at the end of each school term. Using the Web publishing platform Omeka, which is designed to enable users to create collections and exhibits, the University Library also creates instances for classes and students as needed; from these we identified several such websites for capture.

I reviewed all WordPress sites on the initial list received from ITS and the identified Omeka sites to make sure each fit our collecting scope. A few, such as faculty personal pages, were beyond the project's scope. For each one determined to be a publicly available website created by students as part of their coursework, the support staff member sent an email with pre-determined text to the faculty members who had taught the course or the individual student who had created the site, as in the case of honors theses. The email informed the faculty or student of the Library's intention to crawl the specific website and notified them that questions, concerns, and the desire to opt out should be directed to me.

Crafting the text of the email (Appendix A), was a collaborative effort between myself and several other library colleagues. We decided to adopt an opt-out policy because we felt this to be the most practical approach. It seemed highly likely that many faculty members would not respond to opt-in notifications because doing so would be another task, likely one viewed as relatively less important, on their already long to-do lists. Additionally, since we only intended to crawl content that was already publicly available, we did not feel that there were ethical complications to doing so without permission. We decided not to pursue crawling access-restricted sites during this first year since doing so would create the need to work through ethical and additional logistical issues, both of which would mean that the crawling of restricted sites must be opt-in. Also, pursuing access-restricted sites had the potential to increase the volume of sites needing to be captured, and we did not want to overcommit our limited data budget and staff resources. We agreed to revisit this policy periodically and potentially take on this additional work at a later date.

### Archive-It Setup and Crawling

As a first step, I created a Digital Scholarship collection in the Archive-It administrative interface and added the selected sites' URLs as "seeds" within that collection. Before crawling a seed, I reviewed the live site to see if it contained any URLs with different domain names. If so, I expanded its crawl scope so that those

URLs would be captured as well. Then, I ran a test crawl on the seed since the data captured as part of a test crawl did not count against our data budget until the crawl results were saved. Therefore, test crawls enabled me to review a crawl's results and avoid spending our data budget on unwanted or incomplete data. I recommend using the test crawl functionality as a best practice when capturing a seed for the first time.

Archive-It allows users to set limits on the number of documents and/or data captured during crawls and test crawls as well as time limits for their duration. I did not set a document or data limit because I was curious what these quantities would be if unrestricted and knew I could run a second test crawl with restrictions if the results of the first showed this to be necessary. I changed the default time limit from one hour to one day, again hoping to allow for capture of the whole site. One day proved to be sufficient time for Archive-It to finish a complete capture of almost every seed I tested. In the few cases where the crawl reached this time limit before capturing all the site's content, I deleted the test crawl data, and ran a second test crawl on the seed, increasing the time limit by another day. This was enough additional time to allow for complete site capture in all cases.

After receiving email notification from Archive-It that a test crawl had finished, which was an optional feature I enabled, I waited 24 hours before viewing the resulting Crawl Report in the administrative interface to allow the Wayback Machine link it contained to begin working. This embedded Wayback Machine functionality was my main way of reviewing the results and spotting instances where embedded links or URLs with different domain names had not been captured. If the test crawl revealed these issues, I deleted it, added the missing URLs into the seed's crawl scope, and then ran another test crawl. I also reviewed the Crawl Report's list of documents captured to determine if content from undesired URLs had also been captured. If the test crawl results did not reveal either of these issues, then I saved the test crawl data.

There was one instance in which the list of documents captured for one seed with an academic.wlu.edu base revealed that URLs from other seeds also containing an academic.wlu.edu base had been captured for no apparent reason. I investigated this problem with our Director of Library Technology. We suspected it was not an issue with the Archive-It tool but rather something having to do with the site's configuration on our server. We determined that the issue was caused by a plug-in that was active on particular academic.wlu.edu sites. I informed my ITS colleague about this finding, and we agreed that the specific plug-in should not be activated on other sites in the future.

Once the crawl data was saved, I waited an additional 24 hours before viewing the website via our public webpage on the Archive-It site (https://archive-it.org/organizations/998) while also being logged into the administrative interface on another browser window. Being logged in allowed me to enable the Quality Assurance functionality on the public webpage, through which I could conduct patch crawls on missing images, videos, and other interactive elements as needed. I revisited the crawl results on our public webpage at least a day after conducting patch

crawls to verify that they had captured the desired additional content. Occasionally a second round of patch crawling was required.

Online student publications, such as literary magazines and the student newspaper, were also captured as part of our Digital Scholarship collection. The process was the same as above except that no opt-out notifications were sent prior to website capture.

I created two other collections in Archive-It: University History and University Administrative Policies. The University Administrative Policies collection contained five portions of the W&L website that were identified for capture since they documented important institutional policies: General Counsel, Human Resources, Business Office, Office of the Dean, and Office of the Provost. The University History collection contained two portions of the W&L website and one external site that documented campus events, initiatives, and groups: Sigma Pi Sigma, President's Office Timeline of African Americans at W&L, and Mock Convention. The crawling process for these sites was the same as for student publications.

We created a master spreadsheet where the support staff member recorded to whom she had sent notification emails, and I recorded where I was in the crawling and patching process for each seed. Overall, Archive-It did a great job of capturing the content we desired, however it was unable to capture some kinds of dynamic content.

## Metadata creation

The support staff member and I worked with our Metadata Librarian to develop standards for the descriptive metadata we recorded for each crawled site. Archive-It uses the Dublin Core metadata standard in which all fields are repeatable. We added one custom field, Methodology, to the Digital Scholarship collection in which we recorded the kind of research or digital humanities methodologies used in each project. We created a controlled vocabulary list of data values for this field, which was also used in the categorization of projects highlighted on the W&L's Digital Humanities website (https://digitalhumanities.wlu.edu/). All of the fields, except for "Description", create limiting facets on our public webpage on the Archive-It site. In developing our policies, we were guided by the desire to provide metadata in quantity and quality that would be useful to both users and ourselves without expending more precious staff time on metadata creation than truly necessary.

| Field | Contains | Example |
|---|---|---|
| Title | Title of the website as transcribed | The Washington and Lee University Colonnade: The Evolution of the Face of W&L |
| Description | Summary of the site's purpose and content; includes the term during which the website was created, who it was created by, and the name and title of instructor(s) or advisor (s) as applicable | This website was created during Fall Term 2013 by students in ANTH 180, Discovering W&L's Origins Using Historical Archaeology. The course was taught by Donald Gaylord, Research Archaeologist and Instructor of Anthropology. |
| Subject | Word or phrase describing the topic of the site; corresponds to the academic subject(s) reflected in the course designation or student's | Anthropology |
| Creator | Course designation of responsible class or name of responsible individual | ANTH 180 class |
| Methodology | The methodology or methodologies employed in the project; controlled value of either *blog*, *curation*, *data visualization*, *digital edition*, *exhibit*, *mapping*, *storytelling*, *text analysis*, *timeline*, *digital poster*, or *3-D modeling* | Data visualization |
| Collector | Entity responsible for crawling this website and adding it to this collection | Washington and Lee University, James G. Leyburn Library |

Table 1. Seed-level Metadata Guidelines for Digital Scholarship Collection

| Field | Contains | Example |
|---|---|---|
| Title | Title of the website as transcribed; occasionally title was supplied | Office of Human Resources: Washington and Lee University |
| Description<br>*[Note: Field only used for University History collection]* | Brief description of the site and in some cases information about its creation | This timeline is being developed by a special working group established in August 2013 by Washington and Lee University President Kenneth P. Ruscio to explore the role of African Americans to the history of the University. |
| Subject | Word or phrase describing the website's function and content | Administration |
| Creator | Campus office responsible for the content of the website | Office of Human Resources, Washington and Lee University |
| Collector | Entity responsible for crawling this website and adding it to this collection | Washington and Lee University, James G. Leyburn Library |

Table 2. Seed-Level Metadata Guidelines for University Administrative Policies and University History Collections

In spring 2016, the support staff member created metadata for the Digital Scholarship collection in a spreadsheet in accordance with the guidelines above. I then proofed this metadata and made minor adjustments or additions as needed. Lastly, I saved a copy of the spreadsheet as an Open Document System (.ods) file since Archive-It accepts this file format instead of Excel files and uploaded it via the administrative interface. The support staff member then added metadata to the individual seeds in the University Administrative Policies and University History collections through the administrative interface. We determined this method to be as efficient as the spreadsheet method in these cases because both collections contained so few seeds.

## Providing access

All Archive-It subscribers have a public interface via the archive-it.org website. Our Metadata Librarian, Head of Special Collections and Archives, and I sought to provide additional access to these archived websites in a way that would be more

findable for our Special Collections and Archives patrons. We decided to link our public page on the Archive-It site to the Digital Collections section of our Special Collections and Archives webpage.

Soon after the end of our first subscription year, I determined that both Archive-It and DSpace, which is the digital repository platform we use for our institutional repository and provides access to born-digital and digitized Special Collections materials, are compliant with OAI-PHM (Open Archives Initiative-Protocol for Metadata Harvesting). This meant that it would be possible to have Archive-It "talk" to and share its seed metadata with DSpace. Washington and Lee uses the Serials Solutions discovery service Summon, which is also OAI-PMH compliant. Serials Solutions regularly uses its OAI-PMH harvester to pull content and related metadata into our Summon discoveries from both our Millenium ILS and DSpace. I was able to easily enable the OAI-PMH functionality in Archive-It by checking a box on each collection-level metadata page within the administrative interface. Then, I provided DSpace with information about the Archive-It collections from which I wanted it to harvest content. Because all three systems (Archive-It, DSpace, and Summon) were OAI-PMH compliant, it was very easy for us to get maximum return on the time invested in crawling the websites and creating their metadata in Archive-It by adding two additional discovery and access points for the content. These additional access points made the archived websites truly feel like part of the library's collection and ensured that they would not exist as an isolated, hard-to-discover pocket.

## What Was (and Was Not) Captured

### Digital Scholarship collection

In total I captured 105 websites in the Digital Scholarship collection during the course of the year. Archive-It provides the option to arrange the seeds within a collection into Groups, which creates another limiting facet on the public website. I arranged these seeds into three Groups: Course Projects, Honors Theses, and Student Publications. I captured 96 websites that were created by a course. There were eight additional websites created that I was unable to capture before the end of the subscription year. As explained above, one site's test crawl erroneously included content from other sites, which greatly increased the quantity of data gathered. Therefore, I decided not to save this data and to attempt another test crawl only after we identified and determined how to fix this issue. I simply ran out of time to capture the other seven sites, all of which were created during the spring term.

I captured one website from the previous academic year that was created as an honors thesis using the Omeka platform and contained a digital collection and exhibit. One additional honors thesis was created this academic year by a journalism student and used the WordPress platform. I ran out of time to capture this site before the subscription cycle ended since it was not completed until the end of the school year. I also captured four student publications, two of which are in the style of a newspaper and have content added every week. I captured both of those publications

several times over the course of the year with the goal of determining how frequently we would be able to capture them in the coming year.

Many of the Digital Scholarship websites contained embedded audio and video content. Archive-It was very successful at capturing this content although many sites needed some patch crawling to capture it all. Archive-It was typically able to capture embedded data visualizations such as word clouds.

Archive-It did however have trouble capturing some other kinds of content in the Digital Scholarship projects. One course site contained embedded concept maps that students had created using Prezi, and Archive-It was unable to capture this content. Through patch crawling Archive-It was able to capture the still images of embedded 3 -D models, it was not able to capture their rotation functionality. In some instances, it had trouble capturing timelines, but this depended on the platform being used. Early in the subscription cycle it was able to capture timelines created using TimelineJS, but this software tool changed its API during the year and after that point Archive-It was unable to capture content created on this platform. A fair number of course sites have included timeline projects created using a locally developed software tool, and Archive-It was unable to capture this content as well.

Mapping projects proved to be the most difficult for Archive-It to capture. Archive-It was unable to capture any of the mapping content we crawled regardless of platform used. We attempted to capture content created using Omeka's Neatline plug-in, Google Maps, StorymapJS, and a locally developed mapping application. I inquired of Archive-It support staff about this issue and received confirmation that this kind of content is the most challenging for their Web crawler and that they had not yet developed a solution. This was disappointing because many of the sites in this collection contained a mapping component.

### University Administrative Policies collection

Archive-It had no issues capturing the five sites in the University Administrative Policies collection, which contained text, images, embedded videos, and links to PDF documents. For several of these sites, I did have to expand the crawl scope after running an initial test crawl so that the linked PDFs, which had different base URLs, would also be captured.

### University History collection

In the University History collection, we chose to capture the Timeline of African Americans at W&L and Mock Convention websites because both document historic events in the life of William and Lee University. We also captured the Sigma Pi Sigma site, which I became aware of because it was included in the initial list of WordPress sites I received from ITS. I chose to crawl it because I judged the histories of the W&L's fraternities and sororities to fall within scope.

Regarding the timeline, Archive-It was unable to capture the embedded SoundCloud audio clip and linked PDF documents. I was unable to remedy either issue through patch crawling. I was however able to capture the linked PDFs on a second test crawl by expanding the original crawl scope but was unable to capture the SoundCloud audio clip through this method.

### Summary

In total I captured 64.2 GB of data during the subscription cycle. The vast majority of this data was captured as a part of the Digital Scholarship collection. Archive-It is not the perfect solution for collecting, preserving, and providing access to website content and functionality, but it has a relatively high rate of capture success, is easy to use, and the company is actively working on improvements. Therefore, we chose to renew our subscription at the same level. Archive-It rolled over 25% of our unused data from 2015-2016 to our 2016-2017 subscription cycle.

## Lessons Learned

There were several key takeaways from this first year that will enable us to expand and improve during the project's second year. First, we found it useful to always try to anticipate additional uses for collected metadata and create it in such a way that would enable those uses. We planned for our Archive-It seed metadata to live only in the Archive-It system and be visible to the world via our public Archive-It webpage. We hoped to expand access in the future but did not give quite enough thought as to what those specific possibilities might be and how they might be enabled or inhibited by our current metadata decisions. We recognized that the metadata would be able to be repurposed to many systems easily because it utilized the widely-used Dublin Core schema, and our certainty in this, to a certain extent, may have caused us to pay less attention to the format of the content within these fields. Therefore, once we realized it was possible to add the Archive-It seed-level metadata to our DSpace repository and Summon discovery layer via OAI-PMH, we needed to modify some guidelines and change some existing data accordingly.

For individual person website creators, we did not originally specify a standard format for recording names in the Creator field. We are modifying that guideline to specify that names should be formatted as last, first in accordance with the Library of Congress name authority format so that this metadata will be aligned with the rest of the Creator field metadata in the Digital Archive and Summon discovery layer.

We realized the format of the content values in the Creator and Collector fields were opposite and therefore specified in our policy that they should have parallel structure with the name of the highest level of the issuing body (i.e., Washington and Lee University) always first. This format is the same as that used in the Creator field of the existing Digital Archive data.

Technically, most of the content we seek to capture belongs to a University Archives record group. The integration of the metadata about this content into the

Digital Archive, which contains both manuscript and record group materials, caused us to begin recording what record group each seed belongs to in the Relation field.

We also modified our guidelines for the Subject field to specify that all content values should follow existing FAST subject headings because the majority of the Subject data already in the Digital Archive complies with this standard, which is designed to enable many linked data possibilities in the future. Many of the existing values, such as Biology, did not need to be changed, however some did. For example, Administration was changed to Universities and Colleges—Administration.

Second, Web archiving using Archive-It is not particularly difficult, but it can be time consuming, particularly when crawling seeds for the first time. This is largely because, as the methodology above details, many sites need to be reviewed several times during the crawl process in order to ensure that as much of the desired content is captured as possible. The initial website assessment step could be skipped for future crawls of a seed, which would decrease the time needed, but this would likely introduce the chance for content added during the expansion of a site's scope to be missed during a crawl. Individual institutions must decide how much reviewing, scoping, and patch crawling makes sense for them based on available staff time, desired outcomes, and ranking of priorities. The crawling process is flexible in this way, but the decisions made may affect your results substantially. I discovered that I really needed to devote a little time to Archive-It each workday to keep moving forward with the crawling, reviewing, and patching. However, many weeks this was not possible because of the many other hats I wear as a staff member of a small liberal arts college library. Ultimately, I was able to carve out some time each week during the subscription cycle for the project although it was usually several hours either on a single day or divided between two.

The support staff member's time on the project was dedicated solely to metadata creation, and therefore her efforts were concentrated within a large portion of her time for several weeks in a row during June. I do think it was important for me, as the professional librarian/archivist, to do all of the crawling, scoping, reviewing, and patching work during the first year because it gave me the opportunity to determine through iterative efforts the best process for executing these activities. However, I do not think there is anything about the nature of that work itself that should prevent a support staff member from assuming responsibility for it after such an initial period and appropriate training. From a sustainability standpoint, transferring such work to a support staff member once local workflows and policies have been established probably is the best course of action for institutions that do not have staff dedicated solely to Web archiving. Institutions with dedicated staff will likely be positioned to go further faster; they may be able to collect more content and tackle additional preservation concerns related to WARC files and metadata sooner.

Third, and relatedly, adding a rule to a crawl scope that tells the crawler to ignore the robots.txt files it encounters can save administrative time by decreasing the amount of patch crawling needed. For example, the first time I test crawled one of the

Omeka sites, I discovered that none of the site's pictures were captured. I reviewed every page of the site and ran approximately 30 patch crawls to capture the pictures. While initiating the patch crawls, I noticed that Archive-It said the pictures had not been captured because they had been blocked by a robots.txt file. Therefore, when the same thing happened during the test crawl on a second Omeka site, I decided not to save the test crawl data, added an "ignore robots.txt" rule to the crawl scope, and ran a second test crawl, which resulted in the capture of all the pictures. This second approach was easier and less time consuming than the first. I mention this because it could prove useful in certain situations, not to recommend it as a best practice. Since website platforms and authors create robots.txt files for the purpose of preventing certain website files from being crawled, one must consider the ethics of the particular instance when deciding whether or not to ignore a robots.txt file.

Fourth, your data budget will likely go further than you expect. I worried too much about going over data budget and as a result, went under. I wanted to make sure I saved enough room for all the student scholarship that would be created during the winter and spring terms. Therefore, I waited to capture additional University history content until after all of the student projects were captured. The month and a half between the end of the spring term and the end of our subscription cycle did not provide enough time to do so. I will be certain to capture more W&L history regularly throughout our current subscription cycle to avoid repeating this outcome. I will save the data captured by a test crawl unless it is an inordinately large amount instead of hesitating on moderate amounts of data.

## Second Year Plans

We are actively identifying more content for capture. I have reviewed Washington and Lee's list of student organizations and identified those with a Web presence. Last year the library began a Mellon Digital Humanities Undergraduate Fellows program in conjunction with a recently awarded Andrew W. Mellon Foundation grant. We plan to begin capturing the projects created by these students. Based on the amount of data collected during my previous captures of the two student newspaper-type sites, I believe it should be possible to capture both bi-weekly during the next academic year. We are also identifying additional sections of W&L's website and social media presence for capture.

We will consider the possibility of engaging librarians with departmental liaison responsibilities in recommending websites for capture. Several institutions shared their experiences with this approach at the August 2016 Archive-It Users Group meeting in Atlanta.

We currently save a copy of our seed-level metadata locally but would like to investigate additional preservation measures during the coming year. Archive-It is involved in the WASAPI project to develop various APIs, which will aid in data export and exchange between their tool and other systems. Eventually we plan to investigate the possibilities this project provides for easy downloading of WARC files.

## Conclusion

Overall I am very pleased with the accomplishments of our first year using Archive-It and am excited about growing the program next year. Archive-It has given Washington and Lee University the ability to ensure that valuable institutional history and student scholarship will persist into the future with relatively little monetary commitment. Archive-It's technical support staff was very responsive to questions, and the administrative interface had a small learning curve. For these reasons, I believe all institutions could feasibly manage and benefit from an Archive-It program.

# Appendix A

Dear [name of faculty member or student],

The University Library is archiving Digital Humanities projects using a service provided by the Internet Archive called Archive-It. Archived websites will be publicly available at https://www.archive-it.org/collections/6143. Please contact Alston Cobourn, Digital Scholarship Librarian, if you have any questions or concerns, or if you wish to opt your site out of this service. If your website is public, it may be crawled by external services like the Internet Archive/Wayback Machine even if you opt out of this library service. You are receiving this email in regards to [URL(s)].

Sincerely,

Support staff member's signature