Aly DesRochers
Bonnie Gordon
Nicole Greenhouse

Barnard College Alumnae Web Crawl

Working with staff at the Barnard College Archives, we created an Archive-It web crawl of selected Barnard College web pages and social media sites. Shannon O'Neill, Archivist, and Martha Tenney, Digital Archivist, at the Barnard College Archives identified multiple web pages, a Facebook account, a Twitter account, and YouTube content related to Barnard College alumnae to be captured. These pages were chosen because they contain a wide variety of content and will provide a broad picture of Archive-It's capabilities. The Barnard College Archives has allocated funds to acquire their own Archive-It account and begin web archiving within the year, so this select content group will serve as a test case to help them plan for and prepare active crawling.

After our initial discussion with Shannon and Martha, they proposed seed urls and we determined the precise seed format and the scope and frequency of the crawls. There are two Barnard Alumnae homepages (alum.barnard.edu and barnard.edu/alumnae), and they requested that we crawl both to determine whether some content is shared between the two pages and what might be unique about each page. Barnard Magazine is also made available on two separate web pages. We crawled both barnard.edu/magazine and the magazine index at alum.barnard.edu, as the latter is rendered using the Issuu publishing platform, and Shannon and Martha were curious to see how well Archive-It would be able to capture this content. The Barnard Alumnae Facebook Page and Barnard Alumnae Twitter were selected to test Archive-It's ability to capture these social media sites and to figure out how to format the seeds in order to get the best results. The Barnard Alumnae YouTube Channel was included to see whether Archive-It can successfully capture video content. (We later changed the crawl to capture individual YouTube video pages in an attempt to successfully capture the video content -- see below.) We set the scope of the crawls to the default setting, wherein websites that were not part of our seed urls were not crawled.

Shannon and Martha suggested that the social media sites (Twitter, YouTube, and Facebook) should be crawled weekly and the other sites should be crawled monthly. We agreed that monthly crawls would be sufficient to capture the Alumnae Affairs and Barnard Magazine web pages, and as the Facebook and Twitter pages are updated more frequently, weekly crawls would be appropriate. Initially, we were going to set the YouTube channel page to crawl weekly as they had suggested. However, the need to crawl each page individually altered our decision. Upon exploring a sampling of Barnard Alumnae YouTube videos, we noted that there were no comments or commenting had been disabled, so the main purpose of the crawl was to archive the Barnard-generated content (videos) only. Thus it was decided that the individual YouTube pages would only be crawled once. Though YouTube's stylesheet may change, we do not think it is necessary to recapture for those alterations alone.

Once the seed urls, scope, and frequency had been defined, we scheduled a test crawl to evaluate these settings. The test ran for three days and three hours and crawled 2.9 GB of data. From the test's report, we discovered that the Barnard Alumnae Facebook Page, and the Barnard Alumnae Affairs YouTube Channel did not initially crawl due to robots.txt exclusions. We researched Archive-It's tools to overcome

robots.txt exclusions, and contacted Lori Donovan, Archive-It Partner Specialist, to enable the robots.txt override. To ignore robots.txt, we had to modify the crawl scope under the "host constraints" tab. We added www.facebook.com, www.youtube.com, youtube.com, and ytimg.com to be ignored. (ytimg.com is the URL for the YouTube stylesheet -- by overiding this url's robots.txt exclusion, Wayback can render the site properly.) We also altered the YouTube seeds because each video's watch page is not directly nested under the Barnard Alumnae channel page. (For example, the URL for the Barnard Alumnae page is http://www.youtube.com/user/AlumnaeAffairs, while each video's URL is structured http://www.youtube.com/watch?v=xxxxxx.) Because they are in a different directory, each video must be archived as a separate seed URL, so we instead included the unique watch page urls for four of the individual videos. We also scoped the YouTube pages as a "one page only" crawl, to limit the amount of documents archived so that we did not get videos we did not want.

After making these changes, we started our actual crawls. The one-time YouTube watch page crawls completed in six and a half hours, the monthly crawls of the Alumnae Affairs and Barnard Magazine pages completed in two days and 13 hours, and the weekly crawls of Facebook and Twitter stopped when they reached the time limit of 3 days. In total, 12.6 GB of data was captured, though only 4.4 MB of data was captured from the YouTube pages because of additional problems. We then shared the completed crawls with Shannon and Martha and they provided feedback and suggested modifications. Many of their comments reflected our observations. There were some minor issues with the stylesheet captures for alum.barnard.edu and for barnard.edu/alumnae. While these stylesheet issues affect the look of the pages, they do not interfere with viewing or reading content. Because capturing content was Barnard's main goal, they saw this as a minor issue. They also noted that there some variations between captured pages in different browsers and cited the Issuu plugin as an example.

We also observed that, even after disabling the robots.txt block, the Barnard Alumnae Facebook page crawl did not work. Shannon and Martha suggested a possible solution: entering login information into the crawl settings so that it acts as if a logged in user is making requests to the page. Another possible issue with the Facebook crawl that they identified was a robots.txt exclusion. They asked, "is Facebook a lost cause for us?" Their first suspicion was correct. The Alumnae Facebook page's content was password protected, meaning that if one is not logged into Facebook, content cannot be viewed publicly. To work around this issue, we had to manually add a username and password to archive the content. After entering user credentials and overriding a few additional Facebook URLs, akamaihd.net and fbcdn.net (which serve the stylesheet and Javascript, making Facebook renderable by the Wayback Machine), we were able to successfully crawl the Facebook page, though past content (beyond the scroll depth of the page) is not included. This is possibly due to the fact that the crawl timed after reaching our designated three day time limit.

Interestingly, we used Nicole's Facebook username and password to crawl the site, and she was sent a suspicious activities email from Facebook and had her account temporarily locked. When she attempted to log in to Facebook, it asked her if the login attempt from San Francisco, CA, where the Archive-It offices are located, at the crawl's

start time was made by her.[1] It is good to know about this possible issue, so the user whose credentials are used to perform the crawl does not fear a hack of their account.

After the second crawl on the YouTube pages, no one in our group was able to play the YouTube videos from the pages we crawled, though Shannon and Martha were able to view some of them. We attempted to replicate this on several computers with different operating systems and browsers, but could not successfully play any YouTube videos. Shannon and Martha were able to view videos both contained in the pages we crawled individually,[2] and at least one video that was linked to from the top page, but for which we did not set up an individual crawl.[3]

After a final altered crawl, when disabling the robots.txt exclusion of just youtube.com, we were successful in archiving videos. (The first YouTube crawls overrid www.youtube.com and youtube.com, which was incorrect, so we needed to remove www.youtube.com as a host to be overriden.) According to Lori, "the video files themselves are stored on subdomains of youtube.com and are blocked by robots.txt so you need to ignore robots.txt for that base domain in order to capture the video files." Lori was correct, but the crawl was larger than expected; Heritrix ended up crawling not just the videos we commanded, but also other Barnard Alumnae videos from the site. It seems that it archived some videos from the channel site, but did not render them like the live site. Instead, the site takes on the appearance of a site index, meaning that there is still likely to be some robots.txt exclusions blocking. More testing will probably have to be done to determine what videos still have not been crawled from the Barnard Alumnae channel.

Some sites that Shannon and Martha would have liked to have captured were not crawled because they were out of scope. This problem was most pronounced on the Barnard Magazine site. The url of the main page is barnard.edu/magazine. The main page only shows abstracts of articles in the current issue of Barnard Magazine; in order to read the full article the user must click on the link. The Barnard Magazine url is http://barnard.edu/magazine. These links bring the user to the article on the News & Events site. An example url is http://barnard.edu/news/building-base. On the left side of the pages are links to past issues: an example url is http://barnard.edu/category/listing-filter/fall-2011. Thus, because the articles are in the barnard.edu/news directory and the past issues are in the barnard.edu/category/listing-filter directory, all of these pages did not crawl based off of our http://barnard.edu/magazine seed. They suggested adjusting the crawl for greater depth on this page, as they would like to be able to access the linked articles and past issues, but we did not have sufficient time to test this as a solution.

Shannon and Martha were also interested in archiving more material on the Barnard Alumnae Twitter page. Although our crawl of the twitter feed was successful, it

---

[1] Image 1.

[2] http://wayback.archive-it.org/3999/20131025162547/http://www.youtube.com/watch?v=01pKTrgixZo.

[3]

[3]http://wayback.archive-it.org/3999/20131025163019/http://www.youtube.com/watch?v=OXIXhZEJNrQ&feature=c4-overview-vl&list=PL7EDDEE4CEB2D463A.

only archived content going back to June 10, 2011, but the account contains tweets as early as July 7, 2009. Under the default scope settings, Heritrix can only crawl as far back as two scrolls down to the bottom of the page (which enable Twitter to display previous content). To archive past two scroll downs, we had to expand the scope rules by adding a SURT (Sort-friendly URI Reordering Transform) rule to tell Heritrix to archive more documents. We were then able to capture additional past content, though some of the later content has rendering issues -- asiatic language characters appear in lieu of dates and certain Twitter buttons, though all content of the tweets appears to be intact.

This was a very useful activity to conduct with an organization that is planning to begin actively web archiving. We demonstrated the difficulty of archiving social media sites and how to archive them effectively, so they will not have to recreate the multiple test crawls that we performed. They could also see how much data had been archived and started to get a sense of how many gigabytes they should budget for before beginning their crawls. Shannon and Martha were ideal partners; they are archivists with some background on web archiving and ended up asking the same questions about problems and difficulties that we saw.

However, this web archiving experience made several sustainability issues plain to us. As this was a class project and our Archive-It subscription is administered by NYU faculty, we will not be able to transfer management of our crawls to our curators. They will not be able to continue to check the existing crawls for quality assurance, add new seeds (in our case, for additional YouTube videos), create metadata, or perform any of the other necessary maintenance tasks. Though Shannon and Martha will be able to use our findings to establish their own web archiving program, sustainability is a larger issue when using a third party program like Archive-It to perform web archiving. While Archive-It is largely automated after seeds have been defined and initial tests have been performed, it still requires frequent human input to continue functioning properly. Based on the significant time we spent working on the backend of Archive-It, an institution intending to establish their own web archiving project would have to allot a large portion of one (or more) staff member's time (or hire a new dedicated staff member). That person would have to set up crawls, determine the scope and frequency of seeds, what the seed URLs would be, create metadata for the sites, perform periodic quality assurance checks, and possibly develop further access and description methods. Additionally, in our case intellectual property and copyright were not issues because we were archiving Barnard College's own content, for other projects, asking permission to archive could also mean dealing with a legal department or General Counsel to determine a policy. Therefore, web archiving requires a commitment of resources and adequate planning and preparation in order to be performed properly and over the long term.

Aly DesRochers
Bonnie Gordon
Nicole Greenhouse

Image 1: