Jasmyn Castro, Carmel Curtis, Ethan Gates
CINE-GT 1807
10/28/14

# Web Archiving the Educational and Community Outreach of "Nas: Time Is Illmatic"

For our group web archiving project, we decided to contact Martha Diaz, founder of the NYU Hip Hop Education Center, associate producer of the recent documentary "Nas: Time Is Illmatic," and our fellow student last year in the MIAP program. We suspected, given Martha's interest in employing archival materials from the hip hop community in an educational setting, that she would be excited to learn about the Archive-It tool and the potential for expanding her efforts to preserve hip hop culture into web archiving. In a Skype conversation with Martha and Carlos Mare, an artist, teacher and scholar who also works at the Hip Hop Education Center and has been assisting the "Nas: Time Is Illmatic" production team with designing and managing their website, we explained in detail the Archive-It tool and the guidelines of this project.

Enthusiastic about the idea, Martha and Carlos were immediately and primarily concerned with the content contained under the "Education" tab of the "Nas: Time Is Illmatic" site (www.timeisillmatic.com/education/). While this page is still under development, and is ultimately meant to be used as a resource to offer educational activities and resources related to the making of Nas' groundbreaking album "Illmatic" to teachers, student and scholars, at the moment the primary contents (and the materials Martha most desired to preserve) were linked images from the film's Flickr account (which generally posts pictures from community screenings of the film around New York City) and an interactive timeline containing text, images and video files related to the history of hip hop and Nas' personal biography.

The priority in capturing this page, according to Martha and Carlos, was to preserve in some aspect the community and educational efforts that have gone hand in hand with the production and exhibition of "Nas: Time Is Illmatic." It became clear that our initial thoughts that we might document some of the press coverage and reviews surrounding the film itself would not quite fall within the scope that Martha was interested in curating; our focus in gathering content was not necessarily to be on the film itself, but its impact within local communities and hip hop education. Likewise, the content within the interactive timeline (including video interviews with Nas and other prominent musicians, music clips from "Time Is Illmatic" and concert footage) was notable not just in and of itself, but in its educational and interactive context within the timeline. That is, the purpose of capturing the "Education" page, rather than individually crawling the URL for each clip, was to preserve the relationship drawn between these resources by the film's production and design team.

Given that we would apparently be working primarily with one seed URL, we at first agreed with Martha that a very frequent, daily crawl would be unnecessary to sufficiently document the "Education" page; a weekly crawl seemed an appropriate rate which would be able to capture the new material posted from the film's Flickr account, as screening events have been occurring fairly regularly around the city for the past several months. At a later date, once there are fewer screenings and public events, the crawl could possibly even be adjusted down to even a monthly rate to capture updates to the interactive timeline or any of the educational resources that the "Time Is Illmatic" team planned to post.

In our first discussion with Martha and Carlos, we had some concerns regarding the amount of media contained on their requested page. Without having yet performed a test crawl, we were unsure whether capturing the page (or perhaps more accurately, capturing further URLs) would be overstepping a fair share of our class' allotted memory on Archive-It. However, Martha and Carlos were satisfied with keeping the scope and capturing policy of this crawl fairly limited for two main reasons. First, there were some mild concerns regarding intellectual property: though the educational context of the page and their personal relationships with many of the content creators made Martha and Carlos confident that they could continue to post video and audio clips (often taken from television, or from published albums) to the interactive timeline, adding these materials to an archived online collection created another layer of reproduction and access, and for the time being the two did not want to expand their web archiving efforts too far until they had a better grasp on their ability to use copyrighted material. Second, however, they were simultaneously very interested in continuing web archiving efforts beyond this project if possible, either maintaining an archive of the "Nas: Time Is Illmatic" Education page or of other resources related to hip hop culture and education. They were keen on continuing to use the Archive-It software, but in order to keep costs low do not want to use a large amount of memory. It was decided that by performing a relatively limited crawl, we would be able to most accurately serve as models/advisors for the web archiving work that Martha and Carlos could perform for themselves in the future.

After consulting with Martha on what she would like her final outcome to look like, we spent a considerable amount of time conducting several similar test crawls. The first test that we conducted exclusively looked at timeisillmatic.com/education/. As described above, this is the bulk of what Martha wanted archived so we wanted to take a close look at results for crawling this singular website. When we reviewed the report for this test we focused our attention

on what was out of scope. We wanted to make sure that things that were not being captured in this crawl were not things that we actually *did* want captured. One of the points that Martha made clear in initial conversations was that she wanted this to not only capture the documentation but to re-create the experience of being on the webpage. The educational component of "Time is Illmatic" is not one-sided but rather collaborative and interactive. After reviewing what URLs and content were out of scope in this first test, we decided to do a test crawl on timeisillmatic.com/education, without the forward slash. This way nothing would be left out, according to the Archive-It parameters. We were also considering the future use of this URL. Presently, the timeline is the primary aspect of timeisillmatic.com/education. However, when speaking with Martha she mentioned that this URL will become a host to smaller sub-topics. Using the Archive-It Help documentation for assistance, they specifically suggest *not* using a slash if "you want all subdomains of your seed url to be in scope." With this in mind, we conducted another test crawl of timeisillmatic.com/education.

After the first test crawl we also realized that we were using a much smaller amount of our group's allocated storage than we had expected. We consulted again with Martha via email, gauging her interest in possibly adding additional webpages or a social media component to the crawl. She expressed and emphasized an interest in wanting to capture the community's response to "Time is Illmatic" and specifically a response to the educational outreach conducted in conjunction with the movie. We agreed to do a test crawl of the film's official twitter handle (@illmaticmovie), which often posts announcements and documentation from screenings. This would also hopefully include responses from other users to the film's Twitter, making this crawl serve as a record of audience and community reaction to the film. We felt this was in keeping with the original "collecting" policy that we had agreed upon with Martha, preserving part of the local/community aspect of the film's release. There was some discussion of increasing the crawl to a daily rate considering the more rapid accumulation of content on a Twitter account in comparison to other web pages, but for the time being we thought that a weekly crawl would still accurately and satisfactorily reflect the account's history in a web archive. We ran a test on the Twitter URL, twitter.com/illmaticmovie, both with a forward slash and without. After reviewing the report for both instances and consulting the Archive-It help documentation, we agreed that including the forward slash did not leave out anything that was essential.

It was after reviewing the report for twitter.com/illmaticmovie that our group began to reconsider if this seed was in fact an accurate representation of the topic. The twitter account did not capture much of the educational outreach of community response to the movie. We went back to Martha and proposed that instead of capturing just their twitter account that we would also capture the hashtag #illmaticmovie, as well as the film's official Instagram account. This way, we would get a broader look at the community's voice, ensuring that online responses to the film that were not addressed directly at the official Twitter account would still be preserved. It was in this conversation that Martha also expressed interest in crawling the press reports on "Time is Illmatic" through timeisillmatic.com/press. While the press reports currently focus on the reception of the movie, as screenings stop being regularly scheduled the press reports are expected to shift towards accounts of how the film is being used for educational purposes.

We then conducted yet another test of the seeds that would end up being used in our collection: timeisillmatic.com/education, twitter.com/hashtag/illmaticmovie/, and timeisillmatic.com/press/. This test was proving to give the results that we were expecting so we stopped the test and ran the crawl.
While we had discussed with Martha only conducting a weekly crawl, we as a group ultimately decided to go with a daily crawl for our final effort. Since Martha and Carlos are hoping to continue to manage this collection after the duration of our work, and plan to use the reports generated from our group work in formulating a budget. we were concerned that a budget proposal based on reports from a weekly crawl would not sustain the growth that they are anticipating. We decided that while the number of seed URLs would remain limited, we would expand the frequency of capture to a daily crawl to capture the current maximum amount of data storage that would be needed. We feel it would be easier for them to scale back from that point than scale up.

While our test crawls were successful and easy to interpret the results, we experienced some difficulties in our final crawl. For some reason the Twitter account (twitter.com/illmaticmovie) and the Instagram account (instagram.com/illmaticmovie) were not successfully capturing. We were not getting <u>any</u> crawled documents from either of these URLs. We consulted the help documentation on "Why didn't some pages get archived?"[1], went through each of the possible reasons for why these pages weren't archiving and none of the proposed issues seemed to be our specific problem. There were no issues with parts of the site being blocked by robots.txt; the pages were linked; and it seemed unlikely that there was a connection error. For all our understanding, there was no reason for the Twitter and Instagram pages to be out of

---

1 https://webarchive.jira.com/wiki/display/ARIH/Archive-It+How-to+FAQ#Archive-ItHow-toFAQ-Whydidn'tsomepagesgetarchived?

scope. We attempted to compare our successful test crawls to our problematic final crawl. The only difference that we were able to detect was that in our test crawls an uppercase “I” in “Illmatic” was used and in our actual crawls, a lowercase “i” was used. We adjusted the casing and are waiting to see if this has an impact on the results. We have contacted Archive-It for help, consulted fellow classmates, and scanned through Archive-It help documentation and videos.

When we had finished the test crawls and thought that we had reached our final parameters, we reached back out to Martha and Carlos to share our results thus far. They seemed pleased with how manageable the amount of documents captured was and were willing to wait on our final assessment after we complete our troubleshooting. Additionally, we are working on arranging a time in the next couple of weeks when we can meet with Martha and Carlos to give them a more detailed tutorial on how to use Archive-It, so that they can take over the maintenance of this collection. While Martha and Carlos have some knowledge and experience with archiving and preservation, we want to make sure our explanation of the management of this collection is as clear and comprehensive as possible.

We also worked with Martha and Carlos on how to best describe this collection to make it searchable and accessible to interested parties. They emphasized the importance, given their intended educational use, of making this collection searchable. We added as much descriptive metadata as possible, on both the collection and URL level. We did test searches in archive.org to get a sense of common/popular keywords to judge the best and broadest possible tags, as well as possible user variations on similar terms: for example, there are different search results for Hip-Hop (8,221) and Hip Hop (19,564).

Overall, though we encountered some difficulties with the Archive-It software and will continue to investigate the specific issues encountered in our crawl had with capturing Twitter and Instagram, we consider this project a success, if only for being able to introduce Martha, Carlos, and their organization to the potential of creating curated web collections. Our discussions regarding how best to capture community responses online were fruitful, and performing these crawls gave us a better idea of the storage and monitoring practices necessary for web archiving. We hope that Martha and Carlos will indeed be able to successfully integrate Archive-It into their educational efforts in the near future.