Dan Finn; Joey Heinen; Athena Holbrook
CINE-GT 1807
29 October 2013
Prof. Kara van Malssen

Web Archiving Screen Slate

Screen Slate is an independently-run web-listing of cinematic events in and around New York City. Created in 2010 by curator and archivist Jon Dieringer, Screen Slate "aims to raise awareness and accessibility of moving image culture, stimulate exchange between the cinema and art worlds, and illuminate new creative directions." The main site is updated weekly with listings for daily screenings throughout the city. These listings often include links to venue websites, externally hosted images, and video trailers primarily hosted on Vimeo. In addition to the main site, Screen Slate has a directly related Facebook page.

After discussing the scope and frequency of a potential web crawl with Jon Dieringer, we determined that the main Screen Slate site should be crawled along with the embedded hyperlinks to other film organization sites that often feature associated calendars, images, and video trailers. In addition to this primary seed, we added the Screen Slate Facebook page as a secondary seed in order to collect documentation of community interaction and response. We initially set the frequency to "daily" for the first week of the crawl, and agreed to change this to "weekly" thereafter. Given that Screen Slate is updated multiple times throughout the week and links to external sites with regularly updated calendars and content, we feel that it is important to crawl and archive these sites at least once a week. Final decisions regarding the ultimate frequency and scope of ongoing crawls were to be determined after collecting data.

Our first crawl resulted in 357 MB of data, an amount that was fairly congruous with our expectations. The host that was most responsible for this output was also the Screen Slate website (11,086 URLs and 231 MB) which gave us further cause to believe that our seeds were set up correctly and that we were not inadvertently crawling the entire domain of a seed page, or the entire internet for that matter. Along with the html of the original web pages we also received a number of .pdf files that largely appeared to be linked flyers to various events around the city, materials that were in keeping with our desire to collect a curated glimpse of what was happening in experimental film presenting in New York City at this moment in time.

However, not everything was a success in our first attempt, largely due to

two large snares. The first was the fact that no content from the Screen Slate Facebook page was being crawled due to a “robot.txt” block. The other issue was that no video had been collected, a surprise given Screen Slate’s propensity for posting trailers and linking to other websites with video content. The fix for the blocked Facebook page came in adding “fbcdn.net” (the Facebook domain which allows for better monitoring of cookies, stylesheets, and traffic to the site) and “akamaihd.net” (the host for uploaded photo content) as a host. Once a host is added in your Archive-It crawl you can assign a number of constraints to that host such as number of pages run, blocking the host entirely, or choosing from a select number of rules for that host. Ignore robots.txt host constraints were added for hosts fbdcn.net and akamaihd.net, both recommended by Archive-It in order to capture Facebook stylesheets and more accurately capture the layout. And in order to ensure there is not an inordinate amount of material a document limit was set to 6,000. However, this fix did not completely cure our Facebook ails as the second test crawl turned up similar results with Facebook.com not appearing in the web-archived results. We then read that one has to add a SURT (System Update Readiness Tool) which allows you to include all subdomains in your hosts and their related customized settings. This also allows us to enter certain commands such that the specific Facebook style sheets will also be included in the crawl so that the related content has a more intuitive layout. By adding the Facebook SURT we were then able to ensure that the subdomain (i.e. the Screen Slate page) would have the same “block robot.txt function.” After completing these steps, the crawl ran content from the Facebook page successfully.

The second issue with the lack of web video proved to be a little more challenging than the first issue. According to support documentation on the Archive-It webpage, “we are unable to automate crawls of publicly available Vimeo content through the Archive-It service due to the complexity in how Vimeo chooses to construct video links on their webpages. We are currently working on alternate methods of capturing videos from other difficult to crawl video sites which may apply to Vimeo in the future.” Although Screen Slate posts and links to video sources on the main site and Facebook page, these files are hosted on Vimeo and cannot currently be crawled using Archive-It. In order to archive video material, we added the Spectacle Theater YouTube channel as a related seed. Although Spectacle Theater is not directly related to Screen Slate, Mr. Dieringer is an administrator and programmer of the theater and Screen Slate often links to content hosted by Spectacle’s sites. Nonetheless, even after adding this new seed directory, the next test crawl did not result in any video content. We then consulted a web help document that explained the challenges of crawling a YouTube channel as opposed to simply crawling a linked YouTube video. This is

largely due to the fact that there is no way to set limits or parameters as to where a channel ends and the rest of YouTube begins. To commence with this we changed the YouTube host by adding the “ignore robot.txt” function as well as a specific host constraint for the YouTube style sheet (ytimg.com). Next we had to set a document limit (recommended around 1000) such that content would not be crawled beyond the specific channel.

At this point the settings we developed are capturing well, and there is functionality for each seed via the Wayback Machine. The main object of the crawl, screenslate.com, is navigable via the Wayback Machine, with limited discrepancies from the look of the page when viewing the site live. There are some style issues, for instance the image for the main logo at the top of each page, which also serves as the home navigation button, is absent. The link is present but the image is not visible. The cause of this issue is uncertain and will require further discussion with Jon regarding where the image is taken from, and whether there needs to be another host added we have omitted. Another issue is in the venues section of the site. Each venue page provides an embedded Google Maps window; these are not reproduced within the first week of crawls for the page.

However other potentially complicated sections of the page are captured accurately. Many pages within Screen Slate’s site contain a sidebar calendar. Luckily this calendar application is formatted to end at the latest month with an updated schedule. Therefore the calendar does not automatically generate content infinitely through the future or past, resulting in unnecessary crawl bot traps. A sidebar Twitter feed is also captured accurately.

Our latest settings for Facebook are reproducing that page well. The last crawl from October 28th resulted in capturing posts dating back to September 23rd. Due to this and the average updates to the page, we will adjust the crawl frequency for this seed to the monthly option. Trying to navigate from the main Facebook page often leads to redirecting out of Wayback Machine into the live Internet. Therefore we cannot expect to capture more than a snapshot of the main page with every crawl. This should be adequate though since most of the relevant links redirect back to screenslate.com.

The YouTube channel is still the most problematic. The test crawls for YouTube generated a lot of irrelevant content. Even determining this was relatively difficult. The only video we have so far that can be played through Wayback Machine functionality is an .mp4 file. And this was captured from Facebook (via the fbcdn.net host). The irrelevance of the YouTube content had to be determined by retracing some of the URLs listed in the file types section of the crawl reports to the live web pages. Through this method it became clear that the crawler followed several “YouTube rabbit holes,” or the result of following the

links to suggested videos listed at the ends of watched videos. For instance, a lot of Spanish language content was captured, likely beginning from a popular YouTube series “Yo Soy German.” Dozens of CNN content was grabbed as well. In order to limit the rabbit hole effect we again adjusted the Spectacle Theater YouTube channel crawler settings. Instead of seeding the primary user page, we seeded the videos tab of that page, which lists all videos submitted by the user. We combined this adjustment with changing the seed type to News/RSS so the crawler would follow every link on the page but follow no links on subsequent pages. By reexamining the URLs grabbed on the next crawl, this method was more successful.

However functionality is still an issue. Again in order to check which videos were grabbed we were forced to trace the reported URLs to the live internet to see the content of the crawled videos (this method was only possible with .swf YouTube files, the URLs for .flv files could not be accessed). Using any of the three playback methods listed in Archive-It’s help documentation for YouTube crawling proved fruitless. The two methods of viewing via Wayback Machine include in page viewing and following a prompt in the banner on the top of the page. The first resulted in an error message and the second did not become available. Using the “Watch” feature out of Archive-It itself did not work either. At this point we are in a situation where we can see that we are capturing a lot of content we cannot access.

As a final step we have decided to limit metadata to the collection level, partly because the Dublin Core fields are, we believe, flexible enough to explain the seeds we have chosen, to describe the purpose of the project and to list its contributors. We also finalized the crawl frequencies for all three seeds; screenslate.com is weekly, and the Facebook and YouTube pages will be crawled monthly.

Ultimately the project has been successful in terms of our expectations and those of Mr. Dieringer. The main object of the crawl, screenslate.com, is reproducible in Wayback Machine with a high degree of accuracy and navigability. The ability to refer to Screen Slate’s Facebook page is also appreciated by Mr. Dieringer and contextually helpful. The functionality issues with YouTube are disappointing, but the targeted material for the crawl is not central to the overall aims of archiving the Screen Slate web pages. We have provided access links to Mr. Dieringer for all crawled seeds via the Wayback Machine, which he will continue to monitor for quality going forward.