Winnie Schwaid-Lindner
Digital Preservation - CINE-GT 1807
Due December 15, 2017

Whatcha Gonna Do with All That Junk?

Preserving Undesirable Digital Content

Introduction

Throughout the history of the public internet, content has been created for users to view, interact with, and experience. From the start of the World Wide Web, there were websites and email, while multimedia and interactive web components came later. This progression of internet and content flexibility has been remarkable, but from the very beginning, above all genres, formats, and subjects of web-based material, there have fundamentally been two types of content - those that the web-user intentionally accessed, and those that the user didn't.

This later category comes in many forms; its content can be harmless or malicious, and it can enhance users' browsing experience or it can be detrimental - but this content is acquired by the user on a daily basis, whether they realize it or not. Although the idea of unintentionally acquired content invisibly following users throughout the entirety of the internet sounds unappealing to say the least, the existence of this material is actually something you've known about all along - spam. Finn Brunton defines spam (or junk) in a nutshell as "the project of leveraging information technology to exploit existing gatherings of attention"[1], while this paper will expand that definition to be include the mediums that aren't necessarily exploitative, but are still reliant upon a user's attention. Regardless of your chosen definition, it is clear that spam is ubiquitous throughout the development of the internet. As Brunton notes in his book *Spam: a Shadow History of the Internet*;

> "[Spam] has changed our language, economics, and culture and exerted a profound effect on our technologies. It has subtly - and not so subtly - deformed the shape of life online, pulling it into new arrangements that make no more sense than the movement of the galaxies unless you allow for the mass of all the dark matter"[2].

In addition to fundamentally shaping the development of the internet, junk develops as technological advances are made - it started as email, but now is ubiquitous as pop-ups, cookies, and even forced downloads of potentially malicious files. Because of its prevalence, spam also has the unique ability

[1] Brunton, Finn, and Finn Brunton. *Spam: A Shadow History of the Internet*. Cambridge, Massachusetts: The MIT Press, n.d.
[2] Ibid.

to shed light onto the cultural experience of a moment in time, similar to how home movies and advertisements are a window into the 20th Century.

The clear cultural and historic value of this shadowy digital trash means that preservation is an imperative. Practically however, the diversity in spam-types combined with the potentially malicious nature of the materials themselves pose certain challenges for both acquiring and safely preserving the materials at hand. This paper will discuss the different types of junk briefly outlined above and will outline potential acquisition methods and challenges based around currently available software.

What is "Junk", really?

Although spam can often be concentrated on websites that are in questionable legal standing (such as downloading or viewing sources of copyright protected feature length films), they are also a common tool for marketers and advertisers who want consumers to be forced to engage with their campaigns. Junk comes in many forms, but the scope of this research is limited to the four broad categories listed below: spam email, cookies, pop-ups, and forced downloads.

Spam Email

The origin of email spam traces back to early 1992, when the commercialization of the internet was complete, and this type of interaction was made possible[3]. Throughout the progression of the internet, the average user has received multiple types of junk email.

The first is chain mail; these emails were often sent around by friends and family, and requested that the same email be forwarded. Although the message itself varied, (some were political in nature, while others suggested consequences if the email was not re-sent, or rewards if it was), the inherent request was to pass the message along, in an effort to create a "viral" experience, as we describe viral videos today. One example of this type of letter can be seen below:

[3] "Retiring the NSFNET Backbone Service: Chronicling the End of an Era," October 3, 2015. https://web.archive.org/web/20151003183700/http://www.merit.edu/research/nsfnet_article.php.

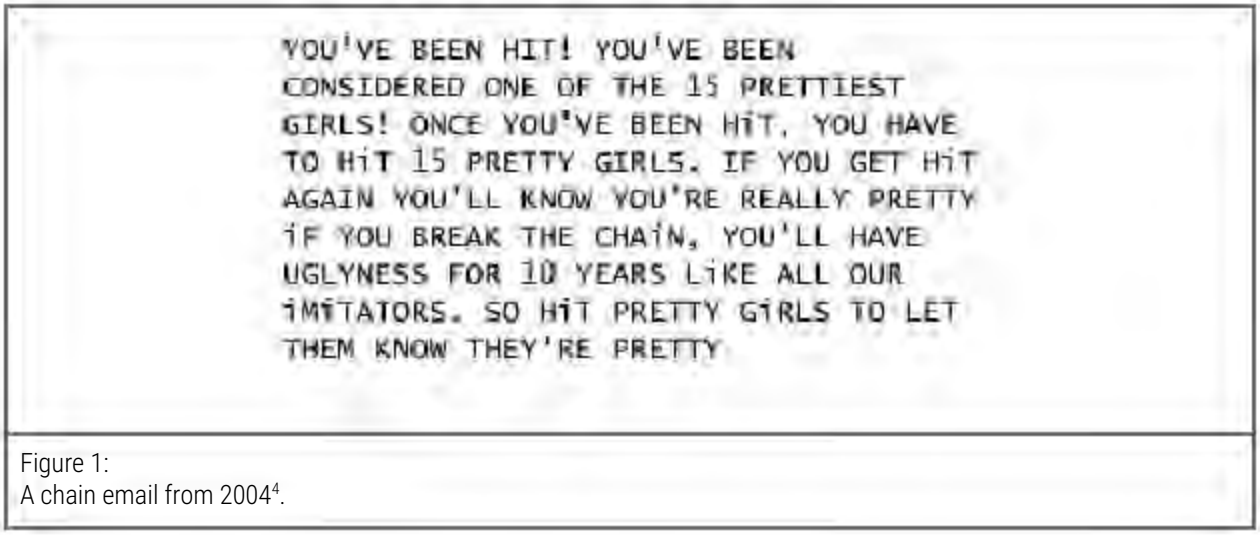


Figure 1:
A chain email from 2004[4].

The second type of spam email is phishing. These emails appear to come from a credible source, such as a bank or a family member, and these are inherently malicious. They usually have the goal of tricking the recipient of the email into either providing private information that can be used for breaking into their accounts, while other older types of phishing emails will often request money be transferred. One such email scheme has gained notoriety as the epitome of basic phishing, and is known as the "Nigerian Prince" scam or "advance fee fraud"[5]:

[4] "Chain Letters Graphics and Gif Animation for Facebook." Accessed December 11, 2017. https://www.funscrape.com/Comments/Chain_Letters.html.
[5] "The Story behind the Nigerian Phishing Scam," PCWorld, March 28, 2010, https://www.pcworld.com/article/192664/the_story_behind_the_nigerian_phishing_scam.html.

SUBJECT: IMPORTANT AND CONFIDENTIAL

FROM: MIKE BONGANTI [MIKE1234@FREEEMAILADDRESS.COM]

I AM BARRISTER MIKE BONGANTI. I AM THE PERSONAL ATTORNEY TO MR. TONY READER OF "WHATCHA GONNA DO WITH ALL THAT JUNK?", A NATIONAL OF YOUR COUNTRY, WHO USED TO WORK WITH SHELL OIL PRODUCING COMPANY IN NIGERIA. HERE IN AFTER SHALL BE REFERRED TO AS MY CLIENT. ON THE 25TH OF DECEMBER, 2003, MY CLIENT AND HIS WIFE AND THEIR ONLY DAUGHTER WERE INVOLVED IN A PLANE CRASH ON THEIR WAY HOME FROM THE HOLIDAYS AND WERE AMONG THE 111 PEOPLE CONFIRMED DEAD.

SINCE THEN I HAVE MADE SEVERAL INQUIRIES TO YOUR EMBASSY HERE TO LOCATE ANY OF MY CLIENT'S EXTENDED RELATIVES, WHICH HAS ALSO PROVED ABORTIVE. AFTER THESE SEVERAL UNSUCCESSFUL ATTEMPTS, I DECIDED CONTACT YOU SINCE YOU HAVE THE SAME LAST NAME, I HAVE CONTACTED YOU TO ASSIST IN REPATRIATING THE FUND VALUED AT $122,000,000 USD(ONE HUNDERED AND TWENTY TWO MILLION UNITED STATES DOLLARS) LEFT BEHIND BY MY CLIENT BEFORE IT GET CONFISCATED OR DECLARED UNSERVICEABLE BY THE INTERCONTINENTAL BANK PLC , WHERE THIS HUGE AMOUNT WAS DEPOSITED.

WHAT BOTHERS ME MOST IS THE FUND, WHICH ACCORDING TO MY COUNTRY'S CONSTITUTION BECOMES THE GOVERNMENTS OWN, IF THERE IS NO ONE CLOSE TO THE DECEASED TO CLAIM THE FUNDS. THE INTERCONTINENTAL BANK PLC HAS ISSUED ME A NOTICE TO PROVIDE THE NEXT OF KIN OR HAVE THE ACCOUNT CONFISCATED IN THE NEXT ONE MONTH. FOR THE FACT THAT I HAVE BEEN UNSUCCESSFUL IN LOCATING ANY RELATIVES OF MY LATE CLIENT SINCE THEN, I HEREBY SEEK YOUR CONSENT TO PRESENT YOU AS THE NEXT OF KIN TO THE DECEASED SO THAT THE PROCEEDS OF THIS ACCOUNT CAN BE PAID TO YOU.

THEREFORE, ON RECEIPT OF YOUR POSITIVE RESPONSE, WE SHALL THEN DISCUSS THE PERCENTAGE THAT IS COMMING TO ME AFTER THE MONEY IS BEEN CREDITTED INTO YOUR CHOSEN ACCOUNT AND THE MODALITIES FOR SHARING IS AS FOLLOWS: 30% SHALL BE FOR ME AND 30% FOR YOU AND 30% WILL BE ALLOCATED FOR THE EXPENCES BOTH PARTIES INCURRED TO MAKING THIS TRANSACTION A SUCCESS.

I CAN ASSIST IN PROCURING THE NECESSARY INFORMATION AND LEGAL DOCUMENTS THAT MAY BE NEEDED TO ACTUALISE THIS PROPOSAL. ALL I REQUIRE FROM YOU IS YOUR HONEST CO-OPERATION TO ENABLE US SEE THIS TRANSACTION THROUGH. I GUARANTEE THAT THIS WILL BE EXECUTED UNDER A LEGITIMATE PLATFORM THAT WILL PROTECT YOU FROM ANY BREACH OF THE LAW.

UPON YOUR ACCEPTANCE TO ASSIST PLEASE SEND ME THE FOLLOWING FOR THE SERIOUS DISCUSSION:

1) YOUR FULL NAMES
2) AGE
3) CONTACT ADDRESS
4) OCCUPATION
5) PRIVATE PHONE NUMBER

I LOOK FORWARD TO HEAR FROM YOU SOONEST. PLEASE NOTE THAT THIS TRANSACTION MUST REMAIN CONFIDENTIAL AT ALL COST.

REGARDS,

BARRISTER MIKE BONGANTI

Figure 2:
A typical example of the Nigerian Prince scam, as generated for this paper on scamgen.com[6].

The last type of email is currently the most common, thanks to the CAN-SPAM Act of 2003[7], which was created to control invasive junk email practices by advertisers and marketing agencies. These are semi-consensual marketing emails, that the intended recipient subscribed to at some point during their lifetime. These emails come with the option to "unsubscribe", but can also be quite invasive and persistent while still complying with CAN-SPAM regulations - in 2016, Groupon, the

[6] "Nigerian Scam Letter Generator." Accessed December 11, 2017. https://www.scamgen.com/.
[7] "CAN-SPAM Act: A Compliance Guide for Business." Federal Trade Commission, September 2, 2009. https://www.ftc.gov/tips-advice/business-center/guidance/can-spam-act-compliance-guide-business.

deals and coupons website, sent each subscriber an average of 388 emails - that's more than one per day[8].

Cookies

Cookies, or HTTP cookies, are "a packet of data sent by an Internet server to a browser, which is returned by the browser each time it subsequently accesses the same server, used to identify the user or track their access to the server"[9], or in plainer terms, a file that tracks a user's path on a website. These are often helpful, as they are what help keep a user logged into their account as they navigate pages within a website, but they can also be used for ecommerce (online shopping) or marketing, and are stored within your browser until deleted.

Pop-ups

Although pop-ups can often be concentrated on websites that are in questionable legal standing (such as sources for downloading or viewing sources of copyright protected feature length films), they are also a common tool for marketers and advertisers who want consumers to be forced to engage with their campaigns. This diversity makes pop-ups preservation wildcards, as they can occur both internally and externally of a browser, can be any size, and can even self-determine whether they will appear in a new window or a new tab. Furthermore, some pop-ups are innocent and informational (i.e. "50% OFF SALE!"), while others are malicious and may even attempt to download files automatically (i.e. "CONGRATULATIONS, YOU HAVE WON $1,000,000!!!"). Two examples of pop-ups can be seen below:

[8] McGoogan, Cara. "These 15 Companies Sent You the Most Spam Emails Last Year." *The Telegraph*, February 19, 2016. http://www.telegraph.co.uk/technology/2016/02/18/these-15-companies-sent-you-the-most-spam-emails-last-year/.
[9] Stevenson, Angus, and Christine A. Lindberg, eds. *New Oxford American Dictionary*. Oxford University Press, 2011. http://www.oxfordreference.com.

Figure 3:
An internal pop-up which prevents the user from interacting with the website allrecipes.com until it is closed.

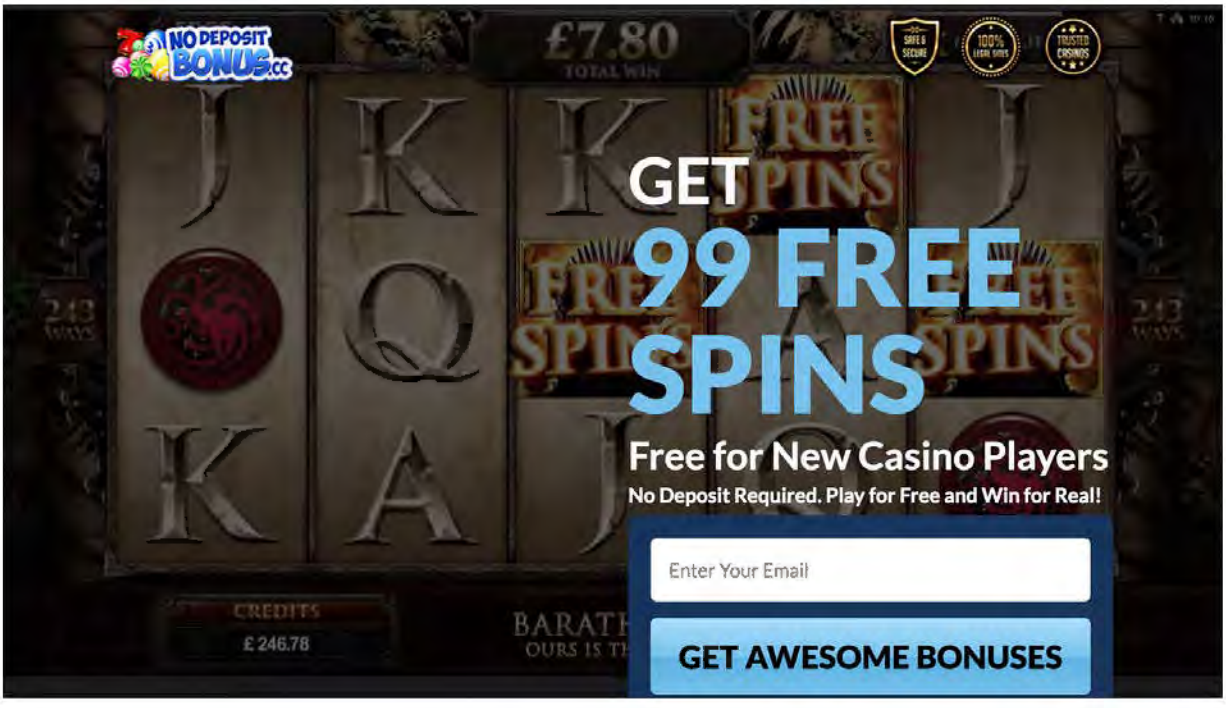


Figure 4:
An external pop-up which loads in a separate window, and has autonomy as a separate website.

Forced Downloads

Forced downloads are when a website will initialize a download without your consent, by using a button that appears to perform an action, such as playing a video, to start downloading a file. On a mac, these files usually have the file extension .dmg, while .exe is more common for Windows operating systems.These downloads could contain any content, but due to the intentionally deceptive method of initializing the download, it's usually assumed that they are malicious.

Junk Acquisition

Because spam is fundamentally an unwanted material, the method of acquisition is intriguing as I have not yet found any software or platform that actually intends to capture junk. Usually junk is the obstacle that gets in the way of preserving content - not the material that you are actually looking to preserve.

Methodology

I selected 2 web-archiving platforms for the website-based spam content. These platforms have different fundamental navigational strategies in order to determine whether acquiring junk is both possible and practical. The first method is GNU wget, a command-line based software that will systematically and automatically "crawl" websites and capture the content from the server that the website is hosted on[10], and the second is webrecorder.io, a browser-based capturing system that will track the user's physical interaction with a website, and record the pages and content that results[11].

Because cookies and spam email are not website-based, I found other ways of preserving the content using non-preservation designed methods.

All acquisition experimentation was conducted on a 2012 MacBook Pro running Sierra (version 10.12.6), and completed either in the command-line or using the Firefox Quantum browser (version 57.0.1).

[10] "GNU Wget." Accessed December 11, 2017. https://www.gnu.org/software/wget/.
[11] "Webrecorder." Accessed December 11, 2017. https://webrecorder.io/.

Experimentation

Cookies

HTTP cookies, as information reflective of an individual's browsing and website interaction, are contained within the browser. The cookies are all compiled in one place, and therefore the simplest method of capturing the data would be to download it as a singular file. I found several extensions for Firefox Quantum that were designed to export cookies, and selected one named "cookies.txt", an extension which "exports all cookies to a Netscape HTTP Cookie File"[12]. The following image is a screenshot example of the resulting file of cookies that correspond to my browsing history:

Figure 5:
A screenshot of the first portion of the HTTP cookies associated with my browsing history, as exported by the "cookies.txt" Firefox extension.

Spam Email

One individual aspect to email is that it requires a user account to access - it's impossible to receive spam without an email address. This means that the method of capturing an email is dependent on

[12] lennonhill. *Cookies-Txt: A Firefox Add-on Which Exports All Cookies to a Netscape HTTP Cookie File*. JavaScript, 2017. https://github.com/lennonhill/cookies-txt.

the email host used. Because of the diversity of junk email and the potential of malicious files to be sent as attachments, I was concerned about finding a method that would keep the functionality of the file intact. One of the most common methods of saving an email for personal use or printing would be to "Save as PDF", which is superficial and therefore inadequate for preserving the functionality and nature of the content itself - a virus or a video can't be represented as a PDF, or represented as content inside a PDF document, either.

Because of this, I turned to .eml files. These maintain the metadata of the email, and can also contain any attachments, regardless of file format. In my experimentation, I attached a video file to an email, and sent it to myself using gmail. The process of saving it as a .eml file[13] contained 5 simple steps for downloading through gmail on a Mac:

1. Open the email you wish to save
2. Click the down arrow that's next to the "Reply" or "Forward" button
3. Choose the option "Show Original" from the menu
4. Click the "Download Original" button
5. Find the downloaded file, and change the extension to .eml, and then approve the extension change

This method produced a file that I was able to open in Thunderbird without an internet connection, and I was able to open and play the attached video file:

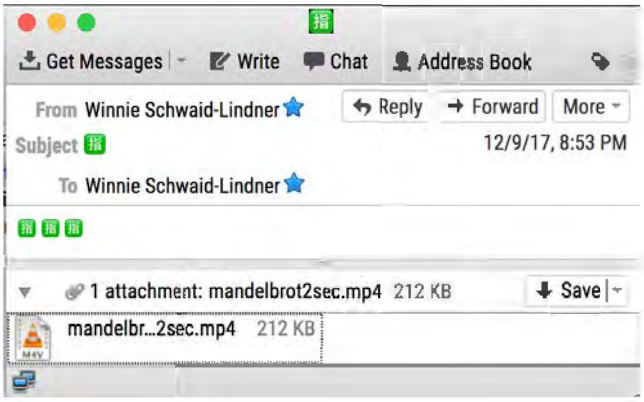


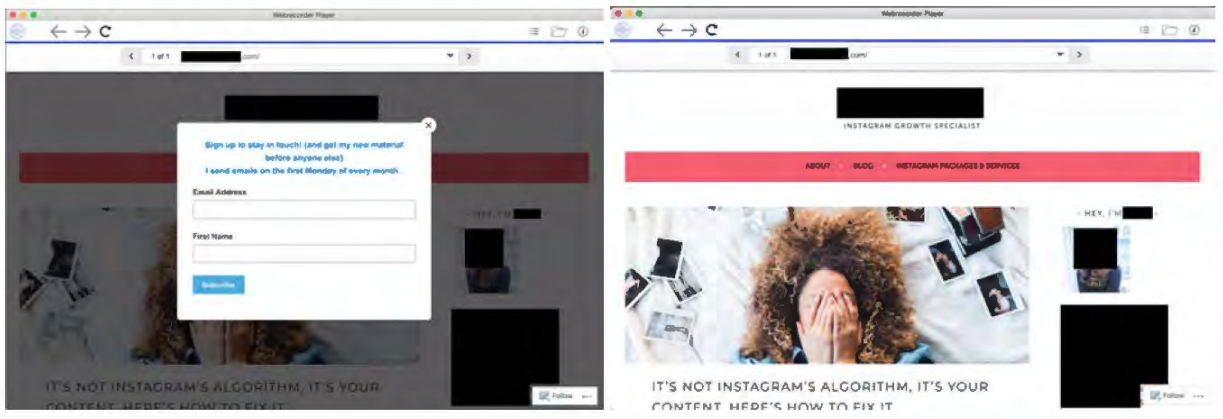
Figure 6:
A screenshot of the opened .eml file, with the playable .mp4 test video attached.

[13] Tschabitscher, Heinz. "Want to Save an Email as an EML File in Gmail? Here's How." Lifewire. Accessed December 10, 2017. https://www.lifewire.com/save-an-email-as-an-eml-file-in-gmail-1171956.

Internal Pop-ups

In order to determine how web-archiving software would handle internal pop-ups, I used both wget and webrecorder.io on the simple blogger and business website that contained a homepage pop-up, requesting that the user subscribe to an email list.

Webrecorder.io successfully captured the pop-up, the content after the pop-up was closed, and the correct order of the two - it was able to recreate the experience of browsing, where upon visiting the page the pop-up would occur, and the user would need to close it in order to continue to navigate throughout the site:



Figures 7 and 8:
Webrecorder.io's capture of a blogger's website with an internal pop-up (left), and the subsequent page when the pop-up is closed (right).

On the other hand, wget had trouble with this type of material - almost immediately an error message was generated as the software attempted to process the pop-up, but it's worth noting that the remainder of the website was captured. The error message can be seen below:

```
2017-12-09 17:19:29 (541 KB/s) - '[redacted].com/index.html' saved [57146]

--2017-12-09 17:19:30--  https://[redacted].com/xmlrpc.php
Reusing existing connection to [redacted].com:443.
HTTP request sent, awaiting response... 405 Method Not Allowed

    0K                                                       2.00M=0s

2017-12-09 17:19:30 ERROR 405: Method Not Allowed.
```

Figure 9:
Wget was unable to capture the internal pop-up of this blogger's website, and produced this error message.

External Pop-ups

When it came to external pop-ups, I selected an online streaming website that was user hosted. This website hosts downloads and streams for many different programs, and an episode of the TV show *Psych* was selected for this experiment as an arbitrary testing example.

This page produced 2 concurrent pop-ups that were opened in separate windows of the browser, and one that was opened as a tab. Webrecorder.io was able to capture all pop-ups, as it opened all derivative pages in additional webrecorder.io windows:

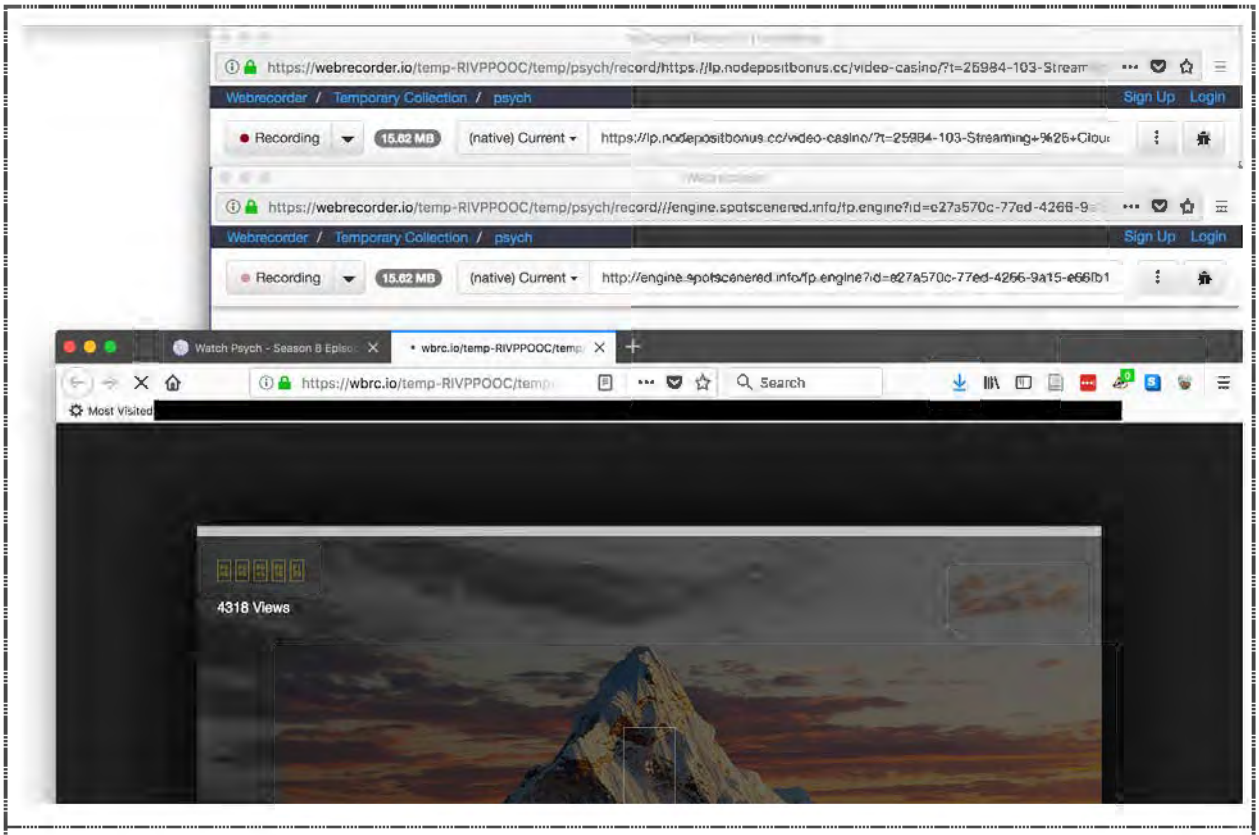


Figure 10:
In the process of navigating the website, 2 windows and one additional tab were opened, and each kept the webrecorder.io domain that allows the platform to capture the material.

However, while webrecorder.io was able to capture the content, it wasn't able to preserve the order - the external pop-ups were saved as concurrent files, but the system wasn't aware of how they should be ordered, unlike the internal pop-ups in the proceeding section.

In comparison, wget was not able to capture any aspect of the website, at all. Immediately an error message was displayed; in addition to not capturing the pop-up, this error message halted the

capture, so that the crawler was unable to reach any part of the website, and resulted in an empty capture.

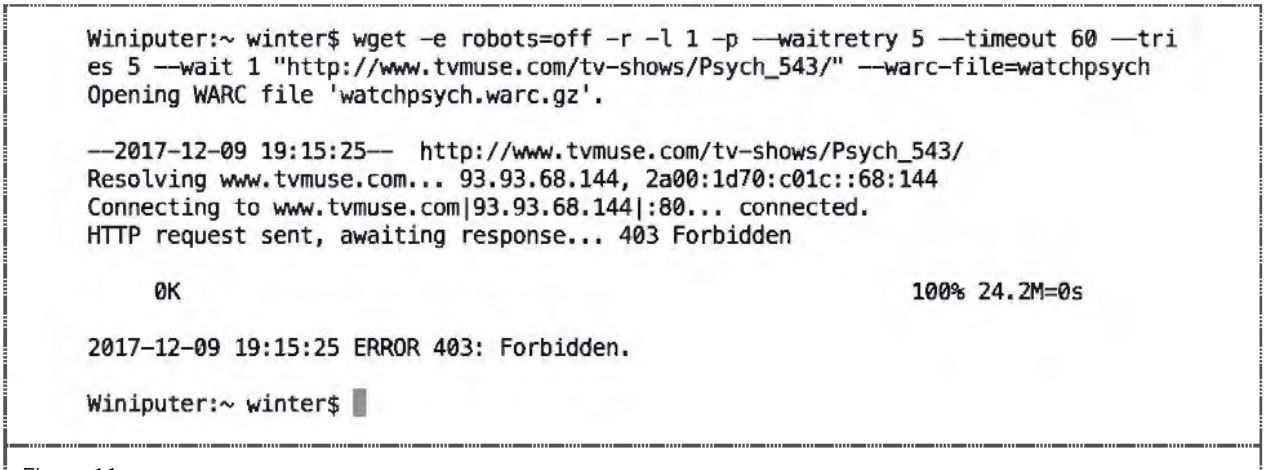


Figure 11:
Upon attempting to use wget to capture external pop-ups, it was unable to access the website, and provided an error message immediately.

Forced Downloads

Because pop-ups are junk that are contained within the realm of your specific browser - whatever animations or actions are required, as annoying as they may be, are fundamentally contained within a webpage, and will be preserved as such. When a file is downloaded through the internet, it exceeds the micro-environment of the web browser that initialized the download, and enters the realm of the computer as a whole.

As the download is started from a browser, I decided to attempt capture through the web-archiving software, in the hope that they would be able to save files as part of the user interaction. For this type of junk, I used a more controlled environment to test the efficiency of the software, as simultaneous pop-ups and downloads that may occur on websites could produce misleading or problematic files, especially considering the results of the external pop-up section above.

Wget was incredibly successful in downloading the files, and was even able to recreate the downloads when “browsing” the captured site, and clicking the download link:

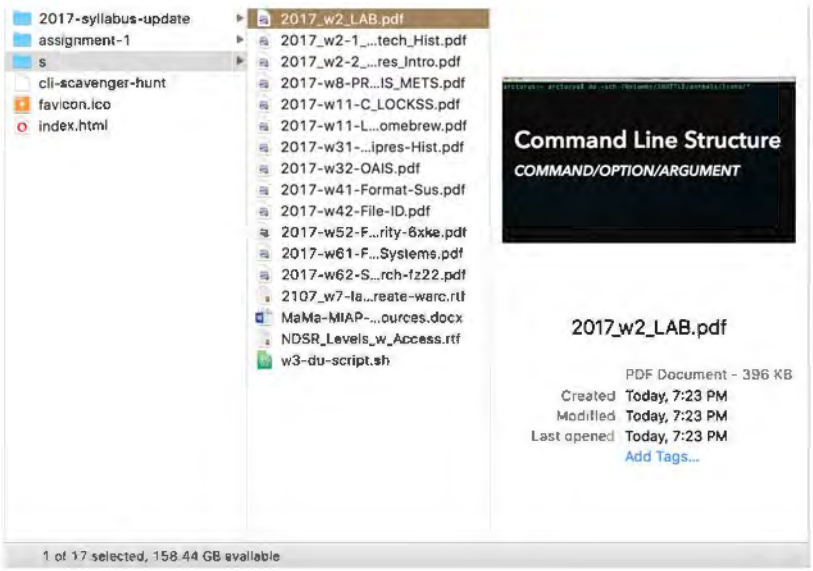


Figure 12:
Wget successfully captured both the files and index from digiprez.com

Although wget worked smoothly, I had mixed results using webrecorder.io for downloads. This system enabled me to download a text file when I was capturing, but didn't save any record of the file download:

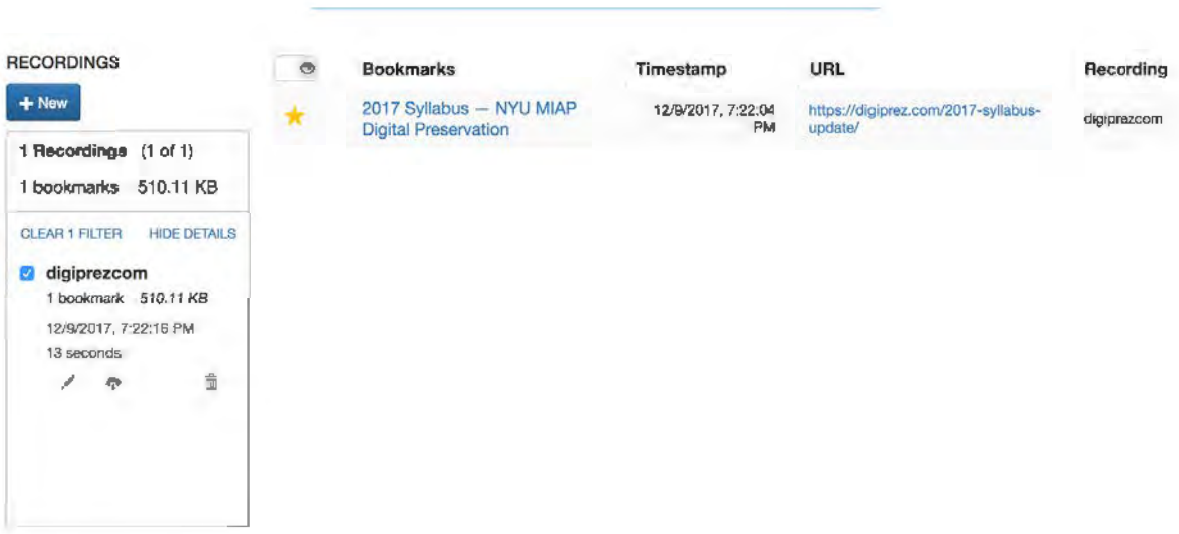


Figure 13:
Webrecorder.io didn't save any indication that a file was downloaded during the recording process.

Although the file had previously been downloaded, clicking on the download link when reviewing the capture resulted in an error - webrecorder.io was unable to recreate the download process at all.

On the other hand, when a .PDF file is clicked on within Firefox (and other browsers such as Chrome, as well), a preview window will display the file within the browser. When that link is clicked while capturing using webrecorder.io, the PDF is similarly displayed as a preview. However, when the link is clicked from the recorded page, webrecorder.io was able to produce a downloaded copy of the file:

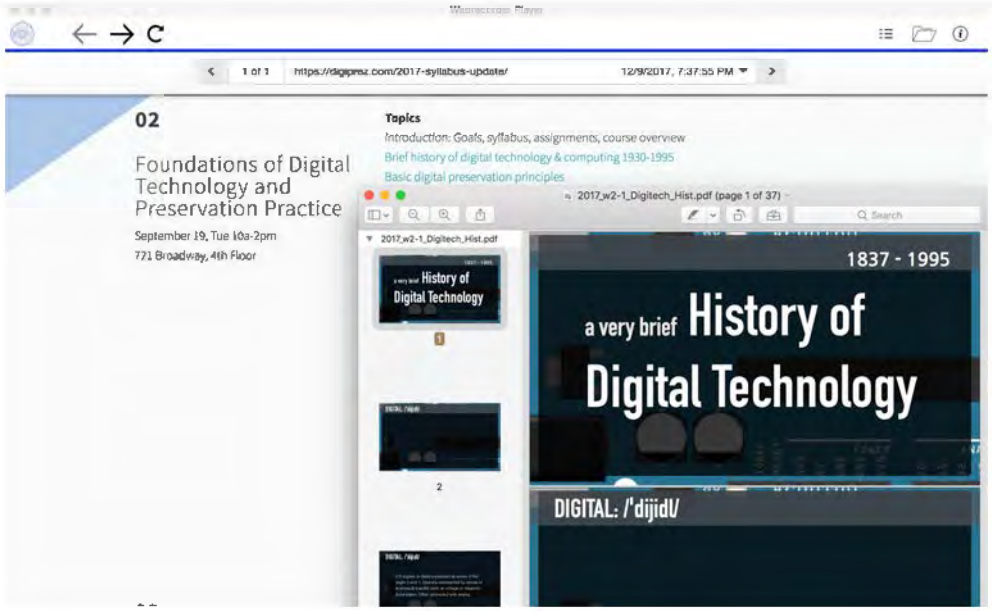


Figure 14:
Webrecorder.io was able to save .PDF files, but not able to save files that don't preview within the browser.

Although certainly successful with .PDFs, the lack of ability to handle any non-browser-previewed files (such as videos or executable files) means that most formats will not be successfully acquired if this method is used.

Junk Retention

Lastly, one natural question of preserving content that is malicious or not intended to be preserved, is that after accessioning these materials must be stored - how can a digital repository safely preserve knowingly malicious content?

The short answer is that malicious materials need to be executed in order to cause damage. Therefore If you are only interested in preservation and not recreating the effects of certain malicious content, there is not necessarily an inherent risk to the network, collection, or repository.

However, there are ways to be safer and more risk-conscious when preserving these materials then just vowing to never open them. As Jonathan Farbowitz notes in *More than Digital Dirt: Preserving Malware in Archives, Museums, and Libraries*: "while the disk image remains unopened the malware is basically inert. Writing the disk image to an LTO tape (as opposed to the hard drive of a computer) provides an additional level of segregation between the malware and a live computer system"[14].

Conclusion

Although untraditional, these pieces of digital junk are vital to the development of digital culture, and could be potentially valuable for future historians and archivists - however, they are also a technological content management challenge. When it comes to preserving these materials, one must consider the diversity in both the content type and goal of spam - from malicious forced downloads to helpful HTTP cookies that stop you from constantly having to log back into your email.

When considering how to best acquire these materials, there is no one-size-fits-all solution. Both wget and webrecorder.io had different features to bring to the table, which means that at this point spam preservation is still not practical for widespread or automated accessioning, but is remarkably possible for targeted sites and content.

Although it is impossible to preserve even close to the amount of junk that exists today, it is imperative to recognize that this type of material is one of the cultural markers of our current technological renaissance, and that it is worthy of historic and cultural attention for years to come.

[14] Farbowitz, Jonathan. "More than Digital Dirt: Preserving Malware in Archives, Museums, and Libraries." New York University, 2016.
https://www.nyu.edu/tisch/preservation/program/student_work/2016spring/16s_thesis_farbowitz_final.pdf.