Frannie Trempe
CINE-GT 1807
15 December 2017

Archiving At-Risk User Generated Content: An Examination of Internet Service Shutdowns

Internet services come and go with alarming frequency. Silicon Valley funding sources dry up, popularity of services waxes and wanes, even legal trouble can arise, forcing a service shutdown. In some cases, the public is given notice to retrieve their data, sparking the frantic scraping and archiving from groups of dedicated web archivists and digital preservation advocates. However, these shutdowns just as frequently occur without advanced warning. As the following three case studies will show, despite the active community working to archive user-generated internet content, a fundamental shift is needed in how app and service developers understand digital preservation and their responsibilities toward users.

This examination of internet service shutdowns will explore three case studies with various outcomes—a successfully archived service, a failed archiving effort resulting in unknown amounts of lost information, and one in progress for a service that will shut down on December 15, 2017. A comparison of these examples—Google Reader, MegaUpload, and the CompuServe Forums, respectively—will illuminate shared themes across service shutdowns and hopefully lead to a better sense of how best to preserve this important user-generated content going forward, as similar services will continue to disappear as long as the internet is active.

Internet services that hold or host user-generated data can shut down or disappear for any number of reasons. Perhaps most obviously, businesses need to continuously turn a profit in the cutthroat and fast-paced contemporary tech landscape. Tech startups and websites are not beholden to their users; the power dynamic here is tilted entirely toward the businesses that hold

information created by their users. To use an obvious, if not dramatic example: if Facebook were to shut down next week and destroy the posts, pictures, and exchanges created by users, untold amounts of cultural and personal history would be lost completely, with no possible recourse. Even though Facebook enables users to download their own data, this is not well-publicized and the implications of digital preservation are not widely known within average social media circles. Sometimes web-based companies shut down and clear out their servers for financial reasons. Often a larger parent company will stay in business but fold one arm, as Google has done so often. Before GeoCities officially shut down in 2009, it had been all but abandoned by Yahoo for years, with little support available; this slow demise is ultimately capped off by a swift shutdown and deletion of user-created sites.[1]

Online services also lose data after redesign or rebranding efforts—such as MySpace in 2013. After years of wild success as a customizable social tool and platform for bands and musicians to engage with their fans in a new way, the company faced difficulties. Facebook, with its streamlined, minimalist aesthetic, had been the social media platform of choice for several years, and though MySpace retained its core users—bands and their followers—their influence had dropped off significantly. Equipped with Justin Timberlake as the face of a new rebranding effort, the website launched a new look, and in the process eliminated significant amounts of user accumulated data without warning, including private messages, blog posts, customized HTML background design, videos, and all friend lists. Since this elimination of data happened without any prior indication, very little of the website's heyday has been archived. In addition to the lack of lead time, much of MySpace (and similar social media platforms) was kept behind a

[1] "GeoCities" *Archive Team*. Last updated 4 December 2017. http://archiveteam.org/index.php?title=GeoCities

username/password wall or not made public due to a user's privacy settings, which would have further complicated archiving efforts—both logistically and ethically.[2]

With so much online content being generated on a daily basis by everyday users, several individuals, institutions, and informal groups have taken up the charge of preserving this important cultural history. Most obviously, the Internet Archive's Wayback Machine serves as a foremost pillar of web archiving. Established in 2001, the Wayback Machine maintains versions of webpages, obtained through crawling software created by the Internet Archive. The Wayback Machine holds numerous petabytes of data and only continues to grow, dominating the landscape of web archiving. Due to their reliable infrastructure and longevity, several if not most other web archiving efforts ultimately push their data to the Internet Archive.

Several communities have arisen in recent years among like-minded internet users regarding data storage and do-it-yourself digital preservation. Twitter serves as one hub for this activity, with active hashtags and pop-up efforts surrounding immediate archiving needs. When local news websites DNAinfo and Gothamist shut down without warning in November 2017 shortly after staff voted to unionize, journalists, programmers, and digital preservation advocates worked to piece together the websites' archives.[3] In the case of DNAinfo and Gothamist, the potential loss of digital information was not just tragic for those seeking to preserve internet history; it also represented the portfolios of countless journalists who may not have backed up their own articles.[4]

---

[2] "MySpace" *Archive Team*. Last updated 17 January 2017
http://archiveteam.org/index.php?title=Myspace

[3] User turtlekiosk. Twitter Post. 22 November 2017.
https://twitter.com/turtlekiosk/status/926289601827934208

[4] Laura Hazard Owen. "Here are three tools that help digital journalists save their work in case a site shuts down." *Nieman Lab*. 21 November 2017.
https://web.archive.org/web/20171213094941/http://www.niemanlab.org/2017/11/here-are-three-tools-

In addition to Twitter efforts, communities focused on digital storage and data recovery continue to thrive on Reddit, Discord, and likely in pockets of the internet not easily found by a simple Google search. One significant example is the Reddit forum r/Datahoarder, whose members advertise themselves as "digital librarians."[5] Much of the subreddit is dedicated to discussions of hard drive capacities and reliability, but there is also a clear focus on preserving at-risk material, much of it user generated; top posts as of December 2017 share news of video website vid.me shutting down, the recovery of data from a defunct amateur erotic literature website, and a discussion of music website SoundCloud's financial troubles and possible shutdown in mid-2017. These communities are deeply interconnected; a cursory look at r/Datahoarder shows overlap between these storage-minded Reddit users and the loosely-organized group Archive Team.

The informal group of web archivists and digital preservation advocates known as Archive Team is not affiliated with the Internet Archive, but they work closely in tandem. Composed entirely of volunteers, Archive Team works primarily toward the preservation of this user-generated online content, especially that which has been identified as at-risk or in the process of dying. Through the group's comprehensive wiki, Archive Team specifically notes their distinction from the Internet Archive (in an ironic turn, the Archive Team website only occasionally loads properly; users are frequently faced with a "resource limit is reached" warning when navigating through pages). This wiki-based website acts as a record of all past and ongoing Archive Team projects, including a "Death Watch" list of particularly at-risk websites;

---

that-help-digital-journalists-save-their-work-in-case-a-site-shuts-down/

[5] "Who are we?" *Reddit.* Accessed 14 December 2017. https://www.reddit.com/r/DataHoarder/

the group has worked, successfully or otherwise, on archiving efforts of hundreds of websites and internet platforms. The public face of Archive Team is Jason Scott, who works at the Internet Archive.[6]

Archive Team uses two major tools in projects to capture and preserve online user content: ArchiveBot and Archive Team Warrior. Both automated tools operate through IRC channels to receive jobs. IRC, Internet Relay Chat, is a longtime platform for online communication; users communicate through an IRC client directly to a server, which in turn directs messages to the proper client and user receiving messages. IRC has been in operation since the late 1980s; it is fitting that Archive Team use such a time-tested technology in its communication. Most Archive Team jobs are completed using the ArchiveBot tool, an automation for the archiving of websites that contain "up to a few hundred thousand URLs." The bot receives commands through a dedicated IRC channel instructing which websites to crawl and where to begin. Archive Team Warrior is a virtual machine that can be installed and run in the background of any capable computer, and similarly receives tasks through IRC.[7]

As with nearly all tools in contemporary web archiving, Archivebot and Archive Team Warrior create WARC files of crawled sites, which reside on the Archive Team servers before eventually being pushed to the Internet Archive. The tools are open source, with source code available to view on GitHub. Similarly, all of Archive Team's communications through IRC are publicly accessible online, facilitating an atmosphere of openness and transparency in the face of

---

[6] "Frequently Asked Questions." *Archive Team*. Last updated 4 December 2017. http://archiveteam.org/index.php?title=Frequently_Asked_Questions

[7] "ArchiveBot." *Archive Team*. Last updated 6 August 2017. http://archiveteam.org/index.php?title=ArchiveBot

the relatively secrecy behind the closing of the websites themselves.[8] This informal and often fragmented nature of these various internet data recovery groups (with the exception of the Internet Archive, which has a prominent physical location and has been around for over twenty years) is unsurprising; many of these websites that shut down are done for reasons of questionable legality or for other nefarious purposes; formal institutions such as university libraries or government bodies like the Library of Congress would be easier to hold accountable if businesses did not want their websites archived than a group of four or five "rogue archivists" working behind screen names through IRC channels or a Reddit forum. This approach is demonstrated further in the treatment of robots.txt exclusions by Archive Team, whose wiki includes a page with the subtitle, "ROBOTS.TXT IS A SUICIDE NOTE;" they find workarounds whenever the exclusions are encountered and do not obey their presence within a website except under extreme circumstances, striving instead to archive everything.[9]

The first of three case studies demonstrating the efforts advocates for internet history take when online services shut down is that of Google Reader. A successful RSS feed launched in 2005, Google announced in 2013 that the service would be eliminated, providing users with just over three months' notice to pull their data and/or acclimate to a new RSS feed service before the disappearance of Reader. This announcement was met with disappointment and scorn from digital archivists and avid Reader users; even though RSS feeds had faced a decline in popularity due in part to the rise of Twitter as microblogging, Reader's ability to track historical feed data, such as edits to posts, made it popular with many dedicate users. Former Google Reader

---

[8] "#archiveteam." *EfNet IRC client*. efnet.org/?module=channels&s=0&ch=archive-team&u=5

[9] "Robots.txt." *Archive Team*. Last updated 17 January 2017. http://www.archiveteam.org/index.php?title=Robots.txt

employees, web archivists, and interested users sprang to action at the news, working to compile as many unique feed URLs as possible through a combination of both crawling and crowdsourcing.[10] A collection of tools to successfully archive Reader content titled "Reader is Dead" appeared, created by former Google employee Mihai Parparita, who himself had worked on the RSS feed platform.[11] This was an incredible boon to the recovery efforts; most shuttering websites do not have the luxury of a preservation-minded former employee who knows the API's ins and outs. Ultimately, 8800GB were uploaded to the Internet Archive, representing a massive success for interested web archivists.[12]

While the efforts to retrieve Google Reader data and unique feeds were largely successful, the shutdown of video and file sharing website MegaUpload is considered a failure by Archive Team. Credited as having taken up four percent of all internet traffic during its peak, MegaUpload servers were taken offline in 2013 without notice due to an ongoing legal battle stemming from copyright issues. The website did not contain exclusively copyrighted material; users could share or upload anything. Even with any notice that the website would be shut down, the massive amount of storage space and technical difficulty required to archive videos, in addition to the human culling that would be necessary to sort out major copyrighted material such as Hollywood films, television shows, or music videos. While a few screenshots and file lists exist of MegaUpload, the site is largely lost to history.[13]

---

[10] "Google Reader/War Room" *Archive Team*. Last updated 25 June 2015. http://www.archiveteam.org/index.php?title=Google_Reader/War_room

[11] Mihai Parparita. *Reader is Dead*. Accessed 14 December 2017. http://readerisdead.com/

[12] "Google Reader/War Room."

[13] "MegaUpload." *Archive Team*. Last updated 16 January 2017. http://www.archiveteam.org/index.php?title=MegaUpload

AOL announced in November 2017 that they would be shutting down the CompuServe forums, a stalwart and early leader in Internet communication. The sheer volume of pages to archive—CompuServe's heyday was during the 1980s into the mid-1990s—presented a significant challenge to Archive Team in working to save this piece of history.[14] Their IRC chat logs, particularly the #compuswerve channel, were a valuable resource in the days leading up to December 15, 2017—the site's official shutdown date.[15] This project includes a large ArchiveBot job, attempting to grab over six million unique URLs, as well as a wpull—the Python alternative to wget. Due to the size, the ArchiveBot job unexpectedly crashed, and just days before the closing date, it looks as though much of the CompuServe forums will not be recovered.[16] The volume and short recovery window, coupled with the fact that Archive Team is just a few enthusiastic volunteers at any given time, albeit with day jobs and other projects to balance made this project a significant challenge.

As of 6:13am December 15, 2017, the CompuServe forums are still visible; one post on a classical music board serves as a reminder that the demise of these websites is not just a loss for cultural history from a broad, scholarly perspective, but can also play a large part in the lives of individual users. In a poignant farewell post, user John Francis writes, "Today's the day, and if CompuServe's announcement holds good, by midnight this forum and all the other forums will be only a memory. For me a long memory, 32 years, rich in friendships and shared experiences. To see it all wiped out is too, too sad."[17]

---

[14] "CompuServe Forums" *Archive Team*. Last updated 18 November 2017. http://archiveteam.org/index.php?title=CompuServe_Forums

[15] "#compuswerve." IRC chat logs. 09 December 2017 – 15 December 2017. https://origin.badcheese.com/~steve/atlogs/?chan=compuswerve&day=2017-12-14

[16] Ibid.

[17] John Francis, "CompuServe's forums to be shut down." *CompuServe*. 15 December 2017. http://forums.compuserve.com/discussions/Entertainment_Forum/Classical_Music/CompuSer

These three case studies are just a small fraction when considering the history of disappeared internet services, a history which continues to this day. In the course of composing this paper, news broke that social media compiling service Storify will shut down and delete all posts in May 2018.[18] Despite the grim outlook, there are valiant efforts being made to save and recover all varieties of online user data, such as those from Archive Team and the Internet Archive. As case studies such as CompuServe show, however, Archive Team needs additional help from the digital preservation community to preserve these slices of cultural history. Their work would benefit from additional computing power and help with organizational efforts surrounding various projects.

On a larger scale, the web archiving community must increase visibility within Silicon Valley tech culture itself—a monumental task, but finding some way to incentivize preservation or provide resources to shuttering businesses to preserve their data seems like a fundamental step in increasing long term preservation of user-generated material. At present, the world of web archiving seems fairly disjointed and separated from the profit-focused, always forward-thinking mindset of Silicon Valley funders and developers, but perhaps this could change with some effort. Ultimately, these businesses have no obligation to the users who may craft large parts of their lives through online platforms.

As the above discussion and case studies show, the disappearance of user-generated content is a major concern in the current landscape of web archiving, but still may not receive the attention it deserves within the larger digital preservation community. The complex nature of

---

ves_forums_to_be_shut_down/ws-tv/1195.4?nav=messages

[18] Shannon Liao. "Storify is shutting down and will delete all posts next May." *The Verge*. 12 December 2017. https://www.theverge.com/2017/12/12/16767880/storify-shut-down-2-livefyre

these projects; the massive volume of URLs, difficulty of grabbing dynamic content, or copyright and privacy issues can all serve as major hurdles in saving the internet as we know it. This study limits itself to traditional websites largely because these few efforts such as Archive Team have focused on those as well. Now with more and more internet traffic working solely through mobile apps, this preservation process will only complicate further. The landscape is not completely dire—this is exciting, new territory, and the collaborative efforts to save data are incredibly reassuring. The cultural value of this user-generated content cannot be overstated, and efforts to protect that data are crucial to the preservation of society as we know it.

Bibliography

"#archiveteam." *EfNet* IRC client. efnet.org/?module=channels&s=0&ch=archive-team&u=5

"#compuswerve." *IRC chat logs*. 09 December 2017 – 15 December 2017. https://origin.badcheese.com/~steve/atlogs/?chan=compuswerve&day=2017-12-14

"ArchiveBot." *Archive Team*. Last updated 6 August 2017. http://archiveteam.org/index.php?title=ArchiveBot

"CompuServe Forums" *Archive Team*. Last updated 18 November 2017. http://archiveteam.org/index.php?title=CompuServe_Forums

Francis, John. "CompuServe's forums to be shut down." *CompuServe*. 15 December 2017. http://forums.compuserve.com/discussions/Entertainment_Forum/Classical_Music/CompuServes_forums_to_be_shut_down/ws-tv/1195.4?nav=messages

"Frequently Asked Questions." *Archive Team*. Last updated 4 December 2017. http://archiveteam.org/index.php?title=Frequently_Asked_Questions

"GeoCities" *Archive Team*. Last updated 4 December 2017. http://archiveteam.org/index.php?title=GeoCities

"Google Reader/War Room" *Archive Team*. Last updated 25 June 2015. http://www.archiveteam.org/index.php?title=Google_Reader/War_room

Liao, Shannon. "Storify is shutting down and will delete all posts next May." *The Verge*. 12 December 2017. https://www.theverge.com/2017/12/12/16767880/storify-shut-down-2-livefyre

"MegaUpload." *Archive Team*. Last updated 16 January 2017. http://www.archiveteam.org/index.php?title=MegaUpload

"MySpace" Archive Team. Last updated 17 January 2017 http://archiveteam.org/index.php?title=Myspace

Owen, Laura Hazard. "Here are three tools that help digital journalists save their work in case a site shuts down." *Nieman Lab*. 21 November 2017. https://web.archive.org/web/20171213094941/http://www.niemanlab.org/2017/11/here-are-three-tools-that-help-digital-journalists-save-their-work-in-case-a-site-shuts-down/

Parparita, Mihai. *Reader is Dead*. Accessed 14 December 2017. http://readerisdead.com/

"Robots.txt." *Archive Team*. Last updated 17 January 2017. http://www.archiveteam.org/index.php?title=Robots.txt

turtlekiosk. *Twitter* Post. 22 November 2017. https://twitter.com/turtlekiosk/status/926289601827934208

"Who are we?" *Reddit*. Accessed 14 December 2017. https://www.reddit.com/r/DataHoarder/