Athena Christa Holbrook
CINE-GT 1807
Prof. Kara Van Malssen
1 October 2013

# ArchiveTeam

Founded in 2009 by Jason Scott, Archive Team is an open collective of volunteers striving to archive and provide access to endangered and extinct web-based digital heritage.[1] This communal outfit is comprised of archivists, programmers, writers, and concerned web citizens tackling a broad range of tasks to ensure proper preservation of web-data before it is irretrievably deleted.[2] The levels of involvement can vary depending on the willingness of the participant. Higher level tasks include manually capturing website data, creating instructional essays and guides on general tenets of personal and professional archiving, hosting mirrored sites, dead sites, and torrents, and developing scripts to improve the efficiency and efficacy of automated web-crawling. Those without extensive archiving or programming knowledge can also get involved as "Warriors" by running Archive Team's Warrior machine from their home computers. Furthermore, archive participants with hosted disk space and torrenting capabilities can volunteer to host the data as backup.[3]

ArchiveTeams collects a diverse range of sites, but primarily focuses on sites with user-driven and user-submitted content.[4] Many of these sites are run or hosted by large corporate entities, such as Yahoo, Google, Myspace, and AOL, but are subsidiary to the overarching function of their parent group. Examples of sites such as this that ArchiveTeam has saved include Starwars.yahoo.com [Yahoo], Google Answers [Google], and MobileMe [Apple]. When a parent group decides to shut down a site that is no longer commercially viable, there is often no concern for the user-base, whether active or passive, and user-created content may be deleted without notice. This often occurs when a site loses popularity or is merged with another company. The social networking site Myspace is a classic example of this scenario. In 2013, after being purchased and redesigned by Specific Media, Myspace deleted all blogs, videos, and private messages shared by users over the course of its decade-long existence without notice.[5] Unfortunately, all of this data was irretrievably lost before ArchiveTeam could step in.

---

[1] ArchiveTeam.org, *Main Page*, http://www.archiveteam.org/index.php?title=Main_Page (29 Aug. 2013).
[2] ArchiveTeam.org, *Who We Are,* http://www.archiveteam.org/index.php?title=Who_We_Are (29 Aug. 2013).
[3] http://www.archiveteam.org/index.php?title=Who_We_Are
[4] ArchiveTeam.org, *Philosophy,* http://www.archiveteam.org/index.php?title=Philosophy (29 Aug. 2013).
[5] ArchiveTeam.org, *Myspace,* http://www.archiveteam.org/index.php?title=Myspace (29 Aug. 2013).

Sites created and run by fans or subcultures are also a major concern for ArchiveTeam. Blog hosts like Xanga, fandoms like FanFiction.net, and alternative wikis like Encyclopedia Dramatica are all examples of sites whose lack of stable infrastructure have led to their high susceptibility for shutdown or near demise. Another important example of a threatened alternative site is The Pirate Bay, a torrenting site fraught with potential copyright litigation. Fortunately, ArchiveTeam has been able to save data from all of these sites.

The primary step towards ensuring the longevity of these websites is community vigilance. ArchiveTeam maintains monitoring projects referred to as "DeathWatch" and "FireDrill". "DeathWatch" is a list of active sites that are currently in danger, ranked by the likelihood of shutdown: "Pining for the Fjords" are those that are already in the stages of data-death, "Preemptive Alarm Bells" for sites at imminent risk of termination, and "Other Endangered Species" for those sites just teetering at the edge of eventual doom.[6] ArchiveTeam also keeps a list of dead sites that they were unable to save, as a reminder of the real threat of potential data loss. "FireDrill" is a list of healthy, active sites that, while not currently "at risk", should be watched intently. A site could appear on this list for a number of reasons including the ability to trust its parent site with data based on past behaviours or the amount of users and hosted user-content. Sites are included on the "FireDrill" watch list if "they solicit so much content, contain so many works and projects by a wide group of people, or have the internet particularly dependent on them".[7]

Most recently, ArchiveTeam finished web-crawls of blog site Xanga and Q&A site Formspring, and is currently focused on its URLTeam project. The URLTeam project strives to save linked data from link-shortening services like bit.ly or Tinyurl. Shortened URLs are a nuisance for a number of reasons relating to their use in spam and virus dissemination and traffic hijacking, but they also present crucial preservation issues.[8] Short URLs are particularly susceptible to link rot due to the transient lifecycle of many shortening sites and increased complications from additional steps necessary for the process to function.[9] URLTeam uses an automated script called TinyBack to grab, "unshorten", and save linked data from shortened URLs.[10]

ArchiveTeam uses two additional automated tools for collecting site data: ArchiveBot and ArchiveTeam Warrior. ArchiveBot is an Internet Relay Chat bot used for small sites that crawls a site from a user-submitted URL and records the

---

[6] ArchiveTeam.org, *DeathWatch,* http://www.archiveteam.org/index.php?title=Deathwatch (29 Aug. 2013).
[7] ArchiveTeam.org, *FireDrill,* http://www.archiveteam.org/index.php?title=Fire_Drill (29 Aug. 2013).
[8] Jonas Jacek, *Why URL Shorteners Are Bad*, Rield.com, http://rield.com/faq/why-url-shorteners-are-bad (12 May 2011).
[9] http://rield.com/faq/why-url-shorteners-are-bad
[10] ArchiveTeam.org, *URLTeam,* http://www.archiveteam.org/index.php?title=URLTeam (29 Aug. 2013).

content therein.[11] ArchiveTeam Warrior works similarly, but is an appliance run through a virtual machine and takes on sites with a much more extensive content.[12] Both of these tools run a *wget* script to grab the full content of sites and record them in Web ARChive format (WARC). This format is used because it records request and response headers in addition to redirects and 404 errors[13] and can be automatically ingested into the Internet Archive's WayBack Machine.[14] This information will later be uploaded to the ArchiveTeam server and the Internet Archive for access. This process allows many dead or dying sites to be viewed in their original integrity using the Internet Archive's WayBack Machine emulator.

At this stage, ArchiveTeam has rescued 51 sites, is currently working on saving 30, and has a "to-do list" of another 85.[15] The project is ongoing for the foreseeable future, as sites are being created and destroyed daily. ArchiveTeam's focus on transparency, education, and public-involvement is a smart and innovative approach to what is in essence the unapproachably mammoth objective of archiving the internet. The use of crowd-sourced aid enables their system to efficiently operate and evolve, and fosters a sense of community action and responsibility.

---

[11] ArchiveTeam.org, *ArchiveBot,* http://www.archiveteam.org/index.php?title=ArchiveBot (29 Aug. 2013).
[12] ArchiveTeam.org, *ArchiveTeam Warrior,* http://www.archiveteam.org/index.php?title=ArchiveTeam_Warrior (29 Aug. 2013).
[13] ArchiveTeam.org, *Wget with WARC Output,* http://www.archiveteam.org/index.php?title=Wget_with_WARC_output (29 Aug. 2013).
[14] ArchiveTeam.org, *Frequently Asked Questions,* http://www.archiveteam.org/index.php?title=Frequently_Asked_Questions (29 Aug. 2013).
[15] ArchiveTeam.org, *Projects,* http://www.archiveteam.org/index.php?title=Projects (29 Aug. 2013).

## Webography


"ArchiveTeam.org." *ArchiveTeam*. MediaWiki, 29 Aug. 2013. Web. 01 Oct. 2013.

Bosker, Bianca. "Jason Scott's Archive Team Is Saving The Web From Itself (And Rescuing Your Stuff)." *The Huffington Post*. TheHuffingtonPost.com, 27 Mar. 2013. Web. 01 Oct. 2013.

Brown, Mark. "Archivists Step in as Google Video Shuts down for Good." *Wired UK*. Wired UK, 18 Apr. 2011. Web. 01 Oct. 2013.

Deleon, Nicholas. "Coming Soon: 900GB Torrent Of (Mostly) Every Geocities Web Site Ever." *TechCrunch RSS*. TechCrunch, 29 Oct. 2010. Web. 01 Oct. 2013.

Garfield, Bob. "The Archive Team - On The Media." *Onthemedia*. WNYC, 23 Mar. 2012. Web. 01 Oct. 2013.

Holt, Kris. "Archive Team Races to Preserve Posterous before It Goes Dark." *The Daily Dot*. The Daily Dot, 13 Mar. 2013. Web. 01 Oct. 2013.

Jacek, Jonas. "Why URL Shortening Services and ShortURLs Are Bad." *Why URL Shorteners Are Bad*. Rield.com, 12 May 2011. Web. 01 Oct. 2013.

Masnick, Mike. "Historic Archive Of Websites From The January 18th SOPA Blackout."*Techdirt.* Techdirt, 12 Apr. 2012. Web. 01 Oct. 2013.

Misener, Dan. "Misener.org." *Misenerorg*. Misener.org, 29 Oct. 2010. Web. 01 Oct. 2013.

Modine, Austin. "Web 0.2 Archivists save Geocities from Deletion • The Register." *Web 0.2 Archivists save Geocities from Deletion • The Register*. The Register, 28 Apr. 2009. Web. 01 Oct. 2013.

Morton, Simon. "Radio New Zealand." *: National : This Way Up : 03 Mar 2012 : The Archive Team*. Radio New Zealand, 3 Mar. 2012. Web. 01 Oct. 2013.

Panzarino, Matthew. "The Archive Team Finishes Downloading All 272 Terabytes of MobileMe and .Mac for Posterity." *TNW Network All Stories RSS*. The Next Web, 26 June 2012. Web. 01 Oct. 2013.

Paul-Choudhury, Sumit. "There Has Been an Error - New Scientist." *There Has Been an Error - New Scientist*. New Scientist, 6 May 2011. Web. 01 Oct. 2013.

Perez, Sarah. "Want To Help Archive Upcoming.org Before Yahoo Shuts It Down? Try This." *TechCrunch RSS*. TechCrunch, 22 Apr. 2013. Web. 01 Oct. 2013.

Schwartz, Matt. "Mobile Devices That See in 3-D." *MIT Technology Review*. MIT, 20 Dec. 2011. Web. 01 Oct. 2013.

Sullivan, Mark. "The 'Archive Team' Rescues User Content From Doomed Sites."*PCWorld*. PCWorld, 12 Apr. 2012. Web. 01 Oct. 2013.

Young, Nora. "Full Interview: Jason Scott on Online Video and Digital Heritage | Spark with Nora Young | CBC Radio." *CBCnews*. CBC/Radio Canada, 29 Apr. 2011. Web. 01 Oct. 2013.

Great work! Very well researched and succinctly but thoroughly summarized. Your presentation was also clear and well organized. I'm glad you took on this project for your assignment. I'm esp glad you could get the Warrior working, and could talk about URLTE.AM in detail (also, bit.ly is in Libya?!). I hope you found it interesting and will continue to contribute to ArchiveTeam beyond just this class project, now that you are a top 10 contributor!

Grade: A