CINE-GT 1807
Fall 2019
Madeline Smith
December 13, 2019

Assignment 2: Final Research Paper
The History of Web Archiving and the Evolution of Archive-It's Brozzler

Since the beginning of the Internet and the World Wide Web in 1991, an unimaginable amount of data and content has been created, shared, posted, and stored on the World Wide Web. When the World Wide Web began, there was no system in place to archive and preserve the Web content, as no one could imagine how massive the infrastructure would become. It was not until the 1990s that users began to recognize the gravity of archiving and preserving all the incredibly diverse born-digital data and content on the Web. Web crawlers are a key component in accomplishing the mission to archive Web content before it is lost for good.

While there have been advancements in web archiving over the years, considering the speed at which digital technology and the varied uses of the Internet are changing, there is still work to be done. One of the more recent web crawling technologies is Archive-It's Brozzler. This report aims to frame Brozzler within the larger landscape of web archiving and web crawler technology. The report will attempt to answer whether Brozzler as a web archiving tool is effective for capturing dynamic websites and the rapidly evolving content of the modern web. The end of this report will be an examination of my own experience attempting to use Brozzler.

In order to understand web archiving and web crawling, a few terms used in describing the evolution of the Internet and the World Wide Web need to be defined. The terms "internet" and "world wide web" are often used interchangeably. However, they do not represent the same concept. The "World Wide Web" refers to "the *content* on the web, including web pages,

databases, and other systems that contain data and information."[1] The "Internet" are "the *systems* that connect content on the web. [These systems] include subterranean and sub oceanic fiber optic cables and the hardware and software that support those connections."[2] The Internet allows servers to directly connect to it. A "server" is a special computer connected directly to the Internet. "Webpages" are files on a server's hard drive. Every server has a unique IP (Internet Protocol) address, which helps computers find each other. Everything connected directly or indirectly to the Internet has an IP address. Computers that are used by everyday users are called "clients." This is because they are not directly connected to the Internet. Instead they are connected indirectly through an internet service provider, or ISP.[3] Whenever information, whether it be an email, images, or a Web page, is transferred across the Internet, the computer breaks down the information into smaller pieces called "packets." Once the information reaches its destination, these packets are then reassembled in their original order. This method of data transmission is called "packet switching."[4]

A Brief History of the Internet and the World Wide Web

The concept for what would become the Internet and the World Wide Web began as early as the 1950s, when multiple scientists were working independently on similar theories and concepts around data distribution and computer networking. In the late 1950s, Paul Baran, at the RAND Corporation in California, started exploring the concept of switching small blocks of data when transmitting them over a digital network. This concept would later be termed "packet

---

[1] Nicole Martin. "Digital Preservation: The World Wide Web [Presentation slides]." October 29, 2019. https://static1.squarespace.com/static/57c2e034ebbd1ac6f2565c18/t/5de7fb5088321b3df06c55dd/1575484244770/2019+w8.1+Web+Intro.pdf

[2] Nicole Martin. "Digital Preservation: The World Wide Web [Presentation slides]."

[3] Aaron. "How the Internet Works in 5 Minutes." *YouTube* video, 4:48. No date. https://www.youtube.com/watch?v=7_LPdttKXPc.

[4] Aaron. "How the Internet Works in 5 Minutes."

switching" by Donald Davies in 1965.[5] Packet switching became one of the underlying technologies of the Internet. Davies, while working at the National Physical Laboratory in the United Kingdom, was separately also working on the concept of message routing. Davies only found out about Baran's work in 1966, after Davies presented on the concept to members of the Ministry of Defence [sic].[6] In 1962, Leonard Kleinrock, at MIT, wrote his PhD dissertation, "Information Flow in Large Communication Nets," on queuing theory, which is a key mathematical background to packet switching.[7] All three men's work on packet switching, in addition to a multitude of other research by a small group of scientists and researchers at various institutions around the world, led to, in 1969, the success of the U.S. Defense Department's Advanced Research Projects Agency Network, better known as ARPANET, and the transmission of the first message on the ARPANET. ARPANET was the predecessor to the modern Internet.[8] ARPANET, in service from approximately 1969 to 1990, was designed to "create a network of geographically separated computers that could exchange information via a newly developed protocol … called NCP (Network Control Protocol)."[9] As Charles M. Herzfeld, the former director of ARPANET stated, "the ARPANET came out of our frustration that there were only a limited number of large, powerful research computers across the country, and that many research

---

[5] Janet Abbate. *Inventing the Internet* (Cambridge, Massachusetts: The MIT Press, 1999), 38.
[6] Davies, Donald Watts. Oral history interview by Martin Campbell-Kelly. Mach 17, 1986, National Physical Laboratory, UK. Charles Babbage Institute. Retrieved from the University of Minnesota Digital Conservancy. http://hdl.handle.net/11299/107241.
[7] Leiner, Barry M., Vinton G. Cerf, David D. Clark, Robert E. Kahn, Leonard Kleinrock, Daniel C. Lynch, Jon Postel, Larry G. Roberts, and Stephen Wolff. "Brief History of the Internet." 1997. https://www.internetsociety.org/internet/history-internet/brief-history-internet/.
[8] Kim Ann Zimmermann and Jesse Emspak. "Internet History Timeline: ARPANET to the World Wide Web." June 27, 2017. https://www.livescience.com/20727-internet-history.html.
[9] Mary Bellis. "ARPANET - The First Internet." The Cold War and ARPANET. http://ocean.otr.usm.edu/~w146169/bellis.html.

investigators who should have access to them were geographically separated from them."[10] "ARPA-funded researchers developed many of the Internet protocols used today."[11]

Access and use of ARPANET was limited to a select group of academic and research institutions. By 1971, the number of separate networks located around the world had grown and there was a need to find a way for all of the computers on these networks to communicate with other networks. This was accomplished by the introduction of the Transmission Control Protocol (TCP), invented by computer scientist Vinton Cerf in 1971.[12] TCP was a way to introduce computers across the globe to each other in a virtual space. TCP was quickly followed by the Internet Protocol, or IP.[13] In the 1980s, scientists used these two protocols to send data between their different computers and networks. The 1990s saw these technologies start to ramp up and the networks and methods for "online" communication and dissemination rapidly expand. In 1991, computer programmer Tim Berners-Lee launched the World Wide Web[14] and the development of the Internet as it is known today began. The World Wide Web moved the computer networking technology from a mechanism and space for researchers and scientists to share data to a whole network of information that was accessible and browsable to anyone with an Internet connection. Users accessed the World Wide Web through an Internet browser, just as users still do today. The first browser to have a graphical user interface (GUI) was Erwise in 1992. Then in 1993, the Mosaic web browser was introduced.[15] Mosaic is the browser responsible for popularizing browsing the web and influenced many of the browsers that are in

---

[10] Nicole Martin. "Digital Preservation: The World Wide Web [Presentation slides]."
[11] Kim Ann Zimmermann and Jesse Emspak. "Internet History Timeline: ARPANET to the World Wide Web."
[12] Life Noggin. "History of the Internet." YouTube video, 3:40. December 15, 2014. https://youtu.be/h8K49dD52WA.
[13] Life Noggin. "History of the Internet."
[14] Life Noggin. "History of the Internet."
[15] Life Noggin. "History of the Internet."

use today, including Google Chrome, Firefox, and Safari. From its beginnings as ARPANET to the present-day Internet and World Wide Web, computer networking and digital data distribution have moved from a technology that was limited to the few to a technology that is meant to be accessible to all or at least connect all who have an Internet connection.

A Brief History of Web Archiving

The introduction of the World Wide Web in 1991 "totally changed the way of publishing and broadcasting."[16] In the almost thirty years since its launch, an incomprehensible amount of information has been generated, stored, and shared via the Internet. The accessibility and pervasiveness of the World Wide Web tends to "foster an impression of permanence,"[17] which is a dangerous impression to have, as the web is highly dynamic and large amounts of data and information are lost every day. It is estimated that around 80 percent of Web content disappears or changes each year[18] and nearly 11 percent of public social media posts are lost after the first year of publishing and continue to be lost at a rate of 0.02 percent *per day* after the one-year mark.[19] As Richard Davis states in his article "Moving Targets: Web Preservation and Reference Management," the web contains "highly volatile content, easily modified or removed by its authors and editors, without any guarantee that previously published version, or any record of the

[16] Masashi Toyoda and Masaru Kitsuregawa. "The History of Web Archiving." https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6182575
[17] Alexis Antracoli, Steven Duckworth, Judith Silva, & Kristen Yarmey. "Capture All the URLs: First Steps in Web Archiving." https://pdfs.semanticscholar.org/2fa3/bde9acbbb21a0d77ef2dbc4ae159b69ace3d.pdf?_ga=2.227425501.133686681.1572972062-956655262.1572972062.
[18] Niels Brügger. "Archiving Websites: General Considerations and Strategies." January 2005. Centre for Internet Research. http://cfi.au.dk/fileadmin/www.cfi.au.dk/publikationer/archiving_underside/archiving.pdf.
[19] Hany M. SalahEldeen and Michael L. Nelson. "Losing My Revolution: How Many Resources Shared on Social Media Have Been Lost?" September 13, 2012. In: Theory and Practice of Digital Libraries. p. 125.

change will persist."[20]These are staggering numbers and all the more support the need for regular and sustained web archiving.

A web archive is "a record of web resources. It may include HTML and images, scripts, stylesheets, as well as video, audio and other elements that web pages and web apps are made of, all in one file."[21] Web archiving is "the complex process of harvesting Web content and then preserving it for the enduring future."[22] When archiving a website, the goal is to "capture and preserve the dynamic and functional aspects of Web pages—including active links, embedded media, and animation—while also maintaining the context and relationships between files."[23] Ideally, the user should be able to interact with the archived Web page as if it were "live." Web archiving should also attempt to capture the user experience. This means that web archiving has to continually be evolving, by consistently updating web archiving and crawling technologies and keeping up to date with the latest Web content.

In 1996, Brewster Kahle founded the Internet Archive, a non-profit dedicated to "building a digital library of Internet sites and other cultural artifacts in digital form."[24] The Internet Archive was the first archive of its kind and thus set the path for web archiving. In 1996, this began with archiving the World Wide Web itself, through the Alexa Internet project. Alexa Internet only captured text, which meant a portion of the World Wide Web was not being archived. The first full crawl occurred in 1997.[25] In 1999, Alexa Internet, both the company and

---

[20] Richard M. Davis. "Moving Targets: Web Preservation and Reference Management." January 30, 2010. In: *Ariadne*, Issue 62. http://www.ariadne.ac.uk/issue62/davis.
[21] Webrecorder. "FAQ." https://webrecorder.io/_faq.
[22] Alexis Antracoli, Steven Duckworth, Judith Silva, & Kristen Yarmey. "Capture All the URLs: First Steps in Web Archiving."
[23] Alexis Antracoli, Steven Duckworth, Judith Silva, & Kristen Yarmey. "Capture All the URLs: First Steps in Web Archiving."
[24] Internet Archive. "About the Internet Archive." https://archive.org/about/.
[25]Courtney Mumma. "Internet Archive: Archive-It and Contract Crawling." November 21, 2016. https://www.slideshare.net/NCDD/internet-archive-archiveit-and-contract-crawling-c-mumma.

the crawler software, were sold to Amazon and Kahle began developing a new crawling software for the Internet Archive that could collect audiovisual media and scanned books, in addition to the content it was already collecting.[26] In 2001, the Internet Archive's Wayback Machine was introduced. It is a search interface for the Internet Archive's archived collections of websites, allowing "people to visit archived versions of Web sites [between 1996 and the present]."[27] In 2003, the Internet Archive released Heritrix, an open source web crawling software that captures media, in addition to plain text. In 2006, the Internet Archive introduced Archive-It, a subscription-based web archiving service that is highly customizable based around the needs and wants of each subscribing organization.[28] Archive-It uses Heritrix crawling software, meaning it can also handle a wide range of content. In 2014, Umbra was introduced by Archive-It to run alongside Heritrix and solve some of Heritrix's previous issues.[29] Fast forward to 2016 when a new organization, Rhizome, in partnership with the New Museum, entered the web archiving technology game, releasing Webrecorder.[30] Seeing the issues that Archive-It's Heritrix has with capturing social media feeds and more dynamic Web content, Webrecorder was presented as "both a tool to create high-fidelity, interactive recordings of any web site browse[d] and a platform to make those recordings accessible."[31] Finally, in 2017, in what is a natural

---

[26] Nicole Martin. "Web Archiving [Presentation slides]." October 29, 2019. https://static1.squarespace.com/static/57c2e034ebbd1ac6f2565c18/t/5de7fb6dc172cb20cebe4400/1575484269669/2019+w8.2+Web+Archiving.pdf.
[27] Internet Archive. "Wayback Machine FAQs." http://web.archive.bibalex.org/web/policies/faq.html#wayback_what_is.
[28] Archive-It Blog. "About Us." https://archive-it.org/blog/learn- more/.
[29] Lori Donovan and Mary Haberle. "Web Archiving for Academic Institutions [Presentation slides]." April 23, 2018. https://digital.sandiego.edu/cgi/viewcontent.cgi?article=1193&context=symposium.
[30] Dragan Espenschied. "Rhizome Releases First Public Version of Webrecorder." Rhizome, August 9, 2016. http://rhizome.org/editorial/2016/aug/09/rhizome-releases-first-public-version-of- webrecorder/.
[31] Webrecorder. "FAQ."

progression from Heritrix and Umbra, as well as a possible response to Rhizome's Webrecorder, Archive-It released Brozzler, its latest web crawling technology.[32]

The Challenges Associated with Web Archiving and Web Crawling

Many of the challenges associated with web archiving and web crawling are the same or are extensions of the same issue. These challenges include: attempting to archive Web pages that are not designed to be archived; continuously competing and racing against time to capture a page before it changes, as the web is constantly changing; capturing dynamic content, especially content that is highly dependent on human interaction or includes Javascript, both of which are often present in Web sites using Flash; archiving pages that contain streaming and downloadable media, such as YouTube or Vimeo, as these media types may not be captured correctly or able to be played back on the archived Web page; crawling password-protected sites, however there are work-arounds and the work-arounds do not yet apply to two step authentication systems or sites with specialty certificates; capturing form- and database-driven content, which will still be accessible, but will generally not function properly in the archived page; encountering "robots.txt" exclusions that limit what a crawler can access and capture on a Web page; capturing and playing-back URLS that contain #s; and capturing the visuals of Web pages faithfully, often by excluding ad banners and images.[33] Web crawlers capture a Web page at a certain moment, freezing the page in time, yet the page will not be updated unless the page is "crawled" again. Therein lies an inherent problem with web archiving and web crawlers. The process must be repeated regularly, and while automation of the process is possible to integrate

[32] Lori Donovan and Mary Haberle. "Web Archiving for Academic Institutions [Presentation slides]."
[33] Jillian Lohndorf. "Known Web Archiving Challenges." Last updated June 2019. https://support.archive-it.org/hc/en-us/articles/209637043-Known-Web-Archiving-Challenges.

into a digital archival workflow, the fact that the process must be continuously repeated is not always known by the organizations who utilize this archival resource.

"The Web is so large that crawling a significant portion of it takes a large number of technical resources. The Web is changing so fast that portions of a website may change before a crawler has even finished crawling it."[34] In addition, crawling technology is constantly being tweaked and by the time archivists have gone through all of the necessary steps to prepare for web archiving and are ready to start, the technology has already changed. As one archivist put it, "Right now, we're 100 percent ready to archive the way the Web was 10 years ago."[35]

The Different Kinds of Web Crawlers Currently In Use

The ideal web crawler attempts to "crawl" by simulating how a user would interact with the webpage and the various kinds of content on the page. However, that is seldom how a web crawler functions, as the technology and capabilities of web crawlers varies and consistency between crawls is not guaranteed because the World Wide Web is ever-evolving and mutating.

In general, web crawlers "identify materials on the live Web that belong in a certain collection, based upon the user's choice of seeds and scope. A 'crawl' can also reference the archived content associated with the action. A 'crawler' refers to the automated agent, also called a robot or spider, that capture the live Web content. 'Crawling technology' refers to technology that helps crawlers capture content."[36] A chosen "seed" is "an item in Archive-It with a unique ID number. The Seed URL tells the crawler where to go on the live web and acts as an access point to archived content."[37] The chosen "scope" is "what the crawler will capture and what it

---

[34] Wikipedia. "Web archiving." https://en.wikipedia.org/wiki/Web_archiving#History_and_development.
[35] A. Bleicher. "A memory of webs past." 2011. In: *IEEE Spectrum*, 48(3), doi:10.1109/MSPEC.2011.571972, p. 37.
[36] Jillian Lohndorf. "Archive-It Crawling Technology." Last updated October 2019. https://support.archive-it.org/hc/en-us/articles/115001081186-Archive-It-Crawling-Technology.
[37] Maria Praetzellis. "Glossary of Archive-It and Web Archiving Terms." Last updated October 2019. https://support.archive-it.org/hc/en-us/articles/208111686-Glossary-of-Archive-It-and-Web-Archiving-Terms.

won't. Scoping refers to options for telling the crawler how much or how little of a seed URL to capture. Archive-It options include seed and collection level scoping."[38]

As stated earlier, in 2003, the Internet Archive released Heritrix as an open source web crawling software that captures media, in addition to plain text, when crawling Web sites to be archived. When initiating a crawl in Archive-It, "Heritrix crawls all seeds in the crawl simultaneously…Heritrix cycles through all the hosts from all the seed sites and embedded content."[39] Heritrix combines scoping, capture, data deduplication, and WARC creation in one process.[40] It produces WARC files, which is the file format successor to the Internet Archive's previous file format ARC.[41] WARC stands for "Web ARChive," whereas "ARC" stands for "ARChive." WARC files store "web crawls" as sequences of content blocks harvested from the World Wide Web. WARC files are open format, aggregate sets of digital resources that make up the Web. They allow for these original digital resources to be extracted from a WARC file, document changes to web content over time, detect duplicate content when compared to previous events, and store associated metadata, so that information about the content's creation is always stored with the WARC file.[42] Heritrix is scalable and flexible. However, it struggles to archive dynamic or interactive content.[43] Heritrix also cannot crawl sites that have robots.txt file exclusion requests and struggles to capture social media Web sites.[44]

In 2014, Umbra was introduced as a potential fix to some of Heritrix's limitations. Umbra is used as a tool, allowing Heritrix to access social media sites like a browser would. It runs alongside Heritrix and mimics the way a browser would access a page. Umbra executes client-

---

[38] Maria Praetzellis. "Glossary of Archive-It and Web Archiving Terms."
[39] Jillian Lohndorf. "Archive-It Crawling Technology."
[40] Lori Donovan and Mary Haberle. "Web Archiving for Academic Institutions [Presentation slides]."
[41] Nicole Martin. "Web Archiving [Presentation slides]."
[42] Nicole Martin. "Web Archiving [Presentation slides]."
[43] Nicole Martin. "Web Archiving [Presentation slides]."
[44] Jillian Lohndorf. "Archive-It Crawling Technology."

side scripts, allowing Heritrix to access previously undetectable URLs. Umbra also enables Heritrix to click or hover to execute Javascript and allows for dynamic scrolling, both of which imitate human interactions with Web sites that Heritrix was not capable of accomplishing prior to the addition of Umbra.[45]

While Heritrix, in combination with Umbra, is the most used web crawler for web archiving, other web crawling technologies include HTTrack and GNU Wget. HTTrack Website Copier is an open source offline browser that "downloads captured Web content onto a local directory."[46] Complaints about HTTrack include the crawler being slow and not being able to handle Flash sites or intensive Java/Javascript sites.[47] GNU Wget is a "free software package for retrieving files using HTTP, HTTPS, FTP, and FTPS (the most widely-used Internet protocols)."[48] It is used through the command line, can run on most UNIX-like operating systems, produces WARC files, and is designed to be used with slow or unstable network connections.[49] This report will not go into great detail about either web crawling technology, as neither are suitable for capturing dynamic Web content, such as social media sites, Wix sites, or sites containing Flash and thus are not entirely relevant to the discussion of the evolution of a web crawler that captures dynamic Web pages faithfully.

Webrecorder, released by Rhizome and the New Museum in 2016, is a free open source software that is very much centered on the individual user and the user's experience and ability to come back to a Web page that was previously archived. Webrecorder "creates an interactive copy of any Web page that is browsed, including content revealed by [the user's] interactions

---

[45] Lori Donovan and Mary Haberle. "Web Archiving for Academic Institutions [Presentation slides]."
[46] Alexis Antracoli, Steven Duckworth, Judith Silva, & Kristen Yarmey. "Capture All the URLs: First Steps in Web Archiving."
[47] HTTrack Website Copier. "FAQ." https://www.httrack.com/html/faq.html#QG0b.
[48] GNU Operating System. "GNU Wget: Introduction to GNU Wget." https://www.gnu.org/software/wget/.
[49] GNU Operating System. "GNU Wget: Introduction to GNU Wget."

such as playing video and audio, scrolling, clicking buttons, and so forth."[50] This means that in order for Webrecorder to capture all elements of a Web page, it must "rely on a human user browsing the live web."[51] Webrecorder, unlike Heritrix and incredibly similar to Brozzler, "focuses on dynamic web content such as embedded video and complex Javascript."[52]

Released sometime between 2016 and 2019, Brozzler is Archive-It's newest crawling technology. It is believed that Brozzler was in Beta testing until August 2019, when it was released to all Archive-It partners.[53] It combines the technologies of a browser and a crawler, hence "Brozzler," which is a combination of the two words "browser" and "crawler." Brozzler was developed to improve the capture of audio and video on Web pages during web crawls, as well as to archive Instagram feeds and Wix sites.[54] Brozzler differs from Archive-It's "standard" crawling technology, Heritrix and Umbra, in "its reliance on an actual web browser to interact with web content before that content is indexed and archived into WARC files."[55] Part of Brozzler's mission to improve the capture of audio and video was to be able to discover URLs generated by Javascript.[56] This is accomplished by "record[ing] interactions between servers and web browsers as they occur, [which] more closely resembl[es] how a human user would experience the web."[57] This is in contrast to following hyperlinks and downloading files, which are the actions of Heritrix and Umbra. Also, Brozzler uses "youtube-dl," a command-line

---

[50] Webrecorder. "FAQ."
[51] Lori Donovan and Mary Haberle. "Web Archiving for Academic Institutions [Presentation slides]."
[52] Lori Donovan and Mary Haberle. "Web Archiving for Academic Institutions [Presentation slides]."
[53] Karl-Rainer Blumenthal. "Feature release: Browser-based web capture for all Archive-It partners." August 27, 2019. https://support.archive-it.org/hc/en-us/community/posts/360050219071-Feature-release-Browser-based-web-capture-for-all-Archive-It-partners.
[54] Sylvie Rollason-Cass. "How and when to use Brozzler." Last updated September 2019. https://support.archive-it.org/hc/en-us/articles/360000351986-How-and-when-to-use-Brozzler.
[55] Rollason-Cass, Sylvie. "What is Brozzler?" https://support.archive-it.org/hc/en-us/articles/360000343186-What-is-Brozzler-.
[56] Levitt, Noah. "Brozzler [Presentation slides]." Keynote Speech presented at the IIPC Building Better Crawlers Hackathon, British Library, London, UK, September 22, 2016. http://archive.org/~nlevitt/reveal.js/#/.
[57] Rollason-Cass, Sylvie. "What is Brozzler?"

program that allows a user to download videos from YouTube and a few other Web sites,[58] to address the mission of capturing audio and video in formats that are stream-able and able to be played back.[59] Brozzler "runs on an instance of chromium browser. [It] opens [a] page in the [Web] browser, takes a screenshot, [and] sends [the screenshot] to warcprox, written as a WARC file."[60] According to Archive-It's website and Lori Donovan and Mary Haberle's presentation about web archiving for academic institutions, Brozzler's functions improve upon some of Heritrix and Umbra's shortcomings. In particular, Brozzler has improved the capture of dynamic and multimedia web content and has had more success capturing and playing-back URLS that contain #s.

My Own Experience Attempting to Use Brozzler

I initially was under the impression I could run Brozzler without having to subscribe to Archive-It. The plan was to run Brozzler on an API, a few complex websites, including a Wix site, and on an Instagram feed. I believed I could run Brozzler because of a section on Brozzler's GitHub site that provided several recipes under "Getting Started" stating they could be used to run Brozzler using the command line. The first recipe was running "brozzler-easy," which the README.rst stated was the easiest way to get started. The first step was installing and starting "rethinkdb," by using the "`brew install rethinkdb`" command in Terminal.[61] I was able to install this. However, the next step required me to install Brozzler, using the "`pip install brozzler[easy]`" command. I ran the command and it came back with "`–bash: pip: command not found`" each time I input the command. After ensuring I had "`pip`" and Python installed, I went to the "Issues" section of the GitHub to see if anyone else was encountering this

[58] GitHub—youtube-dl." What is it?" https://ytdl-org.github.io/youtube-dl/about.html.
[59] Levitt, Noah. "Brozzler [Presentation slides]."
[60] Lori Donovan and Mary Haberle. "Web Archiving for Academic Institutions [Presentation slides]."
[61] GitHub—internetarchive / brozzler. "README.rst." https://github.com/internetarchive/brozzler.

roadblock. It turned out I was not the only one. "primoz-k" had the same issue where "pip install brozzler[easy]" returned the result of "command not found."[62] The user's issue was eventually solved by changing the command slightly to "`pip install "brozzler[easy]".`" I ran this variation on the command with no success. Terminal still said "`-bash: pip: command not found.`" After troubleshooting a little more and continuing to tweak the command with no success, the only conclusion I could come up with is that you must have to run Brozzler with the Archive-It software installed or connected to be able to call forward Brozzler in Terminal. Not being able to test the crawling technology on my own, I had to rely on other user's issues submitted to Brozzler's GitHub. Based on the 11 open issues and 19 closed issues, there are very few negative issues relating to Brozzler. The issues tend to be about wanting to customize the commands for specific purposes, with very few issues concerning Brozzler as a web crawler successfully (or unsuccessfully) capturing audio and video on Web pages. With very little firsthand experience and knowledge to go off of, I am not able to make a definitive decision on the success of Brozzler as a web crawling tool meant to capture dynamic Web pages.

In future research, I would like to find a successful way to test Brozzler and its purported functionalities. As Brozzler has just recently been released from Beta testing, I believe more time is needed to fully work through the possible glitches before a faithful testing can occur. It would also be beneficial to test it against Heritrix and Umbra, as well as against Webrecorder. Understanding that there is very little barrier to using Webrecorder as an independent user, I am interested to compare the user experience of Brozzler to Webrecorder.

[62] primoz-k. "brozzler[easy] not found." April 20, 2017. https://github.com/internetarchive/brozzler/issues/35.

Conclusion

The creation of the Internet and the World Wide Web were catalysts for the creation of the Internet Archive and the beginning of web archiving and web crawling. The evolution of web crawling technology mirrors the evolution of the World Wide Web and the continued proliferation and expansion of what is possible to create, share, post, and store on the Web. Umbra as an update to Heritrix signified the increase in social media usage. The creation of both Brozzler and Webrecorder represented the move towards more dynamic Web pages and the popularity of streaming video and audio. I am still under the impression that there has yet to be a web crawler that is entirely successful at capturing the true essence of a Web page. Each crawler has its issues, which is evidenced by the evolution of the crawling technology, the continuous minor tweaks and updates to each technology, and the current presence of both Webrecorder and Brozzler, which are aiming to solve the same web crawling issues. This leads me to question what will be the next web crawler and what changes to the content on the World Wide Web that technology will signify.

Bibliography

Aaron. “How the Internet Works in 5 Minutes.” *YouTube* video, 4:48. No date. https://www.youtube.com/watch?v=7_LPdttKXPc.

Abbate, Janet. *Inventing the Internet*. Cambridge, Massachusetts: The MIT Press, 1999.

Antracoli, Alexis, Steven Duckworth, Judith Silva, & Kristen Yarmey. “Capture All the URLs: First Steps in Web Archiving.” https://pdfs.semanticscholar.org/2fa3/bde9acbbb21a0d77ef2dbc4ae159b69ace3d.pdf?_ga=2.227425501.133686681.1572972062-956655262.1572972062.

Archive-It Blog. “About Us.” https://archive-it.org/blog/learn- more/.

Bellis, Mary. “ARPANET - The First Internet.” The Cold War and ARPANET. http://ocean.otr.usm.edu/~w146169/bellis.html.

Bleicher, A. “A memory of webs past.” 2011. In: IEEE Spectrum, 48(3), 30-37. doi:10.1109/MSPEC.2011.5719723.

Blumenthal, Karl-Rainer. “Feature release: Browser-based web capture for all Archive-It partners.” August 27, 2019. https://support.archive-it.org/hc/en-us/community/posts/360050219071-Feature-release-Browser-based-web-capture-for-all-Archive-It-partners.

Brügger, Niels. “Archiving Websites: General Considerations and Strategies.” January 2005. Centre for Internet Research. http://cfi.au.dk/fileadmin/www.cfi.au.dk/publikationer/archiving_underside/archiving.pdf

Davies, Donald Watts. Oral history interview by Martin Campbell-Kelly. Mach 17, 1986, National Physical Laboratory, UK. Charles Babbage Institute. Retrieved from the University of Minnesota Digital Conservancy. http://hdl.handle.net/11299/107241.

Davis, Richard M. “Moving Targets: Web Preservation and Reference Management.” January 30, 2010. In: Ariadne, Issue 62. http://www.ariadne.ac.uk/issue62/davis.

Donovan, Lori and Mary Haberle. “Web Archiving for Academic Institutions [Presentation slides].” April 23, 2018. https://digital.sandiego.edu/cgi/viewcontent.cgi?article=1193&context=symposium

Espenschied, Dragan. “Rhizome Releases First Public Version of Webrecorder.” Rhizome, August 9, 2016. http://rhizome.org/editorial/2016/aug/09/rhizome-releases-first-public-version-of- webrecorder/.

GitHub—internetarchive / brozzler. “README.rst.” https://github.com/internetarchive/brozzler.

GitHub—youtube-dl.” What is it?” https://ytdl-org.github.io/youtube-dl/about.html.

GNU Operating System. “GNU Wget: Introduction to GNU Wget.” https://www.gnu.org/software/wget/.

HTTrack Website Copier. “FAQ.” https://www.httrack.com/html/faq.html#QG0b.

Internet Archive. “About the Internet Archive.” https://archive.org/about/.

Internet Archive. “Wayback Machine FAQs.” http://web.archive.bibalex.org/web/policies/faq.html#wayback_what_is.

Leiner, Barry M., Vinton G. Cerf, David D. Clark, Robert E. Kahn, Leonard Kleinrock, Daniel C. Lynch, Jon Postel, Larry G. Roberts, and Stephen Wolff. “Brief History of the Internet.” 1997. https://www.internetsociety.org/internet/history-internet/brief-history-internet/.

Levitt, Noah. “Brozzler [Presentation slides].” Keynote Speech presented at the IIPC Building Better Crawlers Hackathon, British Library, London, UK, September 22, 2016. http://archive.org/~nlevitt/reveal.js/#/.

Life Noggin. “History of the Internet.” *YouTube* video, 3:40. December 15, 2014. https://youtu.be/h8K49dD52WA.

Lohndorf, Jillian. “Archive-It Crawling Technology.” Last updated October 2019. https://support.archive-it.org/hc/en-us/articles/115001081186-Archive-It-Crawling-Technology.

Lohndorf, Jillian. “Known Web Archiving Challenges.” June 2019. https://support.archive it.org/hc/en-us/articles/209637043-Known-Web-Archiving-Challenges.

Martin, Nicole. “Digital Preservation: The World Wide Web [Presentation slides].” October 29, 2019. https://static1.squarespace.com/static/57c2e034ebbd1ac6f2565c18/t/5de7fb5088321b3df06c55dd/1575484244770/2019+w8.1+Web+Intro.pdf.

Martin, Nicole. “Web Archiving [Presentation slides].” October 29, 2019. https://static1.squarespace.com/static/57c2e034ebbd1ac6f2565c18/t/5de7fb6dc172cb20cebe4400/1575484269669/2019+w8.2+Web+Archiving.pdf.

Mumma, Courtney. “Internet Archive: Archive-It and Contract Crawling.” November 21, 2016. https://www.slideshare.net/NCDD/internet-archive-archiveit-and-contract-crawling-c-mumma.

Praetzellis, Maria. “Glossary of Archive-It and Web Archiving Terms.” Last updated October 2019. https://support.archive-it.org/hc/en-us/articles/208111686-Glossary-of-Archive-It-and-Web-Archiving-Terms.

primoz-k. “brozzler[easy] not found.” April 20, 2017. https://github.com/internetarchive/brozzler/issues/35.

Rollason-Cass, Sylvie. “How and when to use Brozzler.” https://support.archive-it.org/hc/en us/articles/360000351986-How-and-when-to-use-Brozzler.

Rollason-Cass, Sylvie. “What is Brozzler?” https://support.archive-it.org/hc/en us/articles/360000343186-What-is-Brozzler-.

SalahEldeen, Hany M. and Michael L. Nelson. “Losing My Revolution: How Many Resources Shared on Social Media Have Been Lost?” September 13, 2012. In: *Theory and Practice of Digital Libraries*. pp. 125–137.

Toyoda, Masashi and Masaru Kitsuregawa. “The History of Web Archiving.” https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6182575

Webrecorder. “FAQs.” https://webrecorder.io/_faq.

Wikipedia. “Web archiving.” https://en.wikipedia.org/wiki/Web_archiving#History_and_development

Zimmermann, Kim Ann and Jesse Emspak. “Internet History Timeline: ARPANET to the World Wide Web.” June 27, 2017. https://www.livescience.com/20727-internet-history.html.