Digital Preservation: CINE-GT 1807
Fall 2017
Caroline Z. Oliveira
12/15/2017

Itsy-Bitsy Spider: A Look at Web Crawlers and Web Archiving

Introduction

Human beings have always felt the need to share information. As societies and technologies advanced, it became clear the world would benefit immensely from a system that would allow ideas and works to be disseminated with great speed, individuals to cooperate on projects regardless of geographical location, and worldwide broadcasting. The first description of what the Internet came to be was recorded in a series of written memos written by J.C.R. Licklider of MIT in 1962, but a year before that Leonard Kleinrock published "Information Flow in Large Communication Nets," a paper about the U.S. Defense Department's Advanced Research Projects Agency Network (ARPANET,) the predecessor of the Internet.[1] Eventually turned into a reality by the Advanced Research Projects Agency (ARPA, later known as DARPA), ARPANET remained in service from the late 1960s to 1990, and it "protected the flow of information between military installations by creating a network of geographically separated computers that could exchange information via a newly developed protocol (rule for how computers interact) called NCP (Network Control Protocol)."[2] This advancement and the events that took place until the 1980s mark the development of the Internet as it is known today, "a widespread information infrastructure, the initial prototype of what is often called the National (or Global or Galactic) Information Infrastructure."[3] Although the Internet had its beginnings over 50 years ago, it was only in the late 1990s that people started realizing the importance of archiving and preserving digital-born materials and today, web crawlers allow individual users and organizations to do just that on a larger or smaller scale. An overview of web archiving and its tools is a necessary step an individual or organization should take to determine which software will best fulfill their specific preservation needs, especially since each technology has its advantages and disadvantages.

---

[1] Zimmermann et al., "Internet History Timeline."
[2] Bellis, "ARPANET - The First Internet."
[3] Leiner et al., "Brief History of the Internet."

Defining Crawling

Crawling or spidering is the process of "exploring web applications automatically [where] the web crawler aims at discovering the webpages of a web application by navigating through the application, [which] is usually done by simulating the possible user interactions considering just the client-side of the application."[4] In a more practical explanation, web crawlers, which are also known as web spiders or web robots, are software that start out by visiting a list of specific Universal Resources Locators, URLs. These URLs are called *seeds* and the crawlers when used for archiving, copy and save the data associated with them. Next, the crawler locates links associated with the URLs and adds them to the seed list. The archived data can then be seen as snapshots of these websites and will not change if the live site gets updated or changed unless the web crawler crawls it again.

The Internet Archive and the Wayback Machine

For many decades, online content was disappearing just as fast as it was being created. In fact, "the average life of a web page is only 77 days,"[5] and according to several studies:

- 80 % of web pages are not available in their original form after one year
- 13 % of web references in scholarly articles disappear after 27 months
- 11 % of social media resources, such as the ones posted on Twitter, are lost after one year.[6]

With the Internet increasingly becoming indispensable in all aspects of human life, Brewster Kahle, Bruce Gilliat, and other technologists realized it was essential to preserve its content. In 1996 they founded the Internet Archive, a 501(c)(3) nonprofit organization that aims to archive content from the World Wide Web on a scale that had never been attempted before. With preservation and access being its main focus, the Internet Archive was recognized as a library by the State of California, which solidified its association with other institutions that share the same objectives. On the other hand, the Archive differs from other traditional libraries for its unique collection and for providing online access to all its materials and "by developing tools and services that help others make content in their own collections more widely available."[7] Today the Archive manages:

- 279 billion web pages

---

[4] Mirtaheri et al., "A Brief History of Web Crawlers."
[5] "Internet Archive Wayback Machine."
[6] Costa, Gomes, and Silva, "The Evolution of Web Archiving."
[7] Rackley, "Internet Archive."

- 11 million books and texts
- 4 million audio recordings (including 160,000 live concerts)
- 3 million videos (including 1 million Television News programs)
- 1 million images
- 100,000 software programs.[8]

Though the Archive has had many contributors since its creation, an important institutional partner in its operations was Alexa Internet. Founded in 1996 by Kahle and Gilliat, the two men behind the Internet Archive itself, Alexa is a "web traffic information, metrics and analytics provider"[9] that promised to:

> Banish '404 not found' messages for its members by retrieving stale pages from the [Internet Archive]. It [offered] guidance on where to go next, based on the traffic patterns of its user community -- putting in sidewalks where the footpaths are. Alexa also [offered] context for each site visited: to whom it's registered, how many pages it has, how many other sites point to it, and how frequently it's updated.[10]

The Archive started using Alexa Internet's proprietary crawler to capture content and in 2001 the two entities made the Wayback Machine, "a three-dimensional index that allows browsing of web documents over multiple time periods,"[11] available to the public.

The Wayback Machine allows users to visit archived versions of websites, but it is not infallible. Certain pages contain broken links, missing images, or could not be archived at all. These are normally caused by:

- Robots.txt: A website's robots.txt document will probably prevent crawling.
- Javascript: Javascript elements are normally difficult to capture and archive, but especially if they produce links without having the full name on the web page. Playback is also a concern.
- Server-side image maps: Similar to other functions on the Internet, if a page needs to contact the originating server in order to make images load properly, it will not be successfully archived.
- Orphan pages -- If there are no links to the page in question, the crawler will not be able to locate it.[12]

---

[8] "Internet Archive: About IA."
[9] "What Is Alexa Internet?"
[10] Dawson, "Alexa Internet Opens the Doors."
[11] "Internet Archive Frequently Asked Questions."
[12] Ibid.

Archive-It

Launched in 2006 by the Internet Archive, Archive-It is a subscription-based web archiving service that assists various organizations to collect, construct, and preserve collections of digital materials. Currently, there are "over 400 partner organizations in 48 U.S. states and 16 countries worldwide"[13] that use this service. Users have access to their collections 24/7 and the ability to perform full-text searches.

Organizations that use Archive-It are able to determine the frequency and scope of crawls, produce reports, and customize Dublin Core metadata fields. In addition, the service "advertises the ability due to collect a wide range of content, including HTML, images, video, audio, PDF, and social networking sites,"[14] with the data collected being stored in multiple servers as WARC (ARC) files.

In fact, two copies of each created file are stored in the Internet Archive's data centers, and Archive-It has partnerships with LOCKSS and DuraCloud, giving users a "trustworthy digital preservation program for web content."[15] Archive-It also prides itself on being able work with organizations to deliver the best plan that meets their web archiving needs and budget.

Heritrix

As mentioned previously, the Internet Archive used Alexa Internet's proprietary crawler, but the software had its limitations. It was not able to perform crawls internally for the Archive and it also did not facilitate cooperation between different institutions.[16] In 2003, the Archive started developing an open-source, extensible, web-scale, archival-quality web crawler written in Java, which they named Heritrix.[17] Heritrix, which is sometimes spelled heretrix, or completely misspelled, aims to preserve and collect data to be utilized by future generations and adequately got its name from the archaic word for "heiress, or woman who inherits."[18] Currently, Heritrix is on its 3.2.0 version, which was released on January 2014, and it is said to be "most suitable for advanced users and projects that are either customizing Heritrix (with Java or other scripting code) or embedding Heritrix in a larger system."[19]

In order to meet the Archive's expectations, this new crawler needed to be able to execute:

- Broad crawling: Broad crawls are high-bandwidth crawls that focus both on the number of sites collected and the completeness with which any one site is captured. Broad crawls

---

[13] "About Us."
[14] "Archive-It."
[15] Slater, "Review: Archive-It."
[16] Mohr et al., "An Introduction to Heritrix - An Open Source Archival Quality Web Crawler."
[17] Rackley, "Internet Archive."
[18] "Heritrix - Home Page."
[19] Mohr and Levitt, "Release Notes - Heritrix 3.2.0."

attempt to take as many samples as possible of web pages given the time and storage resources available.

- Focused crawling: Focused crawls are small- to medium-sized crawls (usually less than 10 million unique documents) in which the crawler focus on the complete coverage of specific sites or subjects.
- Continuous crawling: Unlike traditional crawling, which captures snapshot of specific subjects, "downloading each unique URI one time only," continuous crawling goes back to previously captured pages, looks for changes and updates, fetches new pages, and estimates how frequently changes are made.
- Experimental crawling: The Internet Archive and other organizations wanted to experiment with different crawling techniques, being able to control what to crawl, order in which resources are crawled, crawling while using different protocols, and analysis and archiving of crawl results.[20]

As mentioned previously, Java was chosen as the software language due to its large developer community and open source libraries. Java also "offers strong support for modular design and components that are both incrementally extendable and individually replaceable,"[21] characteristics that would help the Archive accomplish its objectives. When it comes to its architecture, Heritrix was designed to be pluggable and allow customization and contributions from different parties, both which help when performing different types of crawling.

However, Heritrix is not without flaws. Currently, it does not support continuous crawling and it is not dynamically scalable, which means one needs to determine the number of servers taking part in the scheme before one can start crawling. Furthermore, "if one of the machines goes down during your crawl you are out of luck."[22] Heritrix also lacks in flexibility when it comes to output options, only exporting ARC/WARC files. WARC or Web ARChive file format is the successor of ARC, a "format that has traditionally been used to store 'web crawls' as sequences of content blocks harvested from the World Wide Web"[23] and for now, writing data to other formats would require changes in the source code. WARC, an ISO format, ISO 28500:2009, has four required fields:

- Record identifier (i.e., URI): A globally unique identifier assigned to the current WARC record.
- Content length/ record body size: The length of the following Record Content Block in bytes (octets).
- Date: A timestamp in the form of YYYY-MM-DD and hh:mm:ss indicating when the record was created.

---

[20] Mohr et al., "An Introduction to Heritrix - An Open Source Archival Quality Web Crawler."
[21] Ibid.
[22] Pedchenko, "Comparison of Open Source Web Crawlers."
[23] "WARC, Web ARChive File Format."

- WARC record type: The WARC Format Standard supports the 'warcinfo', 'response', 'resource', 'request', 'metadata', 'revisit', 'conversion.'[24]

One of the primary issues with this format, however, is the problem of scalability, "especially for content that needs to be manually inputted to be compatible with Preservation Metadata: Implementation Strategies (PREMIS)."[25]

HTTrack and Wget

Currently, Heritrix is one of the most popular web crawlers for web archiving, but other institutions also use HTTrack and Wget. HTTrack Website Copier is a free offline browser utility, which enables the user to download the contents of entire websites from the Internet to a local directory for offline viewing by "building recursively all directories, getting HTML, images, and other files from the server to [a] computer."[26] However, HTTrack is not perfect. Many of its users complain this crawler is very slow and although the offline browsing works adequately for flat HTML, the user "will need appropriate web-server technologies installed for scripting, JAVA, PHP, or other server-side includes to work."[27]

To operate HTTrack, the user starts out by choosing a filename and a destination folder for the project. The next step is to select an action, which can be to simply download a website, download the website and ask the user if any links are potentially downloadable, only download desired files, download all the sites in pages, test all indicated links, continue an interrupted download, or update an existing download.[28] Before the user is able to start downloading, he or she is able to determine if they want to start the process immediately or at a later date and indicate if he or she wants the server to disconnect after the download is completed or even shut down the computer. Finally, log files generated by HTTrack allow the user to identify potential errors.

Just like other crawlers, HTTrack has its limitations. Currently, it cannot handle:

- Flash sites.
- Intensive Java/Javascript sites.
- Complex CGI with built-in redirect, and other tricks.
- Parsing problem in the HTML code (cases where the engine is fooled, for example by a false comment (<!--) which has no closing comment (-->) detected.[29]

---

[24] "The WARC Format Explained."
[25] Corrado and Sandy, *Digital Preservation for Libraries, Archives, and Museums*.
[26] "HTTrack Website Copier."
[27] Catling, "Review."
[28] "HTTrack Website Copier - Offline Browser."
[29] "F.A.Q."

Though HTTrack's Frequently Asked Questions page is extensive and detailed, it seems that the main concern regarding this crawling tool is its slow speed in completing a project and opening the saved pages. Based on its low cost and functionality, HTTRack is considered an adequate tool for archiving and can be useful to smaller institutions.

Wget is "a free software package for retrieving files using HTTP, HTTPS, FTP, and FTPS the most widely-used Internet protocols [and] it is a non-interactive command line tool, so it may easily be called from scripts, cron jobs, terminals without X-Windows support, etc."[30] In other words, Wget does not require the user to be logged on in order for it to work.

This software was also designed to work with slow or unstable network connections. In case the network is faulty and the download fails, Wget "will keep retrying [to connect] until the whole file has been retrieved."[31]

Brozzler

Brozzler, which got its name from the mash of the words crawler and browser, was developed with a grant from the Andrew W. Mellon Foundation to improve the capture of audio and video in web archiving. According to Levitt, there are two categories of challenges when it comes to capturing audio-visual materials:

- Issues specific to audio-visual media (e.g. streaming formats)
- General web crawling challenges (e.g. discovering URLs generated by Javascript).[32]

To assist with the first issue, Brozzler uses youtube-dl, a command-line program that downloads videos and audio files from streaming websites such as YouTube.com, Vimeo, and Dailymotion. It can be operated on Linux, Windows, and Mac OS X systems, but it requires Python interpreter, version 2.5 or higher, Chromium or Google Chrome browser, and RethinkDB, open-source, scalable database, deployment to work.[33]

For the second issue, Brozzler uses a real browser (chrome or chromium) to load web pages, in fact, it "differs from Heritrix and other crawling technologies in its reliance on an actual web browser to render and interact with web content before all of that content indexed and archived into WARC files."[34]

Brozzler is also designed to work in conjunction with warcprox. Reviews on this software are still hard to find so it will be interesting to see if users will be satisfied with it.

---

[30] "GNU Wget."
[31] "GNU Wget 1.18 Manual."
[32] Levitt, "Brozzler."
[33] *Brozzler*.
[34] Lohndorf, "Archive-It Crawling Technology."

Webrecorder and "Human-Centered" Web Archiving

Founded by the Andrew W. Mellon Foundation, Rhizome released in 2016 the first public version of its Webrecorder, a free online tool that aims to tackle the issue of archiving challenging digital content. Developed by the company in partnership with programmer Ilya Kreymer, Webrecorder is "a human-centered archival tool to create high-fidelity, interactive, contextual archives of social media and other dynamic content, such as embedded video and complex javascript."[35] Since most tools in web archiving deliver static files when crawling websites, they are unable to capture material that keeps changing or being constantly updated, such as social media websites. Furthermore, a regular crawler is unable to capture the whole experience of a website, since it cannot browse a site as a real user would. Although crawlers are faster, they make:

> Decisions based on certain heuristics, and often times it [does not] run javascript, and even if it does, it follows a specific pattern, and [cannot] use the site the way an actual user would. And so, the archive of the site is often incomplete, and different than how an actual user would use the site.[36]

Internet users are advised to sign up for a free account with 5GB of storage space at webrecorder.io. and it is possible to download the Webrecorder Desktop Player Application, which allows the users to browse through their archived material offline. However, one can also utilize the software without registering and create temporary collections anonymously. Web archives created with the software are downloaded as WARC files, which come with all the required metadata mentioned previously in the Heritrix section. In fact, existing standard WARC (or ARC) files can be imported into Webrecorder and added to collections.[37]

The process of creating the recordings seems to be flexible. Each time a user hits the "record" button, the software creates recording sessions, which are the smallest units of a web collection in Webrecorder. The user is able to add multiple recordings to a single collection, rename, move, and even delete files.

Webrecorder is a free open-source software, but users should be aware of its terms and policies. In its "Terms of Use," Rhizome makes it clear it has the right "to terminate or restrict access to [one's] account and to delete or disable access to any links created and/or content stored in connection with the account, in [their] sole discretion, without advance notice, and shall have no liability for doing so."[38] Rhizome also notes that it reserves the right to suspend service without telling its users beforehand. These aspects could certainly generate many issues for an individual or repository that depend on this software to meet his/its web archiving needs.

---

[35] "Rhizome Awarded $600,000 by The Andrew W. Mellon Foundation to Build Webrecorder."
[36] McKeehan, "Symmetrical Web Archiving with Webrecorder, a Browser-Based Tool for Digital Social Memory. An Interview with Ilya Kreymer."
[37] Espenschied, "Rhizome Releases First Public Version of Webrecorder."
[38] "Terms and Policies."

Another aspect that could be considered a disadvantage of the Webrecorder is directly related to its human-centricity. To operate the software, a user has to browse the desired website after pressing Webrecorder's record button. Consequently, the user has to manually open every link, play every video, and access all different parts of the website in order for it to be archived entirely. For this reason, the process can be very time consuming.

Challenges of Web Archiving

Every form of archiving poses challenges and web archiving is no different. Not only is the web constantly changing, but there are also websites that were not built to be archived at all. Although web crawlers have made a lot of progress with their capturing abilities, according to Jillian Lohndorf, web archivists still have to deal with the following issues:

- Streaming and Downloadable Media: Although some crawlers are able to capture this type of media, playback is still an issue, especially on websites that have a large volume of videos and media.
- Password-Protected Websites: These websites were especially designed to avoid outside users from accessing its information. Apparently, there are ways some crawlers (like Heritrix) can access these protected sites, but the user has to use his or her login information to access the content before being able to initiate the crawl.
- Form and Database-Driven Content: If the user has to execute an action to interact with a website (i.e., use its search bar), crawlers might not be able to properly capture its contents. Lohndorf also notes that Archive-It has a particularly difficult time trying to capture websites with POST requests, which "[submit] data to be processed to a specified resource."[39]
- Robots.txt Exclusions: Certain websites might have been designed to prevent crawling all together. These websites in question have a robot.txt exclusion and crawlers tend to respect that by default. However, there are ways users can set up rules to ignore these robots either by contacting the webmaster to ask for an exception or, in Heritrix's case, the crawler's support team.
- Dynamic Content: Many crawlers are able to capture websites with dynamic content, but again, playback remains an issue. There are also certain types of content that are still difficult to archive, including:
    - Images or text size that adjust accordingly to the size of the browser
    - Maps that allows the user to zoom in and out
    - Downloadable files
    - Media with a "play" button
    - Navigation menus [40]

[39] "HTTP Methods GET vs POST."
[40] Lohndorf, "5 Challenges of Web Archiving."

Conclusion

Although it can still be considered a recent endeavor, web archiving has become an indispensable activity in today's world. People are constantly uploading photos, sharing their thoughts and ideas, and seeking answers to their questions in all sorts of websites, all of which make the Internet an important tool in understanding current and past cultural and socio-economic trends. Founded in the 1990s, the Internet Archive aimed to preserve the contents of the web, make all of its collections available to the public, and provide services that would allow other institutions to capture their own materials. Through the years, it has launched the Wayback Machine, Archive-It, and Heritrix, an open-source crawler that has been widely used all over the world. In fact, users currently have several options when it comes to finding a web crawler to meet their preservation needs. Besides Heritrix, a user might find HTTrack, a free offline browser, or Wget, a non-interactive command line tool, more user-friendly and capable of handling smaller operations. Other options include Brozzler, which uses a real browser during crawls, and Webrecorder, an operation that records the user's interaction with a web page. No matter which crawler the user or the organization might end up opting for, it is important to recognize the challenges and limitations related to web archiving in general, such as its difficulties in capturing and executing playback for dynamic content and downloadable media, dealing with websites that contain robots.txt exclusions, crawling password-protected sites, and handling form and database-driven content. However, since technological advances are being made every day, many believe that soon web crawlers will be able to fully capture all different aspects of web pages, allowing organizations to preserve one of the most ephemeral and mutating services in current society, the Internet.

Bibliography

"About Us." *Archive-It Blog* (blog). Accessed November 5, 2017. https://archive-it.org/blog/learn-more/.

"About Us." Archive-It. Accessed December 11, 2017. https://archive-it.org/blog/learn-more/.

"Archive-It." Digital Curation Centre. Accessed December 11, 2017. http://www.dcc.ac.uk/resources/external/archive-it.

Bellis, Mary. "ARPANET - The First Internet." The Cold War and ARPANET. Accessed December 9, 2017. http://ocean.otr.usm.edu/~w146169/bellis.html.

Brozzler: Distributed Browser-Based Web Crawler. Python. 2015. Reprint, Internet Archive, 2017. https://github.com/internetarchive/brozzler.

Catling, Robin. "Review: WebHTTrack Website Copier and Offline Browser (Re-Post)." *Everything Express* (blog), February 20, 2010. https://everythingexpress.wordpress.com/2010/02/20/review-webhttrack-website-copier-and-offline-browser-re-post/.

Corrado, Edward M., and Heather M. Sandy. *Digital Preservation for Libraries, Archives, and Museums*. Second Edition. Rowman & Littlefield Publishers. Accessed November 20, 2017. https://books.google.com/books?id=7GnEDQAAQBAJ&pg=PA292&lpg=PA292&dq=wget+preservation&source=bl&ots=bcUNQ16LpV&sig=asIDMyfd5t45Oy-8aKlLTZwDhy4&hl=en&sa=X&ved=0ahUKEwjNvP6kvM3XAhUE4CYKHf2aA-0Q6AEISDAG#v=onepage&q=wget%20preservation&f=false.

Costa, Miguel, Daniel Gomes, and Mário J. Silva. "The Evolution of Web Archiving." Springer-Verlag, May 9, 2016. https://link-springer-com.proxy.library.nyu.edu/content/pdf/10.1007%2Fs00799-016-0171-9.pdf.

Dawson, Keith. "Alexa Internet Opens the Doors." Tasty Bits from the Technology Front, July 27, 1997. http://www.tbtf.com/archive/1997-07-28.html#s04.

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/.

"F.A.Q." HTTrack Website Copier - Offline Browser. Accessed December 10, 2017. https://www.httrack.com/html/faq.html#QG0b.

"GNU Wget." GNU. Accessed November 19, 2017. https://www.gnu.org/software/wget/.

"GNU Wget 1.18 Manual." Accessed December 11, 2017. https://www.gnu.org/software/wget/manual/wget.html.

Gonzalez, Ricardo Garcia. “Youtube-Dl.” GitHub. Accessed December 10, 2017. https://rg3.github.io/youtube-dl/.

“Heritrix - Home Page.” Accessed November 17, 2017. http://crawler.archive.org/index.html.

“HTTP Methods GET vs POST.” Accessed December 11, 2017. https://www.w3schools.com/tags/ref_httpmethods.asp.

“HTTrack Website Copier.” Accessed November 19, 2017. https://www.httrack.com/.

“HTTrack Website Copier - Offline Browser.” Accessed December 10, 2017. https://www.httrack.com/html/step2.html.

“HTTrack/WebHTTrack.” Accessed December 10, 2017. http://freshmeat.sourceforge.net/projects/httrack.

“Internet Archive: About IA.” Accessed November 12, 2017. https://archive.org/about/.

“Internet Archive Frequently Asked Questions.” Accessed November 17, 2017. http://archive.org/about/faqs.php#The_Wayback_Machine.

“Internet Archive Wayback Machine.” Accessed November 11, 2017. 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.” *Internet Society* (blog). Accessed December 9, 2017. https://www.internetsociety.org/internet/history-internet/brief-history-internet/.

“Leonard Kleinrock And The ARPANET.” Accessed December 9, 2017. https://www.livinginternet.com/i/ii_kleinrock.htm.

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

Lohndorf, Jillian. “5 Challenges of Web Archiving.” Archive-It Help Center. Accessed November 5, 2017. http://support.archive-it.org/hc/en-us/articles/209637043-5-Challenges-of-Web-Archiving.

“Archive-It Crawling Technology.” Archive-It Help Center. Accessed November 4, 2017. http://support.archive-it.org/hc/en-us/articles/115001081186-Archive-It-Crawling-Technology.

McKeehan, Morgan. “Symmetrical Web Archiving with Webrecorder, a Browser-Based Tool for Digital Social Memory. An Interview with Ilya Kreymer.” Accessed December 10, 2017. https://ndsr.nycdigital.org/symmetrical-web-archiving-with-webrecorder-a-browser-based-tool-for-digital-social-memory-an-interview-with-ilya-kreymer/.

Mirtaheri, Seyed M., Mustafa Emre Dincturk, Salman Hooshmand, Gregor V. Bochmann, and Guy-Vincent Jourdan. “A Brief History of Web Crawlers.” University of Ottawa, May 5, 2014. https://arxiv.org/pdf/1405.0749.pdf.

Mohr, Gordon, and Noah Levitt. “Release Notes - Heritrix 3.2.0.” Heritrix. Accessed November 17, 2017. https://webarchive.jira.com/wiki/spaces/Heritrix/pages/13467786/Release+Notes+-+Heritrix+3.2.0.

Mohr, Gordon, Michael Stack, Igor Ranitovic, Dan Avery, and Michele Kimpton. “An Introduction to Heritrix - An Open Source Archival Quality Web Crawler,” 2004. http://crawler.archive.org/Mohr-et-al-2004.pdf.

Pedchenko, Aleks. “Comparison of Open Source Web Crawlers.” *Aleks Pedchenko* (blog), April 6, 2017. https://medium.com/@pedchenko/comparison-of-open-source-web-crawlers-62a072308b53.

Rackley, Marilyn. “Internet Archive.” In *Encyclopedia of Library and Information Sciences*, Third Edition., 2966–76. Taylor & Francis, 2009. https://archive.org/stream/internetarchive-encyclis/EncycLisInternetArchive#page/n0/mode/2up.

“Rhizome Awarded $600,000 by The Andrew W. Mellon Foundation to Build Webrecorder.” Rhizome, January 4, 2016. http://rhizome.org/editorial/2016/jan/04/webrecorder-mellon/.

Slater, Jillian M. “Review: Archive-It.” Marian Library/ IMRI Faculty Publications, September 18, 2014. http://ecommons.udayton.edu/cgi/viewcontent.cgi?article=1002&context=imri_faculty_publications.

Stern, Hunter. “Archiving Rich-Media Content.” Heritrix. Accessed November 17, 2017. https://webarchive.jira.com/wiki/spaces/Heritrix/pages/5735910/Archiving+Rich-Media+Content.

“Terms and Policies.” Webrecorder. Accessed December 10, 2017. https://webrecorder.io/_policies.

“The Cold War and ARPANET.” Accessed December 9, 2017. http://ocean.otr.usm.edu/~w146169/bellis.html.

“The WARC Format Explained.” Mixnode. Accessed November 20, 2017. https://www.mixnode.com/docs/reading-your-data/the-warc-format-explained.

Toyoda, M., and M. Kitsuregawa. “The History of Web Archiving.” *Proceedings of the IEEE* 100, no. Special Centennial Issue (May 2012): 1441–43. https://doi.org/10.1109/JPROC.2012.2189920.

“WARC, Web ARChive File Format.” Web page, August 31, 2009. https://www.loc.gov/preservation/digital/formats/fdd/fdd000236.shtml.

"What Is Alexa Internet? - Definition from WhatIs.com." WhatIs.com. Accessed November 17, 2017. http://whatis.techtarget.com/definition/Alexa-Internet.

Zimmermann, Kim Ann, Jesse Emspak, Live Science Contributors | June 27, and 2017 10:46am ET. "Internet History Timeline: ARPANET to the World Wide Web." Live Science. Accessed December 4, 2017. https://www.livescience.com/20727-internet-history.html.