Allie Whalen
Digital Preservation
Final Paper
December 12, 2014

Archiving Web-Based Advocacy and Activism: An Introduction to Web Archiving and A Case Study on Columbia University's Human Rights Web Archive

Advocacy is central to upholding the cause of human rights, and activist efforts are commonly manifested through print publications, blogs, and websites that function as avenues for free speech and impact. At the time when print publications existed as the sole medium for human rights documents, individuals and organizations depended on resources and physical distribution to expand awareness. The advent of digital has brought forth the internet as a core tool for human rights advocates to exchange dialogue, evidence, and information through a global open forum. As a result, organizations and individual proponents of human rights have developed an extensive web presence that continues to grow; whether the material is posted by a long-standing human rights organization or a blogger, the amount of content on human rights within the web is expanding and changing rapidly.

My paper will focus on archiving human rights websites and the unique characteristics and challenges of collecting controversial, graphic, and censored content. A case study on Columbia University's Human Rights Web Archive will serve as my primary resource to illustrate the types of human rights content on the web, the specific issues for archiving this content, and how such a web archiving program is implemented and managed. In addition, I will use Adrian Brown's book, *Archiving Websites: A Practical Guide for Information Management Professionals,* to outline the general development, methods, selection, cataloging, delivery, preservation, and operations of web archiving, as well as evaluate ongoing projects and initiatives and provide suggestions for archiving human rights websites. Specific sections of the web archiving outline go into greater depth than a general overview for the purpose of relating these elements to the case study at HRWA.

I. An Introduction to Web Archiving

The first website became publicly available on the world wide web in 1991 and so far in 2014 the number of active websites have exceeded 1 billion.[1] The internet's prevalence in global society, accelerating pace, and fast-changing content have engendered contemporary approaches to digital preservation and set archivists and librarians on alert to the imperativeness of collecting websites. The lifespan of a website is debated by internet analytics companies, web developers, and professionals in the field of digital preservation and web curation. Alexa Internet, an internet analytics company, measures that a web page is accessible for about 75 days before it is changed or deleted.[2] Web developers estimate that a website's design, structure, and function remain for 2 to 5 years before the site is redesigned or rebuilt.[3] Digital librarians and web archivists, such as Brewster Kahle, founder of the Internet Archive, gauge the average lifespan of a web page at 44 days.[4] These diverging estimations point to a core issue of preserving websites: how do we confirm that a webpage has 'died'? On the Library of Congress's Digital Preservation Blog, Mike Ashenfleder lays bare that, "Determining the average lifespan of a webpage is complicated not just by the infrastructure required to analyze a plausibly representative sample of links across the web but also because it's easy to conflate 'the average lifespan of a webpage' with other closely-

[1] "Total Number of Websites." *Internet Live Stats,* http://www.internetlivestats.com, (2014)
[2] "A Look at Website Lifespans." *Bismarck Tribune,* http://bismarcktribune.com, (2014)
[3] "A Look at Website Lifespans." *Bismarck Tribune*
[4] "A Look at Website Lifespans." *Bismarck Tribune*

related concepts that are, in actuality, much more difficult to measure."[5] Some of these concepts include non-resolving links, ephemeral issues, and loose estimates about when websites disappear, all of which will be evaluated later in this paper. These outlined challenges of determining the average lifespan of a webpage, along with other obstacles, contribute to the evolving understanding of web archiving and new approaches to digital preservation.

a. Development of Web Archiving and the Internet Archive

With these challenges in mind, web archiving was developed in an effort to ensure that historical output via the internet was preserved. Adrian Brown writes, "Web archiving has rapidly evolved to become an international, multidisciplinary concern, spawning a multitude of research and practically based programs."[6] This turn in traditional archival collection and directed digital initiatives began with the establishment of the Internet Archive in 1996 that remains the most notable and widely used web archiving service.[7] This section will briefly describe the Internet Archive as a primer for later in this paper when the implementation of I.A. will be evaluated through the case study on the Human Rights Web Archive. Similar web archiving programs include the Web Archiving Service (WAS) as well as the Internet Memory Foundation's private web archiving program, Archive The Net, both of which have their own benefits and limitations.[8] The purpose of I.A. is to "offer permanent access for researchers, historians, scholars, people with disabilities, and the general public to historical collections that exist in digital format."[9] With an agenda for universal access, I.A. successfully archives and makes available texts, audio, moving images, software, and archived web pages in its collection, which contains over 2 petabytes of data and over 150 billion web pages.[10][11] I.A. developed out of Alexa Internet, a web cataloging company, with the goal of providing permanent access through building a digital library.[12] In its former years, I.A. utilized Alexa Internet's proprietary web crawling software although direct use of Alexa's harvesting service was not available to the organizations that enlisted it.[13] Thus, I.A. developed a new open source Java application for web crawling and storing, Heritrix, to gain control over their own harvesting and develop new techniques.[14] I.A. also created the open source program, the Wayback Machine, as a means of access to the digital collections, complete with a time-based index, interface, and captured websites that can be browsed within their historical context.[15] The establishment of the Internet Archive, with the Heritix crawler and the Wayback Machine, set the precedent for software development and initiatives to come.

Web archiving programs have emerged on national, international, and institutional levels that all range in scale and approach. Some of these programs include the National Library of Sweden's Kutlrarw3 project (that contributed the open source NWA Toolset for access and delivery systems), HTTrack (an open source web crawler), NEDLIB (Networked European Deposit Library, a web crawler developed by the National Library of the Netherlands), DeepArc (a tool for capturing deep web content), MINERVA (Mapping the Internet Electronic Resources Virtual Archive, a web crawler developed by the Library of Congress), and more.[16] In addition, a

---

[5]"The Average Lifespan of a Webpage." *The Signal Digital Preservation,* http://blogs.loc.gov, (2014)
[6]*Brown, Adrian. Archiving Websites: A Practical Guide for Information Management Professionals. (London: Facet Pub., 2006),* 8.
[7]*Brown, Adrian. Archiving Websites,* 8.
[8]Perricci, Anna. "Interview with Web Archiving Coordinator at Columbia University Libraries," (2014).
[9]"Read More." *Internet Archive: About IA,* https://archive.org/about/, (2014).
[10]"The Code4Lib Journal – Archiving the Web: A Case Study from the University of Victoria." *The Code4Lib Journal Syndication,* http://journal.code4lib.org/articles/10015.
[11]"Read More." *Internet Archive*
[12]*Brown, Adrian. Archiving Websites,* 9.
[13]*Brown, Adrian. Archiving Websites,* 9.
[14]*Brown, Adrian. Archiving Websites,* 9.
[15]*Brown, Adrian. Archiving Websites,* 10.
[16]*Brown, Adrian. Archiving Websites,* 11-20.

multitude of groups have formed to develop tools and standards for web archiving as new programs surface, which will be mentioned later in this paper.[17] The International Preservation Consortium (IIPC) has been crucial to the collaboration of various international programs to develop tools, standards, and best practices of web archiving. Such collaborations have led to the development of the open-source web archiving toolkit that has aided in the creation of new methods and the interoperability of advanced technology.

b. Methods and Technology

Brown writes, "The web archiving process can be viewed as a workflow, whereby web resources are selected, collected, preserved, and delivered to users."[18] The various elements of web archiving are closely linked and often function in relation to one another, for example access may depend on the type of collection method employed. Technological resources factor largely into the decisions web archivists make for collection methods. Technology is the backbone of web archiving therefore understanding web servers, browsers, and crawlers, types of websites, harvesting settings, and limitations are crucial to the success of a web archiving program. A web server is a computer program that stores and retrieves content, such as HTML pages, and is identified using a uniform resource locator (URL). [19] A web browser is a computer program that requests content from the server in the form of HTTP (hypertext transfer protocol) requests to enable a user to view and navigate web content.[20] A web crawler is a computer program that automatically browses web pages and retrieves content.[21] Essentially, the server stores content, the browser renders content, and the crawler browses that content. The use of a website derives from a series of transactions between the server and the browser through a URL.

Web servers and browsers are especially important in terms of web archiving because these elements can decide the method in which a website is captured; Websites can be archived using the web server (server-side) that hosts the content or web browser (client-side) that retrieves the content.[22] This decision is usually made in consequence of the nature of the website being captured – if the website is static (or made up of multiple pre-existing web pages) then client-side collection methods will be likely be taken while dynamic websites (or web pages that are generated quickly from smaller, separate elements of content) will usually require server-side methods.[23] Further, types of website content also factor into methods of web harvesting an archive employs, with the exception of transactional archiving that is event rather than content-driven. These types include direct transfer, database archiving, transactional archiving, and remote harvesting.[24] Each of these collection methods have particular strengths and limitations, such as the high degree of authenticity derived from direct transfer or lack of capturing the original look and feel of a website through database archiving.

---

[17] "About IIPC." *International Internet Preservation Consortium (IIPC),* http://netpreserve.org/about-us.
[18] *Brown, Adrian. Archiving Websites,* 5.
[19] *Brown, Adrian. Archiving Websites,* xiv.
[20] *Brown, Adrian. Archiving Websites,* xiii.
[21] *Brown, Adrian. Archiving Websites,* xiii.
[22] *Brown, Adrian. Archiving Websites,* 44.
[23] *Brown, Adrian. Archiving Websites,* 45.
[24] *Brown, Adrian. Archiving Websites,* 46-61.

Types of Web Harvesting	Method	Content/Event Driven	Client-side/ Server-side	Strengths	Limitations
Direct Transfer	Data is copied directly from the source	Content	Server	Most authentic rendition	Resources needed to effect transfer, technological sustainability, website owner cooperation
Database Archiving	Data is standardized in order to archive it	Content	Server	Provides standard and easy to import collecting and database content, multiple supporting technology not required,	Limited technological support, issues defining timing and nature of collection, doesn't preserve the 'original look', website owner cooperation
Transactional Archiving	Transactions between the server and browser are collected	Event	Server	Collects evidence for how a website is used, can collect content regardless of the source	Doesn't collect content that has never been used, website owner cooperation, impacts server performance
Remote Harvesting	Data is captured from content hosted on remote servers	Content	Client	Easy to use, flexible, widely applicable, available software tools, simple infrastructure, archive in control	Backlog, issues configuring crawls, bandwidth, unable to collect dynamic content and binary objects

Table 1

c. Selection and Quality Assurance

Collection methods and technology are closely tied with the process of selection for web archiving, which specifies how the web archive will employ harvesting techniques in accordance to its collection scope. The selection process is broken down into steps: 1) defining a selection policy, 2) context 3) methods, 4) criteria, 5) boundaries, and 6) maintenance.[25] Factors that help to define these six steps include the mission of the collecting institution, intellectual property rights, and the institutional resources available to the web archiving project.[26] Challenges of the selection process include the structural, temporal, and informal qualities inherent to the web.[27] For example, one of the prominent characteristics of the web is its interconnectedness; the structure of hyperlinks present issues for web archiving where an archived web page is preserved but any linked web pages are not therefore resulting in page not found errors. In order to clearly explain the selection process, I have composed a graph that describes each section.

[25] *Brown, Adrian. Archiving Websites,* 24.
[26] *Brown, Adrian. Archiving Websites,* 24.
[27] *Brown, Adrian. Archiving Websites,* 25.

Steps	Considerations	Decisions			
Context	The organizational and global context, the defined needs of the organization & needs of the international web archiving community	What to collect			
Methods	The intended breadth and depth of the archive	What collecting method to use:	Approach	Strengths	Limitations
		Unselective	Collect everything possible	Captures interconnectedness of web, avoids subjectivity	Expensive, time consuming, possible legal issues
		Selective	Identify specific web resources	Higher granularity, manageable legal issues	Higher subjectivity
		Thematic	Semi-selective, captures every possible website that fits the designated theme	Catalog and access benefits for browse and search	Judging the rate of content change, subjectivity
Criteria	What web content relates to the scope of the collection	Content			
	If the crawler should be allowed to gather external resources in addition to designated websites	Extent			
	Lifecycle, existence, and possible evolution of a web resource, rate of content change, risk assessment for future technological obsolescence, significance of certain topics that may require higher frequency at specific times (during an event or crisis)	Timing & Frequency			
Boundaries	How far back crawler should go into the server's file directory	Define parameters for collection			
Maintenance	Monitor the status of websites confirm activity	Implement necessary changes to the selection			
Selection Policy	Consider the above factors	What best fits the organization and its needs			

Table 2

Three ways of carrying out quality assurance tests include pre and post-collection testing as well as issue tracking which helps to identify potential issues, ensure that harvesting is being enacted successfully through testing, and recording issues in a standard log to be resolved.[28] In order to explain these issues more concisely, I have composed a graph that illustrates types of issues, what they result from, how they affect the integrity of an archived website, and what preventative measures can be taken to ensure accurate content retrieval and capture.

[28] *Brown, Adrian. Archiving Websites,* 70-73.

Overall Type of Issue	Specific Issue	Cause	Long-term Effect	Prevention/ Resolution
Incorrect Navigation Functions/Missing Content	Incorrect or dead hyperlink	Content of a link may already have been removed or moved to another URL	No content or different content than intended was captured and archived	Links must be checked in advance before capture
	Incorrect or not captured hyperlink target	Hyperlink target URL is absolute and continues to link directly to the fully-qualified domain.	Absolute hyperlinks continue to link to the live version of the URL instead of the archived version.	Check to make sure the domains for all hyperlink intended for capture are on the collection list.
	Crawler blocked by robots.txt	Rule to restrict access to 'images/' folder	Captured website has missing images	Ignore the robots.txt and re-capture
	Inability to capture flash-driven menus	Web crawler can't capture dynamic or embedded hyperlinks	Linked content is not captured and navigation links can't function	Collect content manually through the live website
	Copyright Protection	Password protected webpages	Content can't be accessed or captured.	Configure web crawler to provide password or obtain content directly from website owner
Text-Only Versions Captured	Graphics of a webpage not captured	Websites offer graphical and text-only versions to websites	Style and additional content of a website are not captured and fracture authenticity	Change the web crawler to mimic a web browser to see the graphical interface
Multilingual Content	Web crawler captures pages in mixture of languages and the language toggle button does not function in archived version of the site	Multilingual content is provided through links that run through server-side script or cookies storing preferences	Archived version of website can't recreate server-side functions	Collect separate snapshots of each language webpage or configure crawler to follow or avoid language toggle button
Multimedia Content	Remote harvesting does not capture streaming audio or video	Dynamic content has restricted download, such as Flash	Crawler is unable to capture content	Obtain content directly from website owner
Crawler Traps	A set of web pages creates an infinite number of URLs (documents) for the crawler to find	A crawl could keep running and finding 'new' URLs that it had not encountered previously for an infinite amount of time	Uses up large amounts of document or data budget that could be used to archive valuable content	Run a test crawl and review the resulting URLs. Apply host restraints to modify the URLs that will be crawled

Table 3

d. Cataloging, Delivery, and Access

A web archivist's understanding of how a collection was harvested contributes to the potential granularity of a web archive's catalog and accessibility. Cataloging is necessary to manage and provide user access and many times web archives seek to integrate their collections into the larger catalog of an institution. Web archivists apply similar standards to cataloging as traditional libraries, archives, and museums, such as MARC21, although similar methods of cataloging aren't appropriate for maintaining web archives. For instance, descriptive metadata is best captured through manual creation, i.e. when an archivist inputs content information directly into the catalog record. Standard descriptive metadata can be captured through encoded HTML webpages – such as Dublin Core that includes information such as date, title, creator, etc. – although this metadata is not always reliable and is sometimes recycled and applied to multiple webpages.[29] Technical metadata can be generated automatically and retrieved through logs that

[29] *Brown, Adrian. Archiving Websites,* 79.

describe the time of capture and how the content was captured – such metadata can be further applied to long-term preservation.[30] The Internet Archive enacts a limited approach to metadata collection due to the high volumes of data the program gathers, capturing only the date of harvest and whatever metadata is automatically generated at the time of capture. Conversely, programs such as Pandora, that have a more selective collection scope, are able employ manual metadata entry for each captured website.[31] Generating metadata, particularly descriptive, for archived websites is a primary challenge for building catalog records since content on the web is both expansive and rapidly changing.

Delivery and access for users is reliant on the ability to search or browse for content. Robust catalog records aid in discovery, although, as previously stated, this is not always a viable option for web archives. Thus, tools such as full-text search engines across entire collections, browse options through hierarchical subject classification, external search engines, and relating archived content to external content, both live and archived. Managing the distinction between live and archived content is equally important for users and can be executed through standard messages that appear on every archived page or URL prefixes. Further factors in enhancing delivery include citation and assigning unique references and highlighting lost or disabled functions. Brown writes, "A delivery system must be designated to accommodate its user community's preferred methods of access, including appropriate mechanisms for searching, browsing, analyzing, reusing, and citing archived resources."[32] Access depends upon an organization's ability to accommodate for its user needs, which involves the organization's access to resources, available staff, and legal permissions to make archived content publicly available.

e. Legal Issues

Legal issues affect collection, preservation, and dissemination of archived websites through copyright, privacy, and content liability issues.[33] An overarching issue for web archives is the global context in which the internet functions. In other words, web archives face specific legal challenges because content is posted online from different countries around the world in various jurisdictions with their own legislation. Copyright affects a web archives ability to copy, alter, or disseminate archived web content.[34] Copyright is also a challenge for preservation since long term access often requires some sort of migration, such as transferring born-digital files to new formats. Approaches to mitigate copyright infringement include identifying intellectual property owners to obtain explicit permission, relying on legislative mandates for public records, or adopting the policy of retrospective take-downs; although, this is not always an easy task when it comes to the internet where content is posted freely and at times anonymously.

Privacy and liability issues also contribute to legal concerns over web archives. Archives usually hold large amounts of information pertaining to individuals within their collections however such collections may be permitted for historical research purposes. While web archives are advised to be aware of privacy rights, many websites already fall within the public domain and are less likely to have privacy rights issues as a result of capture. Liability issues, on the other hand, are of a greater concern to web archives that may capture content related to defamation, obscenity, or the promotion of illegal activity.[35] This is especially true for web archives such as the Human Rights Web Archive that collect controversial, censored, or illegal content. Careful selection and collection methods, quality assurance, limited access, take down policies, and enabling archival legislation are considered effective steps towards legal defense, but the

---

[30]*Brown, Adrian. Archiving Websites,* 79.
[31]*Brown, Adrian. Archiving Websites,* 79.
[32]*Brown, Adrian. Archiving Websites,* 144.
[33]*Brown, Adrian. Archiving Websites,* 144-55.
[34]*Brown, Adrian. Archiving Websites,* 148.
[35]*Brown, Adrian. Archiving Websites,* 152-5.

recommendation for web archivists to seek legal help if in doubt remains the most important.[36]

f. Preservation and Management

Legal issues, cataloging, quality assurance, and technology all factor into a web archive's main goal of continual access, which is the essential purpose of preservation.
In particular, digital formats have specific issues related to more general preservation concerns, such as technological obsolescence and storage. There are two possible types of preservation that an organization can take on including passive and active preservation. Passive preservation involves securing the storage of digital objects and preventing accidental damage or loss through redundancy and backup (such as the LOCKSS project, also known as Lots of Copies Keep Stuff Safe), disaster recovery, content management, and careful selection of storage mediums based on longevity, capacity, viability, etc.[37] Active preservation surpasses passive preservation by going a step beyond secure to storage to transferring formats to new technology to ensure continued access, software, hardware, and operating systems emulation as well as migration just before obsolescence, on demand, and through normalization (the standardization of file formats).[38]
Further, the Open Archival Systems Reference Model (OAIS) addresses such issues as preservation metadata, storage, ingest, management and access through its system for integrating digital preservation into the already established larger operations of an archive.[39]

Particular digital preservation concerns for web archiving stem from attempting to capture the hierarchy of interconnectedness inherent to the internet. A website is made up of numerous individual webpages, and each webpage contains various components such as HTML and images. These linked aspects of the internet present challenges for capturing and preserving the full extent of a website so that the archived version can serve as an authentic record. In his book, Brown lists two fundamental goals for preserving complex objects: "First, it must provide some means of describing the relationships between the components of the object and, second, it must ensure that any preservation actions that result in a change to one component are reflected by updating all the related components "[40] For example, hyperlinks are links that are encoded within a website therefore the underlying folder structure of the website needs to be preserved, along with the website, to capture relational metadata. Additionally, if a website that contains HTML pages is captured and later the file format of an image presented <img> within the HTML page is migrated to another format (GIF to PNG), then this process should be similarly carried out for all the other images that appear on the website in order to preserve these relationships. The techniques and implementation of digital preservation continues to evolve alongside advancements to digital technology thus one of the most important aspects of ensuring digital preservation is remaining aware and engaged with new developments and tools. Keeping up to date with changing technology and best practices for digital preservation and web archiving factor into the management and operations of a web archive. A web archive can be operated on an in-house, contracted-out, and consortium level depending on the resources and staffing available to the organization. Open source tools and software and at least one staff person with related experience is recommended for implementing a web archiving program and involves the development of new skills and collaboration between individuals at an organization.

II. Archiving Human Rights Websites

The task of archiving human rights websites holds particular characteristics beyond those of general web archiving; these aspects result from the quick changing, sometimes clandestine

---

[36]*Brown, Adrian. Archiving Websites,* 156-8.
[37]*Brown, Adrian. Archiving Websites,* 103-8.
[38]Brown, Adrian, *Archiving Websites,* 110.
[39]Lavoie, Brian F. *Technology Watch Report: The Open Archival Information System Reference Model: Introductory Guide*. Rep. Office of Research OCLC
[40]Brown, Adrian. *Archiving Websites,* 122.

nature of creating, sharing, and collecting human rights documents via the internet. Alongside the tools, resources, and expansion of access brought forth by the web, were new methods of restriction, censorship, and surveillance. Online censorship functions in a somewhat elusive form that is capable of penetrating the infinite tunnels of the internet and achieving stealth surveillance while censorship during the print period existed as a more of tangible force that involved the physical confiscation of materials. The methods in which governments implement censorship vary according to the state of politics, economy, and technology in a particular region. Tessa Fallon, the Web Collection Curator at CUL, writes in her article "Archiving Human Rights on the Web," that "Leaving aside practical challenges which exist for every website, in many places there is also the possibility of sabotage or attempts to remove a human rights-related website by opponents, religious, ideological, governmental or otherwise. Examples include denial-of-service attacks or in the most extreme case, the cutoff of all Internet service providers (the Internet "kill" switch)."[41] Government internet censorship occurs at different times, to different groups of people, and for different reasons, for instance during political change, military invasions, and crisis situations.[42] Activist groups, bloggers, and journalists are targeted for their dissemination of privately guarded information or evidence.[43] Online outlets for human rights carry sensitive material used to raise awareness and promote action, and these characteristics contribute to the complexity of preserving such web-based material.

The root challenge for archiving human rights websites results from the nature of the content that is essentially controversial and graphic photos, videos, audio, and text. For instance, human rights websites can contain content with evidence of gender-based violence, police brutality, political corruption, violations of children's rights, and much more. Organization-hosted or personal websites with highly sensitive images and material are more vulnerable to blocking, censorship, or direct attacks. For this reason, some sites are hosted anonymously or frequently move URLs in order to circumvent blocking, censorship, or persecution. The fast-changing aspect of human rights websites results from the necessity to keep out of reach of government control and monitoring these movements presents a primary concern for web archivists. In an interview with Alex Thurman, the Web Archivist at Columbia University's Human Rights Web Archive, he expressed that as the collection grows it becomes increasingly hard to keep track of which websites are still live.[44] HRWA currently captures over 600 websites at a range of frequencies that makes it difficult to monitor activity, especially with a limited staff. Other organizations that actively harvest human rights websites encounter this issue of tracking websites that are inherently covert and unstable due to their content. Another tracking and monitoring challenge for human rights web archives is establishing and maintaining contact with website owners, especially during volatile or impacting moments. Web archives – such as the Human Rights Web Archive and the Human Rights Documentation Initiative at the University of Texas (HRDI) – work in partnership or under the umbrella of other human rights organizations and archives, such as the Center for Human Rights Documentation and Research at Columbia University, the Bernard and Audre Rapoport Center for Human Rights and Justice at the University of Texas, Witness, and Human Rights Watch. Collaboration among organizations and individuals has resulted in tools, projects, and initiatives towards the progression and evolution of human rights documentation, archiving, and web archiving. In the following case study, I will evaluate how web archiving tools are being implemented to preserve human rights websites and provide recommendations for enhancing the web-based collections.

---

[41]"Archiving Human Rights on the Web - WITNESS Blog." *WITNESS Blog,* http://blog.witness.org/2012/01/archiving-human-rights-on-the-web.
[42]Hussain, Howard, and Argawal, *State Power*, 11.
[43]Hussain, Howard, and Argawal, *State Power*, 11.
[44]Thurman, Alex. "Interview with Web Archivist on Human Rights Web Archive," (12 Nov. 2014).

III. A Case Study on Columbia University's Human Rights Web Archive

The Human Rights Web Archive (HRWA) began gathering websites relating to human rights in 2008 and has so far collected over 600 sites, as previously stated. HRWA exists within Columbia University Libraries and was developed by the Center for Human Rights Documentation and Research, which supports multidisciplinary work in the field of human rights. HRDR remains the official repository for several organization-based human rights archives, such as Amnesty International, Human Rights Watch, and the Committee of Concerned Scientists.[45] HRWA is also a key component of the Web Resources Collections Program at Columbia, a collaborative initiative to incorporate web archiving into continuing collections development, and organizes community workshops and panels with the help of its web archiving coordinator. Implementation of HRWA began when librarians at Columbia University congregated to discuss what important resources research libraries were not capturing. From this meeting, librarians determined that while plans were set in place to manage commercial web resources, there was a gap in collecting for non-commercial resources, a specific urgency to attend to human rights websites at increasing risk of disappearing. A majority of small human rights organizations rely solely on the internet and either do not produce print materials or their materials are difficult for a library to obtain. This decline or lack of publications results from the limited resources available to certain countries, the fast-changing information and news of activism and war, and the advantage of an increased audience via the internet. HRWA's mission is to acquire content and information that is unable to otherwise be acquired. The archivists and specialists at CUL were previously aware that such content existed on the web through previously acquired print publications from various organizations. Following this decision to collect human rights websites, HRWA received funding through three Andrew W. Mellon Foundation Grants to implement the web archiving program and was able to hire two web collection curators to manage the program. HRWA uses the Archive-It service through the Internet Archive which incorporates the use of the open source web crawler, Heritrix, to carry out webpage captures. Archive-It has proved an effective tool for capturing and providing access, linking archived websites that reside at the Internet Archive to HRWA's own portal on CUL's website.[46]The web archivist, along with the help of other CUL personnel and selection guidance of HRDC subject specialists, has implemented an active, progressive, and impactful web archiving program that can be evaluated by using the sections outlined above.

a. Collection Methods and Selection

HRWA uses remote harvesting to gather web resources, relying primarily on client-side capturing to retrieve content from outside servers. Remote harvesting was selected based upon the nature of the websites being collected and the content-driven purpose of the web archiving program. For instance, the flexibility and user control given to web archivists through remote harvesting enables HRWA to change crawl boundaries and frequencies upon demand, which is especially important for documenting human rights websites that may require higher capture rates and extended crawling at influential moments. During the Arab Spring, HRWA was able to collect websites that reported on the protests happening across the Middle East and North Africa and capture the events and shared voices live as content was posted to the internet.[47] Increasing both the amount of websites and frequency at which they were crawled during the Arab Spring enabled HRWA to document the organizations'' sites and the power of the movement. [48]

However, remote harvesting has also limited the effectiveness in which HRWA can provide access to its archived collections. For instance, backlog accumulations have effectively

[45]"About" *About the Center for Human Rights Documentation & Research,* http://library.columbia.edu/locations/chrdr/about.html.
[46]"FAQ" *The Columbia University Human Rights Web Archive,* http://hrwa.cul.columbia.edu/faq.
[47]"Archiving Human Rights on the Web - WITNESS Blog.
[48]"Archiving Human Rights on the Web - WITNESS Blog.

inhibited catalogers from providing descriptive metadata for each archived web resource thus disabling full-text index searching within the collection. So, if an HRWA user wanted to search across various website content for a person, such as Nelson Mandela, the user would have to browse various websites through the provided subjects tab to find possible keywords that could relate to him, such as "Africans – Civil rights" or "Apartheid".[49]
Additionally, remote harvesting methods have difficulty capturing dynamic content and binary objects, such as streaming audio or video. A majority of human rights documentation that reaches the internet arrives in the form of audio or video that typically captures human rights violations and serves as strong evidence for global awareness and possible convictions. In 2011, HRWA received feedback from its users that out of the 369 crawled websites at least 77 sites published materials on YouTube or had their own YouTube channels.[50] Remote harvesting causes issues for capturing steaming audio and video that restricts the crawler from capturing Flash-based content, which accounts for much of the data uploaded to YouTube. Fallon writes, "While the live sites are currently available to everyone, what are the chances of being able to find the same particular pages, documents, and video in five years?"[51] The inherent susceptibility of web data to such factors as "link rot" and the specific vulnerability of human rights content to take-downs urge the importance of carrying out a comprehensive capture of each archived website so that such websites can serve as later references.

Managing web harvesting issues is aided through the process of selection, which helps to better define the context, scope, criteria, boundaries, and maintenance of a web archiving program. HRWA's selection process can be best illustrated using the selection process table outlined in section I. 10 subject specialists at CUL and Cornell University Libraries carry out the content selection process for HRWA; each specialist has expertise in a specific region and language that qualifies him/her to select websites for inclusion in the archive.[52] HRWA additionally accepts website nominations from researchers, student, scholars, and human rights advocates as well as the website owners themselves, and nomination forms for each type are available on the Archive's website. [53] Similarly, the Human Rights Documentation Initiative (HRDI) at the University of Texas has several activists, scholars, and organizations that help to identify web resources on human rights and welcomes suggestions and proposals for websites and projects.[54] Further, HRDI also captures content through Archive-It and exercises a subject-based thematic approach to its collection methods, although the subject headings of HRWA and HRDI differ which will be evaluated in the cataloging section of this paper.

---

[49]"HRWA" *The Columbia University Human Rights Web Archive,* http://hrwa.cul.columbia.edu.
[50]"Archiving Human Rights on the Web," WITNESS Blog.
[51]"Archiving Human Rights on the Web," WITNESS Blog.
[52]Thurman, Alex, "Interview with Web Archivist."
[53]"FAQ," *The Columbia University Human Rights Web Archive*.
[54]"About the HRDI," *University of Texas Libraries,* http://www.lib.utexas.edu/hrdi/about.

Steps	HRWA Considerations	HRWA Decisions			
Context	What is needed for human rights organizations? What types of human rights resources exist on the web? What is needed for the international community of human rights archives? The higher level organizational policies of Columbia University and the Human Rights Documentation Center	Archive web-based information on the interdisciplinary, wide-ranging, and networked field of human rights disseminated through publications, reports, media, and other content. Reasons for archiving include the risk of content disappearing rapidly and building documentary record of human rights advocacy and research			
Methods	Capture as much information as possible because more data will provide greater reference for the future but more data will also be harder to catalog. Collect websites that relate to the mission of CUL and the Center for Human Rights Documentation	What collecting method to use:	Approach	Strengths	Limitations
		Thematic	Capture sites based on relevance of the website to current research, teaching, and advocacy, perceived risk of a website disappearing, and likelihood that a website will not be archived or preserved by other means. Organizations whose paper archives are held at Columbia are another priority for web archiving	Opportunities for web curation, easy to browse	Specialist subjectivity, hard to search since more information is difficult for granular cataloging
Criteria	What types of web content relate to human rights?	Acquire content related to human rights that cannot be acquired otherwise (i.e. print), acquire web content for all organizations that the Center for Human Rights Documentation currently archived in other formats, and websites created by non-governmental organizations, national human rights institutions, tribunals and individuals			
	The importance of links to human rights content on the internet and how information is publicly disseminated (news feeds, dialogue, rallies)	HRWA captures external links that are included on the selected websites. Depends on the nature of the website, set the crawler path to up to 5 or 6 directories up within the directory structure (using Drupal software) but be careful of false directories.			
	Prepared collection guidelines, relevance of the website to current research, teaching, and advocacy , nature of individual websites, high rate of content change, risk of take-down	Regularly scheduled intervals. Apply frequency depending on the nature of the website (semi-annual or quarterly for less updated sites to monthly for more updated sites; possibly daily if crawls are ad-hoc in response to current events)			
Boundaries	Not able to capture all human rights web content but want to crawl as much data as possible, should align with CUL's existing collections, HRWA's Archive-It data budget	Define scope so that collected sites coincide with CUL's collection rather than extending captures beyond designated sites (less permissions and technical harvesting issues, less data usage)			
Maintenance	Human rights websites have high rate of changing URLs often or having content removed	Run frequent test crawls to check for content, robot.txt files, or issues capturing binary objects.			
Selection Policy	HRWA operates under CUL, subject-based thematic collecting methods, human rights related content and the nature of human rights websites	Web collections should aim to fit within the larger archival collections within CUL. Content should be related to human rights and human rights organizations and should be regularly monitored on behalf of the sensitive nature of the site content.			

Table 4

c. Cataloging and Delivery

As previously stated, HRWA uses Archive-It to capture websites; these archived websites are made publicly available on HRWA's website via the Internet Archive. Each website also contains a MARC21 catalog record that is available on CUL's online library catalog (CLIO) and OCLC's Worldcat database, including links to both archived and live versions of the sites. Metadata is automatically generated through Archive-It and is received within one hour of a crawl. Automated metadata work files include information such as file type (html, pdf, website), capture date, and domain. However, automated descriptive metadata is not possible with current web archiving software and data volumes exceed the manageable amount of content for manual entry. Limited metadata restricts the ability to search within catalog records, however, HRWA arranges content within subject headings to provide hierarchical browse terms to aid user discovery.

On the homepage of the HRWA portal, a user is able to see 6 possible browse options including featured websites, websites, URLs, subject, places, and languages. Each of these subject options can be organized alphabetically, in columns, and by count (with the exception of websites and URLs). The featured option includes a curated selection of websites based on a certain subject, such as websites related to civil society. The websites and URLs options provide full lists of each website. The subjects, places, and languages options include further selection choices, such as sub-subjects for child abuse, nationalism, and political prisoners, sub-places for Africa, Russia, and Iran, and sub-languages for Albanian, Spanish, and Arabic. Further, the results of each specific subject option can be narrowed down by geographic focus, language, where the organization is based, the organization type, and subject.[55] For each website, a new page appears with a link to the archived version and a MARC record including fields for creator, organization type, organization based in, geographic focus, subjects, summary, and languages. When a user clicks to view a capture, all past captures for the website are listed and accessible. Each capture displays a yellow header indicating that the user is viewing an archive web page through CUL and Archive-It, the date of capture, the related collection (in this case 'Human Rights'), and notification that the webpage may be out of date. Other web archives for human rights sites include similar yet different styles of user access. HRDI's archived web resources page lists all of the websites with their MARC record, which can be browsed by coverage, thesaurus term, language, and a keyword search option.[56] The Tamiment Library Web Archive includes several individual web archives collected based on subjects such as Anarchism, Feminism and Women's Movements, and Civil Rights and Human Rights.[57] Each archive includes home (with a brief description of the archive, search bar, and quick facts), about, site list, search, help, and contact. Site lists are provided based on the website's organization, such as AK Press and Lesbian Herstory Archives.[58] HRWA provides a separate portal for its web collection, while web resources captured by HRDI and Tamiment are available on the main website which doesn't allow as many options for information and browsing. However, because all three of the web archives use Archive-It similar MARC records are available indicating a general dependency upon automatically generated metadata within the web archiving sphere.

d. Legal Issues

Similar to basic cataloging functions for web archiving, are the legal issues of copyright, privacy, and liability. HRWA has implemented a precise permissions policy for their collections: 1) notify all organizations/website owners of HRWA's interest in capturing their website, 2) if the website owner does not reply, HRWA assumes it is permitted to capture the content and sends a follow-up notification indicating this assumption and the commencement of

[55]"HRWA" *The Columbia University Human Rights Web Archive*
[56]"Archived Web Resources," *University of Texas Libraries,* http://www.lib.utexas.edu/hrdi/hr_archive, (2014).
[57]"Web Archive" *NYU Libraries,* http://www.nyu.edu/library/bobst/research/tam/webarchive.html.
[58]"Web Archive" *NYU Libraries,*

capture, 3) HRWA will remove a site from the collection list upon the website owner's request.[59] From the 500-800 organizations that HRWA has contacted, about 40% of requests receive replies and about 55-60% of requests are not responded to.[60] Permissions are an important element of web archiving because they provide the ability to override robot.txt files that often restrict access to important content such as images and style sheets.[61] In an interview with Web Curator, Alex Thurman, he stated that permission requests can be challenging, for instance, acquiring permission from websites hosted in volatile locations such as Syria should be considered before the regular permissions policy is set in place because of the extended amount of time it may take to receive a response and the act of burdening these website owners with bureaucracy during hostile times.[62] Tamiment Library, that works with similar institutions takes another approach to permissions, which does not include any request to website owner's for permission to ensure timely captures and follows any take down requests from owners.[63] Cornell University follows the same policy as HRWA of two notifications to website owners, and HRWA keeps a record of all sent permission requests and replies in their FileMaker Pro database.

Copyright, in addition to permissions, is another legal issue that web archives deal with, and most archives that use the Archive-It service follow its 'opt-out' approach that "allows a website owner/content provider to remove access from the archive, and/or prevent their content from being captured by putting up a robots.txt exclusion on their website."[64] For its own policies, HRWA removes any material that violates copyright or other intellectual property rights and encourages contact by any individuals or organizations that suspect copyright infringement. [65] However, copyright exists as a larger issue since legislation differs between jurisdictions and create disconnects for what qualifies for copyright infringement. The varying notions of copyright around the world, in particular, affect the collection of human rights websites that come from many parts of the world and at times can be based in one region with information and content from another.

f. Preservation

Legal issues, along with the other aspects of web archiving previously stated in this paper, affect the preservation of archived websites. HRWA does not yet have a system in place for long-term preservation and continues to rely upon the Archive-It facilities for storing and preserving their captured data. Similar situations extend across web archives that have not yet developed a foundation to enact their own preservation plans. However, HRWA, along with other web archives, is involved in numerous initiatives and projects to support the growth of web archives, collaboration, and the enhancement of digital preservation.

III. The Growing Community of Web Archives and Recommendations

Some prominent projects and efforts to support web archiving include HURIDOCS (an organization dedicated to developing information tools and techniques to human rights organizations), Memento (an effort to provide cross-institutional search capabilities for archived websites), and the International Internet Preservation Consortium (IIPC) (an organization to build tools, standards, and best practices for web archiving and promote international collaboration). Several working groups exist to foster collaboration, such as Researcher Requirements Group, Access Tools Group, Metrics and Testbed Group, Deep Web Group, Content Management Group. Such collaborations have been successfully in developing tools, such as Heritrix,

---

[59]"FAQ" *The Columbia University Human Rights Web Archive*
[60]Thurman, Alex. "Interview with Web Archivist
[61]Thurman, Alex. "Interview with Web Archivist
[62]Thurman, Alex. "Interview with Web Archivist"
[63]Greenhouse, Nicole. "Interview with Web Archvist at Tamiment Library," (2014).
[64]"The Code4Lib Journal: Archiving the Web"
[65]"FAQ" *The Columbia University Human Rights Web Archive*

DeepArc, WARC (Web Archiving File Format), Wget, HTTrack, NutchWAX (Nutch with Web Archive Extensions), Xinq (XML Inquire), and more. Continued collaboration is required to advance the development of tools, software, and standards as well as spur new actions. Additional books, such as Adrian Brown's *Archiving Websites: A Practical Guide for Information Management Professionals,* can be used to increase awareness about web archiving, outreach, and problem solving. Extended practical guidance, trainings, trainings, workshops, and information for new web archiving projects will simultaneously promote collaboration and engender new web archiving efforts.

Web/Bibliography

"About" *About the Center for Human Rights Documentation & Research*. N.p., n.d. Web. 12 Dec. 2014. <http://library.columbia.edu/locations/chrdr/about.html>.

"About the HRDI." *University of Texas Libraries*. N.p., n.d. Web. 12 Dec. 2014. <http://www.lib.utexas.edu/hrdi/about>.

"*About IIPC*." *International Internet Preservation Consortium (IIPC)*. *N.p., n.d. Web. 12 Dec. 2014. <http://netpreserve.org/about-us>*

"*A Look at Website Lifespans*." *Bismarck Tribune*. *N.p., 27 Jan. 2014. Web. 12 Dec. 2014. <http://bismarcktribune.com/news/columnists/keith-darnay/a-look-at-website-lifespans/article_1d879ae6-851a-11e3-8bd1-0019bb2963f4.html>*.

"Archiving Human Rights on the Web - WITNESS Blog." *WITNESS Blog*. N.p., n.d. Web. 12 Dec. 2014. <http://blog.witness.org/2012/01/archiving-human-rights-on-the-web/>.

"*Archived Web Resources*." *University of Texas Libraries*. *N.p., n.d. Web. 12 Dec. 2014. <http://www.lib.utexas.edu/hrdi/hr_archive>*.

Brown, Adrian. Archiving Websites: A Practical Guide for Information Management Professionals. London: Facet Pub., 2006. Print.

"FAQ" *The Columbia University Human Rights Web Archive*. N.p., n.d. Web. 12 Dec. 2014. <http://hrwa.cul.columbia.edu/faq>.

Greenhouse, Nicole. "Interview with Web Archivist at Tamiment Library." Personal interview. 14 Oct. 2014.

"*HRWA*" *The Columbia University Human Rights Web Archive*. *N.p., n.d. Web. 12 Dec. 2014. <http://hrwa.cul.columbia.edu/>*.

Lavoie, Brian F. *Technology Watch Report: The Open Archival Information System Reference Model: Introductory Guide*. Rep. Office of Research OCLC, n.d. Web. <http://www.google.com/url?q=http%3A%2F%2Fwww.dpconline.org%2Fdocs%2Flavoie_OAIS.pdf&sa=D&sntz=1&usg=AFQjCNEWmAuYodEBX1wDQqhESvLNZdjdag>.

Perricci, Anna. "Interview with Web Archiving Coordinator at Columbia University Libraries." Telephone interview. 14 Nov. 2014.

"Read More." *Internet Archive: About IA*. N.p., n.d. Web. 12 Dec. 2014. <https://archive.org/about/>.

State Power 2.0: Authoritarian Entrenchment and Political Engagement Worldwide, Edited by Muzammil M. Hussain and Philip N.Howard, Ashgate, (2013).

"*The Average Lifespan of a Webpage*." *The Signal Digital Preservation*. *N.p., n.d. Web. 12 Dec. 2014. <http://blogs.loc.gov/digitalpreservation/2011/11/the-average-lifespan-of-a-webpage/>*.

"*The Code4Lib Journal – Archiving the Web: A Case Study from the University of Victoria*." *The Code4Lib Journal Syndication*. *N.p., n.d. Web. 12 Dec. 2014.* *<http://journal.code4lib.org/articles/10015>*.

Thurman, Alex. "Interview with Web Archivist on Human Rights Web Archive." Personal interview. 12 Nov. 2014.

Thurman, Alex, and Lily Pregill. "More Podcast, Less Process: The Web Archivists Are Present." Interview by Joshua Ranger and Jefferson Bailey. Audio blog post. Keeping Collections. N.p., n.d. Web. <http://keepingcollections.org/more-podcast-less-process/>.

"Total Number of Websites." - *Internet Live Stats*. N.p., n.d. Web. 12 Dec. 2014. <http://www.internetlivestats.com/total-number-of-websites/>.

"Web Archive" *NYU Libraries*. N.p., n.d. Web. 12 Dec. 2014. <http://www.nyu.edu/library/bobst/research/tam/webarchive.html.>.