An Evaluation of Web Archiving Programs in the U.S. Relevant to International and Area Studies

THE EXAMPLE OF LATIN AMERICA AND THE CARIBBEAN

By Jeffrey Garrett
Contents

Executive Summary ........................................................................................................................................anskjfadskfa dfasdfsdf asdf df as df a d f a d f as df
An Evaluation of Web Archiving Programs in the US Relevant to International and Area Studies: The Example of Latin America and the Caribbean ......................................................................................................................3
I. Introduction: Latin America Takes to the Web .........................................................................................5
II. Conducting Research in the Live and the Past Web of Latin America: A Hypothetical Example ....................................................................................................................................................................................11
III. Archiving the Latin American & Caribbean Web .................................................................................15
   A. Web Archiving at the Library of Congress .....................................................................................15
   C. Web Archiving at the University of Texas at Austin: LAGDA and HRDI ...............................28
IV. An Independent Use Analysis ..............................................................................................................34
V. Conclusions and Recommendations: Accelerating the Integration and Mainstreaming of Web Archiving .............................................................................................................................................................................38
   A. Metadata ...........................................................................................................................................40
   B. Finding Aids .....................................................................................................................................40
   C. Citation Standards & Tools ..............................................................................................................41
   D. Self-archiving .................................................................................................................................43
   E. Credibility Enhancement ................................................................................................................44
   F. Education, Outreach, Promotion ..................................................................................................46
   G. Data Mining & Whole-Collection Analysis .................................................................................46
   H. Promoting Collaboration ................................................................................................................49
VI. Cited Literature ....................................................................................................................................50
Executive Summary

In 2016 The Andrew W. Mellon Foundation awarded CRL funding to develop an “integrated, self-sustaining, international cooperative framework to support area and international studies (AIS).” The chief goal of the Global Collections Initiative is to expand electronic access to primary source documentation and data from major world regions, where the information landscape differs from that in the U.S. and Western Europe. The initial phase of the project focused on one region: Latin America and the Caribbean. A major focus of the initiative has been access to materials existing only in digital form. This report evaluates efforts in the U.S. to archive open web content from the Caribbean and Latin America for future use by researchers.

The web is booming throughout this world region and has become a fertile resource for research and publication across the disciplinary spectrum. At the same time, the ephemerality of web content, the result of deletion, migration, alteration, or adulteration, collectively known as “reference rot,” has led to a crisis in scholarly communication. Conducting, sharing, and reading web-based research is “like trying to stand on quicksand.” (Lepore 2015) While largely resolved for journals, this crisis “has so far not been adequately addressed for . . . web-at-large resources.” (Klein et al. 2014)

Part I of this report introduces both systemic and region-specific issues of web use (and abuse) which are at the root of the problem; while Part II, using the example of a hypothetical scholar researching the Landless Workers Movement in Brazil, shows how these issues impact actual research in ways unimaginable in the pre-web era.

In Part III, three prominent area studies–relevant archival programs are examined: the Library of Congress Web Archiving Program (LCWA); Columbia University’s Human Rights Web Archive (HRWA); and then two programs at the University of Texas at Austin, the Latin American Government Documents Archive (LAGDA) and the Human Rights Documentation Initiative (HRDI). For each, we describe: history; governance; scoping and selection; metadata and search; in-house use analysis; and self-assessed challenges and future hopes. In Part IV, we consider what external evidence exists that these archives are actually being used. We discover that “views” do not necessarily translate into scholarly citations, but that methodological problems make any comprehensive use analysis—recognized by the National Digital Stewardship Alliance as a high priority—difficult at this time. (Bailey et al. 2017, 27) Above all, lack of consensus on how and even whether to cite from web archives compromises any analysis of their research relevance.
Finally, in Part V, this report describes opportunities for moving ahead. These include efforts to standardize metadata across the library/archives divide; to develop better finding aids and expose them to web crawlers; to improve citation standards for web content—especially in popular style manuals and bibliographic software used by students, scholars, and publishers; to introduce certification standards for web archives, thereby enhancing their credibility among skeptical scholars; to push education, outreach, and exchange outside the bubble of web archivists both on campus and beyond, for example at discipline-specific professional meetings; to promote data mining and whole-collection analysis; and finally to advance inter-institutional—and international—collaboration, leveraging the strengths of multiple partners.
An Evaluation of Web Archiving Programs in the US Relevant to International and Area Studies: The Example of Latin America and the Caribbean

I. Introduction: Latin America Takes to the Web

After a slow start in the 1990s and early 2000s, internet use in Latin America and the Caribbean exploded a little over ten years ago, now exceeding that of the United States and Canada.\(^1\) As with so much having to do with very large numbers, a graphic can make this rapid growth dramatically clearer. The image on the left shows city-to-city internet connections between North America, Europe, Africa, and South America in 2007. It is immediately apparent that the bulk of international internet traffic was taking place between North America and Western Europe. The graphic on the right, created by the same researcher using the same mapping algorithm, reflects the density of internet connections just four years later, in 2011. Quite suddenly, South America and the Caribbean appear “on the map” as participants in world internet traffic, vastly outpacing Africa, though still lagging behind Asia’s even more explosive growth.\(^2\)

---

\(^1\) According to the monitoring site Internet World Stats, of 4.2 billion Internet users in the world (as of June 30, 2018—up 300 million since June 2017), 10.4% reside in Latin America and the Caribbean, while 8.2% are in North America. For comparison, almost half of the world’s Internet users—49.0%—are in Asia, with that percentage increasing. (Argaez 2018, Graham and Norsworthy 2019, 235)

\(^2\) Images created by Chris Harrison of the Human-Computer Interaction Institute at Carnegie Mellon University. As Harrison explains, the displayed data “only reflects density of connections, and not usage—hundreds of people may
Open web content, consisting of mainly human-intelligible commercial, political, cultural, and scholarly websites, additionally the accessible subset of social media use, blogs, discussion forums, and much else, makes up a large portion of internet traffic—in Latin America perhaps even more than in North America, since to our south “much digital publishing is not channeled or distributed through commercial publishers but is instead only taking place on the freely accessible web.” (Graham and Norsworthy 2019, 229) To an extent even greater than in other parts of the world, the web in Latin America and the Caribbean is rapidly becoming the primary venue for information generated by the news media, governments, NGOs, and cultural organizations—in other words, the type of information that has traditionally provided the basis for the historical record. There can therefore be no doubt that the capture, preservation, and exploitation of web content in Latin America and the Caribbean is of great importance to students and scholars of this world region—even while harvesting and harnessing this content for scholarly use is still in its infancy. And yet, as Harvard historian Jill Lepore put it in a widely-discussed article for The New Yorker in 2015, the web as a medium of memory is “elementally ethereal, ephemeral, unstable, and unreliable.” (Lepore 2015)

We therefore cannot wait until sometime in the future to preserve and archive web content—because it is disappearing before our very eyes. There is some debate about it, but generally the average lifespan of a web page worldwide is held to be between 44 and 120 days. (Guy 2009, Brügger 2012, 318) As Lepore says, basing research and scholarly discourse on information available on live websites today is “like trying to stand on quicksand.” She continues: “The footnote, a landmark in the history of civilization, took centuries to invent and to spread. It has taken mere years nearly to destroy.” (2015) It should be obvious that we need a memory apparatus that captures and preserves content when it happens rather than assuming it will be there months or years later.

There are reasons why preserving web content in Latin America has been slow getting off the ground. Some of these reasons are inherent to the web itself—and are familiar to most—while others are specific to Latin America and its history. On the systemic side, first, there is the ephemerality Lepore describes: Unlike traditional, physical libraries and archives, which are places of memory, on the web new content constantly overwrites the old, leaving not a trace of what had been there before.³ German media theorist Stefan Heidenreich has characterized the memory-averse web even more drastically than Lepore: “The web steals our history. We forget

³ Tim Berners-Lee, who gave us the Hypertext Transfer Protocol (HTTP), the core DNA of our modern-day Web, is regretful today that he neglected to build memory—a time axis—into his invention. As he confessed to Lepore in an interview: “I was trying to get it to go. Preservation was not a priority.” (Lepore 2015)
who we are. We become victims of the great chaos of the simultaneous.” (Heidenreich 2010)

Second, there is still a prejudice held by many contemporaries that most content on the web has no heft, that it is more akin to idle conversation than content that merits preservation. (Stirling et al. 2012) When the Library of Congress announced in December 2017 that it would no longer archive Twitter comprehensively, public reaction ranged from indifferent to sarcastically positive: New York Magazine, for example, ran an article under the headline “The Library of Congress Will No Longer Archive Your Garbage Tweets.” (Feldman 2017) Of course, this perceived lack of archival “worthiness” has plagued ephemeral formats for centuries—pamphlets, posters, playbills, newspapers—reflecting a disregard for ephemera and, conversely, a high regard for libraries and archives as Valhallas of enshrined/entombed knowledge rather than active memory institutions. It can all be reduced to the outdated view of the mission of historiography and the humanities and social sciences more generally as teleological, as “summing up” and awarding laurels (and apportioning blame) rather than “looking for heterogeneities, messiness, lines of influence, struggles, forces and conflicting interpretations.” (Michel Foucault as summarized in Brügger 2016, 1060)

On the systemic side, third and finally, there are the ethical concerns affecting the preservation of much currently relevant web content, especially social media: “It can seem improper to some to archive anything relating to this kind of individual activity.” (Stirling et al. 2012) These concerns become even more acute for web content in the human rights domain, where personal data regarding perpetrators, victims, and informants may become exposed to public view in a way print archives usually are not. The attendant moral, legal, and political issues now being raised are aggravated by the globally perceived sense that big American tech firms are predatory data gatherers, unconcerned with personal privacy and safety. Such concerns could easily scare off many smaller players both here and abroad whose efforts to collect web-based text and data are essential—but are becoming more complicated than ever.

Then there are several specifically Latin American issues standing in the way of web content preservation. One is the absence of a strong archival tradition, in large part a legacy of centuries of colonial rule. (Graham and Norsworthy 2019, 226f.) For the Spanish-speaking countries in the Western hemisphere, archives and other important records were maintained for centuries not in-country, but instead at the seat of colonial power in Spain. For possessions in the New World,

---

5 Or, as Neil Postman said of historians more pithily, “. . . they find autopsies easier to do than progress reports.” (Postman 2011, 5)
6 This could explain today’s “larger focus on preserving internal or institutional content and . . . concomitant decrease in the perceived mandate or ability to archive materials outside institutional purviews.” (Bailey et al. 2017, 29)
from the late 18th century forward, archives were consolidated in Seville, at the Archivo General de Indias. Perhaps this lack of inherited archival institutions contributes to the fact that to date, only a single Latin American country, Chile, has joined the International Internet Preservation Consortium (IIPC), which is dominated by European, North American, and East Asian members.

Also painfully relevant is Latin America’s history of autocratic and dictatorial rule—relevant because the culture of autocracy, for many reasons, is hostile to memory institutions such as archives and libraries. The fact that erasure of the past is so easy on the web is a gift to rulers seeking to eradicate the memory of their predecessors and, perhaps, of their own brutality in coming to and retaining power. When the Honduran military overthrew the elected government of Manuel Zelaya in June 2009, the entire content of its web presence—speeches, government plans and reports, and details about the administration’s achievements—was summarily deleted as well. (Norsworthy 2016, 17) Sometimes, erasure of websites is done openly, at other times covertly—since afterward, who can tell? David A. Bliss, digital processing archivist at the Human Rights Documentation Initiative (HRDI) at the University of Texas at Austin, shares a case in point, also from Honduras, of what may have been an act of cloaked, targeted memory erasure in a web context. The Comité de Familiares de Detenidos Desaparecidos en Honduras (Committee of Relatives of the Disappeared in Honduras), or COFADEH, represents the families of alleged political opponents of the regime who were “disappeared” during the dirty wars of the 1980 and early 90s. See Figure 3 on the following page: above left is the original COFADEH website; above right the same URL after it had been hijacked and overwritten with gibberish—ostensibly a random act by hackers in Indonesia. Bliss and his colleagues, however, strongly suspect that this was an intentional, targeted act. Fortunately, before vanishing from the live web, the websites of both the Zelaya government and COFADEH had been archived by separate projects of LLILAS Benson at the University of Texas at Austin and are accessible to researchers.

---


and human rights advocates in Honduras and everywhere in the world. (Norsworthy 2016, 17; and David A. Bliss, personal communication)\(^9\)

In Latin America as elsewhere, valuable websites can also be hijacked by commercial actors who do not play by the rules. For example, beginning sometime in the last few years and lasting until recently, visitors to cipamericas.org, the website of the CIP Americas Program based in Mexico City, found themselves redirected to a website offering cannabis derivative products. This affected the entire “/archives/” subdomain of CIP Americas as well as the web page of the program director, Laura Carlsen—showing that sites that advertise themselves as “archives,” a term that suggests they are safe and permanent, are for that reason alone not any less susceptible to attack.\(^10\) Older books citing frequently very valuable CIP Americas web content as well as books in press at the time of the hijacking now contain dead links—or unintended links to cannabis ads.

Faced with this problem, the parent organization, the Center for International Policy based in Washington, DC, decided that their original site, cipamericas.org, was too corrupted to be reclaimed, so they migrated to a new domain, americas.org, where their “archived” content can once again be found. (Center for International Policy 2018)\(^11\)

---

\(^9\) For more on website disappearances in Latin America—including the fate of the Zelaya government pages and the Zelaya administration’s afterlife as a website—see Trevor Owens’ interview with Kent Norsworthy at (Norsworthy 2012).

\(^10\) In Part V (“E. Credibility Enhancement”), we will consider efforts to attach greater credibility to the notion of a “web archive” through audits and certification.

\(^11\) Illustrating the problem is Rebecca Galemba’s *Contraband Corridor*, about trade and illegal trafficking along the Mexico-Guatemala border, whose links to CIP Americas content are now dead—just months after this book was published.
have solved one problem but has caused others, since until 2007 americas.org was the online home of the now defunct Resource Center of the Americas in Minneapolis and later its successor, La Conexión de las Américas. (Stinnett 2010, 72-73) Numerous URLs referencing online content created and posted by these two no longer existing organizations can still be found in published scholarship, but instead of pointing to content, they now just yield an “Error 404 Not Found” message. 12 Many similar tales of link rot and content drift can be found by tracing the history of crawls on the Internet Archive for any given domain or specific URL. 13

Even without bad actors involved, not having a curated and stable web archiving infrastructure to rely on can compromise sustainable access to primary source material relevant to Latin American and Caribbean studies. Research on Cuban literature, for example, cannot overlook that in the absence of a robust and open publishing industry, much new work is circulated primarily online, via blogs and webzines. As Northwestern Latin Americanist Emily A. Maguire describes it, these sources are often backed up, if at all, by individual readers on “flash drives, kindles, etc. (and even, sometimes, in hard copy) before, during, and after circulation online.” (Maguire 2012, 20 and personal communication) Recognizing how fragile the domestic distribution infrastructure is (and how spotty the coverage by the Internet Archive), individual scholars in the United States and some other countries, e.g., Argentina, have used their personal or their institution’s websites to store and share copies of Cuba’s literary production. 14 Yet how comprehensive is this coverage? How durable and protected from intrusion and rot? Will these sites be taken down or hidden away from crawlers when their individual faculty champion retires or relocates? Many supposed “archives,” especially those compiled and maintained by individual scholars and academic departments, are more akin to the scrapbooks, filing cabinets, and floppy disks scholars have traditionally used to save their primary sources, and are threatened by the same causes of rot which consume the sites they were created to preserve.

published (Galemba 2018, 265). Galemba’s bibliography in fact contains numerous references to other web-based articles where—probably just at the last minute—the note “link no longer available” had to be added. This story is told in greater detail—with illustrations—at (Garrett 2018b), and is by no means unique.

12 One example among many found by looking for occurrences of “americas.org” in Google Books: the link to “www.americas.org/item_17139” in chapter 3, note 40, of Alejandra Roncallo’s The Political Economy of Space in Latin America. (Roncallo 2014)

13 Go to https://web.archive.org/web/20150103072651/http:/www.americas.org/ to follow the tangled history of the domain americas.org through successive crawls.

14 One such scholar’s site documenting Cuban science fiction is that of physicist Daniel W. Coon at St. Lawrence University. http://it.stlawu.edu/~koon/cuba/CFCubana/CubanSF.html. This site does not appear to have been updated since 2009 and we do not know whether it is intended as a permanent archive. The library at the University of South Florida maintains an archive of the Cuban SF magazine Disparo en Red, which was active between 2004 and 2008, http://digital.lib.usf.edu/disparo. This appears to be a permanent archive—but do we know for sure?
Despite all this volatility, we should not single out the Latin American web as somehow unique for erasure, hijacking, and other forms of website abuse and manipulation—after all, link rot, not to mention attempts to eradicate or alter web content, take place in the United States as well, including websites controlled by the U.S. government. (Sinclair 2018) In fact, some forms of content loss encountered elsewhere in the world have not occurred in Latin America and the Caribbean, for example the loss of entire nation-state domains—called ccTLDs, or country-code top-level domains—when countries cease to exist, as happened in multiple cases in Eastern Europe during the 1990s. (Ben-David 2016)

Overall, we see that the web on its own cannot be an enduring home for information: it is by its very nature the opposite of an archive. An archive must instead be artificially created—like it has been for both paper and electronic scholarly journals and monographs by initiatives such as JSTOR, Portico, and others. As a 2014 study published by the Public Library of Science (PLOS) points out, “while solutions have been put in place to combat reference rot for references to journal articles, the problem has so far not been adequately addressed for references to web-at-large resources that are increasingly used in journal articles and that point to a wide range of web content, distinct from journal articles.” (Klein et al. 2014) This is therefore a crisis affecting (and afflicting) the preservation of open web content—and all the scholarship that cites it.

II. Conducting Research in the Live and the Past Web of Latin America: A Hypothetical Example

By way of introduction to the present and the future of web archiving for this region as exemplified by three large-scale archival programs—which we will be considering in the following section—let us work through a hypothetical case involving web archives, in this way also highlighting several important terms and problems. Let’s posit a scholar researching policies of the Dilma Rousseff presidency in Brazil affecting the Landless Workers Movement (MST), an important social movement in Brazil and elsewhere in the world for the implementation of agrarian land reform. This scholar begins by looking for ideas and prospective primary sources, some of which will be on the web. As we know is common from a number of recent studies of the practices of actual researchers (Rutner and Schonfeld 2012, 17-

15 “In assessing the ability to reconstruct any domain’s past, one should keep in mind that the use of the (archived) Web as a primary source can only date back to 1996, the year of the establishment of the Internet Archive. In that sense, the Web does not remember the history of ccTLDs deleted prior to 1996, such as the Web of Czechoslovakia.” (Ben-David 2016, 1114)
18, Riley and Crookston 2015), she begins with broad scans in omnibus databases, such as Google, Google Scholar, Google Books, JSTOR, ProQuest Global Dissertations, and others. She finds an intriguing master’s thesis by Maria A. Chavez of the University of Kansas entitled “Não é apenas sobre nós: Food as a Mechanism to Address Social and Environmental Injustices in Mato Grosso, Brazil.” (Chavez 2015) There she finds references to a relevant policy document from 2014 bearing the title “Mais Mudanças, Mais Futuro” (“More change, more future”). The footnoted location of the original document, programadegoverno.dilma.com.br\textsuperscript{16}, no longer exists on the live web—an instance of link rot. What to do? She succeeds in finding the document on the live web, or at least a document bearing the same title, but should she cite this location? She decides for two (good) reasons not to: for one, she doesn’t know if the document is identical to the original or might not have been redacted in the interim, e.g. at

\begin{quote}


\end{quote}

\begin{figure}
\centering
\includegraphics[width=\textwidth]{figure4.png}
\caption{Reference to a policy document in the bibliography of an M.A. thesis on the Landless Workers Movement in Brazil. (Chavez 2015, 130)}
\end{figure}

some point after the election or after Dilma Roussef was impeached (content drift).\textsuperscript{17} Second, she knows that there is no guarantee that a link to the live web will work over time to aid future researchers in reconstructing, reviewing, confirming, or challenging her findings. Our researcher then goes to the Library of Congress Web Archives, where she knows that there is a rich and publicly available archive for the 2010 presidential election in Brazil, but there appears to be no

\begin{footnotes}
\item[16] Chavez links to this original location, recording her access date as November 16, 2014: https://programadegoverno.dilma.com.br/wp-content/uploads/2014/07/Programa-de-Governo-Dilma-2014-RGB1.pdf (Chavez 2015, 130). See fig. 4.

\item[17] In fact, she cannot know for sure whether the document Chavez downloaded on November 16, 2014, was the same document as was on the site before the election over a month before.
\end{footnotes}
similar collection created for the 2014 Brazilian presidential election.\(^\text{18}\) Finally, however, she searches the Internet Archive using the original URL, and following one redirect, there it is: the page imbedding the policy document was crawled at 17:31:42 GMT on September 28, 2014—exactly one week before the Brazilian general election of October 5, 2014. She has her document, she believes, and it appears to have archival authority, and, based on the persistence policies of the IA, she can hope that it will also have a permanent location findable by later scholars without encountering a 404 Not Found message—or an ad for cannabis oil products—at the end of the trail. The only discomfiting fact is that the time-date-stamp of the linked-to PDF encoded in the archival URL is 15:26:48 GMT on October 15, 2014—ten days after the election—even though the crawl of the web page it is linked from is stamped September 28, 2014. So despite all her archival diligence, and the long path she has taken to obtain this version of the document, in the end our author still has no guarantee of her document’s authenticity. In the literature, this is often called a *time skew*.\(^\text{19}\) The problem derives from the fact that an archived web page is not really a “snapshot” of a web page at all, at least not in the original photographic meaning of the word—as it is nonetheless frequently called among web archivists—but a “mixed display” (Ankerson 2015, 39-40): a composite reconstruction using crawls of different elements of a live web page undertaken at different times. As Gordon Mohr, former chief technologist for web projects at the Internet Archive, explained in a blog post several years ago:

Even on a frequently and deeply collected site, all the different independent resources that make up one “page” may be collected hours, days, weeks, or months apart. The site could change design or ownership over that time, and so the page you see may be a mosaic of disparate elements. Less-frequently or less-deeply-collected sites, or sites that could not be fully collected due to technical limitations (such as crawler-blocking robots.txt), could be assembling resources together from years apart—the Wayback is always doing a “nearest” date match, but “nearest” could be years earlier or later. (Mohr 2010)

Since the Internet Archive does all the crawls for Archive-It partners, among them the Library of Congress, Columbia, and the University of Texas, time skew is endemic to a very large percentage of web archives in the United States and Canada. Time skew has non-trivial consequences for the research credibility of “archived” web pages. Susanne Belovari discusses

---

\(^{18}\) According to Abigail Grotke, web archiving lead at the Library of Congress, LC did indeed crawl and archive the 2014 election in Brazil: it just hasn’t been gathered as a collection yet, though this is planned for the near future.

\(^{19}\) Computer scientists, however, use “time skew” (or “clock skew”) to refer to something quite different, namely time lags, usually measured in milliseconds, between joined computers. See, for example, https://social.technet.microsoft.com/Forums/en-US/cacc38f4-9410-4691-91f8-7481f9bd7517/time-skew-problem?forum=winserverDS.
how it violates basic archival principles, destroying “original order and provenance, which have been at the very foundation of valid historical research and archival policies for centuries”:

Imagine studying the historical development of the State Department in 1877 by looking through its paper records and suddenly encountering records dating from 1932. (Belovari 2017, 75)

This is a significant flaw in web archiving technology, at least from the perspective of researching historians. Perhaps this helps explain why—if the evidence gathered and presented in this report (s. Part IV) is indicative—so few researchers use and footnote the archival version, if they think to look for it and succeed in finding it, instead preferring to footnote a (once) live website that may no longer exist or whose content has changed (“drifted”) over time. Citing the original source may satisfy the scholarly requirement to document one’s sources, but in the new research environment of resources gathered on the open web, the practice often leaves the task to the reader to find—or not to find—the authentic, original content.

How to deal with time skew and other problems deriving from the artificial and composite nature of archived web pages is not a topic treated in two leading style guides consulted for this report: the *MLA Handbook* and the *Chicago Manual of Style.* They largely disregard the existence of an archived “Past Web” and the different type of web-archival competence a researcher needs to navigate and exploit this Past Web—which will arguably be essential for scholars to use—and to cite—in the future. One of the purposes of this report is in fact to identify the limitations of web archiving as currently practiced relevant to international and area studies, describing in the final section (V.) efforts underway in the web archiving community to address and resolve them. In addition to continuing to refine the technology, we will also recommend teaching what Megan Sapnar Ankerson has called “digital archive literacy,” an acquired skill which she says is “. . . essential for using the Wayback Machine for historical research.” (2015, 40)

---

20 *MLA Handbook*, 8th edition (2016): “While URLs define where online material is located, they have several disadvantages: they can’t be clicked on in print, they clutter the works-cited list, and they tend to become rapidly obsolete. Even an outdated URL can be useful, however, since it provides readers with information about where the work was once found. Moreover, in digital formats URLs may be clickable, connecting your reader directly to your sources. We therefore recommend the inclusion of URLs in the works-cited list, but if your instructor prefers that you not include them, follow his or her directions.” (2016, 48) Belovari reports that some researchers, feeling uncomfortable quoting from the live Web, actually take (real) snapshots of their computer screens, using screen capture or smartphones, creating “scrapbooks” of their online sources to pull out if challenged—a charmingly anachronistic touch in this digital world. (Belovari 2017, 71) This practice is explicitly endorsed by the new edition of *CMOS.* (2017, 749) For a discussion in greater detail of how style guides and manuals recommend citing the “Past Web,” see (Davis 2016).
III. Archiving the Latin American & Caribbean Web

We consider now three programs among those setting the trend in IAS web archiving, again with a focus on Latin America and the Caribbean. These are: The Library of Congress’s Web Archive (LCWA); Columbia University’s Human Rights Web Archive (HRWA); and the Latin American Government Documents Archive (LAGDA)—along with its close affiliate, the Human Rights Documentation Initiative (HRDI)—at the University of Texas at Austin. We will want to see how they are coping with the issues identified so far and with what success their archived content is entering the teaching, research, and publication mainstream.

A. Web Archiving at the Library of Congress

1. A Brief History

National libraries around the world have recognized since the 1980s that their depository responsibilities need to be extended to the digital realm, especially where materials historically submitted to them in print form for legal deposit are now only available as digital files. Starting in 1996, some countries, among them the UK, Australia, Sweden, and Denmark, passed legislation to allow or mandate the collection and preservation of their nation’s digital output. In the United States in the year 2000, Congress established the National Digital Information Infrastructure and Preservation Program (NDIIPP) to develop a national strategy to collect, preserve, and make available to the public significant digital content. At the same time, the Library of Congress established its first digital archiving program, called MINERVA (Mapping the INternet Electronic Resources Virtual Archive), today simply called The Library of Congress Web Archives (LCWA). MINERVA worked together with the Internet Archive to archive the 2000 presidential election and then, responding to the 9/11 national emergency, domestic and foreign websites chronicling world reaction to these events, acting quickly to harvest and preserve this content before it disappeared. Over 30,000 websites were captured at that time: the September 11, 2001 Web Archive is today the Library of Congress’s single most visited web archive collection.

21 Our description of web archiving activities at the Library of Congress is based on publications by LC Web archiving team lead Abigail Grotke (e.g., Grotke 2015, Grotke 2011); on a phone interview with Grotke and LC collection development analyst Michael Matos on December 21, 2017, and a meeting at the Library of Congress on January 24, 2018, which included Grotke, Matos, additionally Debra McKern, former director of the Library of Congress Rio de Janeiro Office, who has also served as RO (recommending officer) for Latin America, nominating seeds (i.e. high-level domains) for inclusion in LC’s web archives; and Grace Thomas, one of the Web archiving team’s digital collections specialists. To all of them, the author extends his sincere thanks for contributing to this report.

Today, LCWA still contracts with the Internet Archive for crawling services, but—unique among IA's hundreds of other partners—harvested content is not made available through Archive-It or the Wayback Machine, instead the archived content is loaded on Library of Congress servers and is available only through the Library of Congress’s portal. Moreover, much content is available only on the premises, analogous to the treatment of deposited print publications. Unlike many other countries, however, the Library of Congress has never been granted a legal mandate requiring publishing entities and individuals to deposit their digital output, and conversely, the Library of Congress is not legally required to archive websites. This has led to a complex system of permission requests, to be described below.

As of 2018, there are about 100 event and thematic collections administered by LCWA, with detailed information—though not necessarily public access—provided through the gateway at [www.loc.gov/webarchiving](http://www.loc.gov/webarchiving), with the actual web archives grouped by subdomain at webarchive.loc.gov. There are currently 51 foreign collections, of which only three have to do directly with Latin America and the Caribbean: the Brazil Cordel Literature Web Archive[^23] and the archives of the two Brazilian presidential elections of 2010 and 2014, the latter of which has not yet been officially posted (and hence was unknown to our hypothetical researcher in II., above.)

According to Grotke, citing LCWA’s most recent annual report, the total size of the archive is now 1.3 petabytes.

### 2. Governance and Selection

The Library of Congress makes its collection development policy statements public, but there is a separate supplemental policy governing web archiving across all subject areas.[^24] There we learn about the process of nominating a site, the role of Recommending Officers (ROs) in selecting seeds and curating sites, and that the usual practice at the Library of Congress “is not to acquire individual web sites one-by-one, but as part of a named subject, event, or theme-based collection.” (Library of Congress 2017, 1) Policies governing the selection and archiving of foreign sites (of importance to Latin Americanists and other area studies researchers) are given special attention in a set of “Supplementary Guidelines” to LC’s Collections Policy Statements as follows:

> Foreign web sites are collected on a highly selective basis. To avoid duplication of effort, recommenders of international sites should verify that the content is not already being archived and made publicly available by the host country. Exceptions to this policy can

[^23]: [https://www.loc.gov/collections/brazil-cordel-literature-web-archive/](https://www.loc.gov/collections/brazil-cordel-literature-web-archive/)
be made if there are concerns over the long-term accessibility of a foreign website. Proposals for new collections of foreign web sites should include a statement in the justification addressing how the foreign web sites are “of most immediate concern to the people of the United States.” (Library of Congress 2017, 3)

It is perhaps surprising to those unfamiliar with the Library of Congress, considering both its name and its popularly assumed mission and the fact that the Library collects a great deal of material on behalf of the Congressional Research Service, that neither Congress itself nor the executive branch (including the Department of State) exercise direct influence on the selection of seeds for LCWA. The “primary drivers” of selection, according to LC collection development analyst Michael Matos, are the 209 recommending officers (ROs) at the Library of Congress and at regional offices around the world. Just as international and area studies web archives do not constitute a separate, defined segment of LCWA, neither are ROs necessarily responsible for particular countries or world regions. There are discipline-specific ROs (Law, Folklife, and many others) as well as format-specific ROs (Serials). To both attract and vet proposals for new seeds, a group of eight ROs, with Matos as convener, was formed in 2017. In addition to the disciplines and formats just mentioned, this vetting group includes representation from the Congressional Research Service (CRS), thereby indirectly providing for some legislative branch input into the selection process. According to Matos, proposals are being actively encouraged since the Library

---

25 In fact, Congress exercises no direct influence on what the Library of Congress collects in any format—aside from general—usually statutory—mandates that define broad areas like Law, Broadcast, and American Folklife. (Bernard Reilly, personal communication)
has emphasized the growth of web archiving as part of its overall digital collecting plan. “We recognize that there is a lot on the internet that is within scope and not being actively archived by anyone, and we currently have the capacity to add additional projects without compromising our core web archiving efforts (Federal websites, elections, etc.).” (Matos, personal communication)

Once a project is identified and approved, additional scoping takes place which goes beyond single seeds. The Web Archiving team identifies additional URLs on other domains to also crawl if those URLs publish content by the same producer. For instance, if a site owner publishes content on third party sites, such as Twitter and other social media accounts, LCWA instructs the crawler to try to archive that as well when capturing the primary site that was nominated for archiving. These captures are not comprehensive.

On the topic of Twitter and social media, much has been made of the Library of Congress Twitter archive, first acquired in 2010 with tweets going back to 2006 and with the charge to include all public tweets going forward.26 As was widely publicized in late 2017, the collecting mandate is no longer comprehensive, reflecting many concerns on the part of the Library of Congress, among them the need to honor delete requests, but especially the complexity facing researchers who want to use the vast amounts of data that have been collected for a host of different and fascinating projects. Permission must be granted by both the Library of Congress and Twitter for any such use. For this reason, most researchers seeking to use Twitter data go directly to Twitter itself or to commercial services such as Dataminr and Gnip, licensed by Twitter, typically to mine feeds for certain topics, opinion research, trends, and other patterns. For more on data-mining in general, see Part V (G. Data Mining & Whole-Collection Analysis).

3. Support and Collaboration

Not having the sweeping digital depository mandate of national libraries in other countries, such as Britain, France, Denmark, among others, the Library of Congress relies on the willing cooperation of site owners, both domestically and abroad. The Library has a notification and permissions process based on the country of publication and the type or category of the nominated site, and two requests are addressed in email letters that are sent out to most site owners: one for notification or permission to crawl; and another for notification or permission to provide access outside the Library’s premises. Although on the LC web preservation page,27 site administrators are encouraged to show sensitivity to the need to archive by removing blockages

26 It should be noted that this project is organizationally entirely separate from LCWA.
27 For very helpful advice to institutions or individuals wanting to facilitate (or at least not hinder) preservation and copyright protection of their Web content, see “The Library of Congress Guide to Creating Preservable Websites.” (Library of Congress 2018)
such as robots.txt files from their pages, the Library of Congress (unlike the Internet Archive) mostly does not honor robots.txt or delete requests—as the attorneys for MSNBC host Joy Reid, to their consternation, discovered last year. (Nelson 2018a, b)

Matters are much more complicated with foreign partners. With them, crawling permissions are based on U.S. law (since the crawling activity is happening in the United States), but the access permissions are based on the laws of the country that the site is published in. Often explicit permissions are required—and this turns out to be a bigger challenge than almost any other. As Grotke says, for foreign sites “the biggest non-technical problem is locating valid contact information—and then getting someone to respond if we can even find them.” Internationally, LC regional offices are a critical asset in securing this information and obtaining permissions, leveraging contacts in the regions where they are located.

Seeking permission from foreign site owners involves two steps. The first is notification of LC’s desire to crawl and requests permission to do so—if that permission is required. The second is a request for permission to display archived content publicly, if permission is needed. On occasion, depending on the category and country, site owners are simply notified of both. If only notice is required, the process ends there and the archiving and provision of public access move forward. If permission is required for either archiving or access, the LC moves forward only if appropriate permissions are secured. If not, archiving or access or both may be cancelled.

Debra McKern, former head of the LC’s office in Rio de Janeiro, talked about the difficulty of this process in Brazil:

For the political party sites and the press sites, the site owner is not revealed, so you’re relegated to using the “contact us” link. The likelihood that our request would be culled from the thousands of emails received by those sites is 0%. There’s nothing to set us apart, to alert them that a reply is necessary. (McKern, personal communication)

And yet ultimately a remarkable level of coverage has been achieved for the Library of Congress’s archives of foreign elections and of the international response to 9/11.

4. Metadata and Search

The Library of Congress’s Web archives are created and stored in the Internet Archive ARC (legacy) and Web ARChive (WARC) container file formats. Individual archived web sites are

cataloged using the Metadata Object Description Schema (MODS). Preliminary keyword, title, and subject metadata are extracted from the archived web sites and in-house workflow tools to create preliminary MODS records. As resources allow, records are subsequently reviewed and enhanced by catalogers who assign controlled subjects from Library of Congress Subject Headings (LCSH) or Thesaurus of Graphic Materials (TGM). In some cases, minimal MODS records are created with the extracted data only: the hope in mid-2018 was to release a backlog of about 4,000 of these minimal records across 40+ collections in the near future.

In addition, a MARC record for each collection is available in the Library of Congress Online Catalog at loc.gov so that the collection can be found along with other library materials in the catalog. Metadata included in collection-level records in Library of Congress Online Catalog looks like this:

```
245 $a Collection title $h [electronic resource].
520 $a General description of the collection content and number of web sites and date range when web sites were captured
6XX $a Collection-level subject heading (usually several 6XX fields)
856 $a http://hdl.loc.gov/loc.natlib/collnatlib.12345678
```

For discovery purposes, it is of course unfortunate that the vast web archival holdings of the Library of Congress are not (easily) co-searchable with other sites crawled and archived by the Internet Archive; also, that so much material is not available for research access outside Library of Congress premises. However, policies governing in-house-only access to Library of Congress Web Archives content are no different than those affecting the LC’s equally vast physical collections—nor are they substantially different from those of other national libraries enjoying broad collecting and depository mandates.

---

29 This presentation of LCWA metadata and cataloging practice closely follows https://www.loc.gov/webarchiving/cataloging.html. Accessed March 17, 2018. See that page as well for detailed information on website-level MODS data.
30 As Grotke explains, this statement is nominally true—unless you know where to look. LCWA, Internet Archive, and others are “Memento compliant,” which means LC-archived URLs with Wayback Machine URLs are searchable along with others at the Memento website: http://timetravel.mementoweb.org/.
On the positive side, collection-level and website-level description at LCWA are provided to a high standard, greatly facilitating discovery. Google searches for “Brazilian presidential election 2010,” for example, return three links to the matching Library of Congress Web collection within the top ten results. More detailed discovery is also possible thanks to copious metadata which is then picked up by major search engines, as a search for information on the PSDB in Rio de Janeiro (=Partido da Social Democracia Brasileira–Rio de Janeiro) during the 2010 election demonstrates (see Fig. 6). Going to the detailed description of the archived web site at https://www.loc.gov/item/lcwa00097012/ reveals that (some) Twitter and YouTube content has been scoped in—but also that “access [is] restricted to onsite users.” This at least answers to complaints voiced by Belovari and others that web archives with access restrictions should provide detailed information on what a researcher might find onsite, allowing him or her to decide whether making the journey to the web archive’s physical location will be productive. (Belovari 2017, 70)

5. In-house Use Analysis and User Feedback

Use data at LC Web Archives is collected using the Adobe Marketing Cloud, which counts Total Visits, Unique Visitors, Page Views, Average Time Spent on Site, Bounce Rate (= the rate of visits to a page that did not include a click to another page, download, or other request of the server), and several other data types. The number of total visits in 2017 for the Library’s Web Archive interface at loc.gov/websites was 180,238, or roughly 500 a day. The average time spent on the site is 2.62 minutes and the average number of page views per visit is 1.73. These latter two numbers seem low, but in the absence of comparison data we will refrain from making judgments.

The Adobe Marketing Cloud analysis tool also returns a list of the most popular LC web archives. Of the top ten, the September 11, 2001 Web Archive is, as mentioned above, the most popular, followed closely by the United States Elections Web Archive. Several IAS–relevant sites are in the
top ten: the Philippine General Election 2010 Web Archive (#7); the Indian General Election 2009 Web Archive (#8); and the Crisis in Darfur 2006 Web Archive (#9). No Latin American or Caribbean collection ranks in the top ten in terms of number of views.

At this stage of web archive development, and in recognition of the fact that the value of archived content will continue to grow over time as live websites being archived disappear, the Library of Congress Web Archiving Team is primarily focused on building the archive rather than on performing downstream use analysis. This is not unusual across the country.31

According to Grace Thomas of the LC Web Archiving Team, the Library of Congress is beginning to experiment with creating data sets that will allow researchers to use the archives in new ways, such as the network analysis projects that researchers like Ian Milligan and Peter Webster have undertaken.32 Not much is known about specific use of archived content in the LC Reading Room or elsewhere in the country or world. As an independent investigation suggests (see IV., below), the LC web archives are only infrequently cited in published research.

6. Challenges, Future Plans, Blue Sky

As the 2016 NDSA Survey revealed, many research universities are expanding their web archiving operations, but they are focusing more on archiving internal, institutional content rather than on the much more complicated (and therefore expensive) archiving of external—and especially foreign—web pages, as important as this might be for international and area studies.33

Given the constant budget pressures on the Library of Congress as a government agency—all Library activities, including OVOP’s activities and budget have been scaled back repeatedly since 9/11—one might think that the focus on area and international studies-relevant web archiving might also be in retreat. Yet according to LC’s Matos, this is not likely to happen. As he says, the Library of Congress has always had a strong focus on foreign-language materials and on gathering information and knowledge from the world beyond our country’s borders, and this is not likely to change anytime soon. The Library of Congress will continue to collect expansively and looks forward to working with partners across the globe, especially through organizations such as the IIPC, the International Internet Preservation Consortium.

31 As concluded in the 2016 NDSA Survey: “Given the relative youth of many programs, as well as the fractional nature of staffing and other resource limitations, lack of knowledge of downstream use is perhaps not surprising. Of the 80 responses [to the question: “Do you have active researchers utilizing your web archives?”], 19% (15 [respondents] answered ‘Yes,’ 30% (24) answered ‘No,’ and 51% (41) answered ‘Don’t know.’” (Bailey et al. 2017, 27)

32 See the blog “Web Archives for Historians: Historians Who Use, Think About, and Work with Web Archives” at https://webarchivehistorians.org/. We will return to the data mining potential of web archives for new and largely quantitative analysis in the final section of this report.

33 See the passage quoted in note 6 above from (Bailey et al. 2017, 29).
B. Web Archiving at Columbia University Libraries: The Human Rights Web Archive

1. A Brief History

Columbia University Libraries began exploring web archiving in 2008 out of a recognition that freely available websites were an increasingly important but ephemeral research resource that university libraries were not actively collecting. Robert Wolven, then Associate University Librarian for Collections and Bibliographic Description, served as PI for a one-year exploratory grant followed by a three-year program implementation grant, both from The Andrew W. Mellon Foundation. By 2013, Columbia was funding its own Web Resources Collection Program, which now includes large thematic web collections in the areas of human rights, historic preservation and urban planning, and New York City religions, in addition to archiving the University’s institutional web domain.

Columbia’s first and still largest collection is the Human Rights Web Archive, a collecting focus inspired in part by a 2007 CRL-cosponsored conference on human rights records and documentation at Columbia. Organizationally, it is an initiative of the Columbia University Libraries’ Center for Human Rights Documentation and Research (CHRDR). It represents an effort to preserve and ensure access to freely available human rights resources created mainly by non-governmental organizations, national human rights institutions, and individuals. Its origins date back to 2007—around the time of a CRL-cosponsored conference on human rights records and documentation at Columbia—when a project team of CUL staff began using Delicious to identify and gather human-rights-relevant websites, prioritizing sites deemed to be at risk. (Wolven 2010) Over 500 sites were tagged, including a few dozen suggested by CUL area studies selectors. Experimental archiving of selected sites, testing (and shortly thereafter subscribing to) the Internet Archive’s Archive-It service, began in May 2008.

---

34 The following presentation of Columbia’s Human Rights Web Archive has been compiled based on an initial phone interview with project coordinator Pamela Graham on December 15, 2017; a list of questions with answers supplied by project staff in January 2018; then an onsite meeting on January 26, 2018 with Graham and additionally Alex Thurman, Web Resources Collection Coordinator; Samantha Abrams, Ivy Plus Web Collection Coordinator; Kate Harcourt, Director of Original and Special Materials Cataloging (OSMC) at Columbia University Libraries (CUL); and Stephen Davis, Director, Digital Library and Scholarly Technologies, CUL. The author expresses his gratitude to all of these individuals for the generous support they have given this project and more generally CRL’s Global Collections Initiative.

35 For a full description of the project, refer to https://hrwa.cul.columbia.edu/about.

Project work on HRWA transitioned to programmatic work in 2010. As of early 2018, the project had collected 15 terabytes of data and was conducting active harvests of about 700 seeds.

2. Governance and Selection

HRWA is not an administrative entity. It is the largest of four thematic web collections being built by CUL’s Web Resources Collection Program (WRCP). Administratively, the program is based in the Original and Special Materials Cataloging department (OSMC) within Columbia’s Collections Division. A high priority at Columbia is mainstreaming web archiving with all other collecting and archival activity being undertaken by the Libraries. This is reflected in HRWA’s thorough integration with other administrative units of CUL and the campus at large, starting with the fact that Pamela Graham is head of both CHRDR and HRWA, additionally serving as the director of the Libraries’ Area Studies/Global Resources Division. Her current title is Director of Humanities & Global Studies and Director of the CHRDR. The CHRDR in turn is connected to the Global Studies Division of the Libraries, but also draws upon staff and support from the Rare Book & Manuscript Library, which houses and provides access to traditional archival collections. These include such notable holdings as the (paper) archives of Human Rights Watch and Amnesty International USA. The Director of CHRDR also serves as a board member of the University’s Institute for the Study of Human Rights (ISHR), which has been the main program and academic initiative in alignment with CUL’s HRWA collecting. This results in frequent interaction, both formal and informal, between HRWA staff and faculty and students: an important source of useful intelligence about existing as well as potential new seeds for the web archive.

Integration of physical and digital resource selection and management is key to policy and organizational structure at CHRDR and therefore also at HRWA. The importance of conjoining traditional collection expertise with the selection and management of seeds at HRWA is made clear in the 2019 publication—already frequently quoted in this report—“Archiving the Latin American Web: A Call to Action,” co-authored by Pamela Graham with Kent Norsworthy of the University of Texas at Austin:

Knowledge of existing publishing streams (including print) forms a basis for understanding the broader cultural production landscape and traditional modes of dissemination; in turn we can identify publishing that sits outside the mainstream, i.e. gaps and silences that are not systematically documented. Examples include websites of

37 Until 2011, Graham was also the librarian for Latin American & Iberian Studies.
marginalized social groups or movements, or emerging writers or artists who only disseminate their work online. (Graham and Norsworthy 2019, 230)

In addition to input from subject specialists at Columbia University Libraries “with regional and language expertise” as well as from faculty and staff associated with ISHR and CHRDR, the Human Rights Web Archive has also solicited—with some success—seed nominations from “researchers, students, scholars and human rights advocates” regardless of affiliation. “We are especially interested in hearing from human rights organizations that want to nominate their own websites.” (Human Rights Web Archive 2012)

3. Support and Collaboration

As mentioned earlier, essential support for HRWA during the early years came from three successive grants from The Andrew W. Mellon Foundation to enable specific initiatives and innovations. There is also consortial support for web archiving overall: as of mid-2018, Ivy Plus Libraries funded one full-time staff member, the Web Resources Collection Librarian (currently Samantha Abrams), and one part-time (50%) bibliographic assistant. While administratively based at Columbia, funding for these positions is provided collaboratively by participating Ivy Plus institutions—at present 11 of the group’s 13 members: Brown, Chicago, Columbia, Cornell, Dartmouth, Duke, Harvard, Johns Hopkins, Penn, Princeton, and Yale. Both the Ivy Plus librarian and the HRWA coordinator participate in the Ivy Plus Web Advisory Group, which reports up to the Ivy Plus Collection Development Group.

The key to the success of the project, however, has been principally the “forward-looking stewardship” (Graham and Norsworthy 2019, 224) given from within CUL itself. Although many institutions rely almost entirely on the Internet Archive not only for crawling but also for search features and interoperability with other institutional collections, Columbia determined early on that to be both attractive and responsive to researchers, “what we really needed was a customizable local access point.” (Alex Thurman) So even before working with the Internet Archive, CUL’s Digital Library and Scholarly Technologies Division created HRWA’s architecture, including keyword indexing—which must be frequently refreshed—and full integration of metadata. Metadata creation, which was and still is a new field in the (equally new) web

38 Examples documenting the importance of expert knowledge of literary scenes and publishing patterns in scoping web archiving projects are to be found in several sections of this report, e.g. the discussion of Cuban science fiction publishing on page 10 above. (Cf. Maguire 2012)

39 No funding is coming from the Mellon Foundation at present, which is, of course, the way the Mellon likes to see programs mature.

40 Ivy Plus maintains its own presence on Archive-It, administered by the Ivy Plus librarian: collections (and sites contained within) can be nominated by anyone at a participating Ivy Plus library. Collections are maintained in Archive-It by the Web Resources Collection Librarian. For more information, visit https://archive-it.org/home/IvyPlus.

41 This support and interest was manifested at our January 26, 2018, meeting at Butler Library by the presence and active participation of two senior administrators.
archiving community, required the full support of the Original and Special Materials Cataloging Division, which it received. Of course, once begun, as with all local IT projects of scale, established IT and cataloging infrastructure must be maintained and updated—also locally. This is especially true for HRWA, the only Columbia web collection with a local discovery portal, as all others rely exclusively on the Archive-It access portal.

4. Metadata and Search

As observed already, Google-searchable finding aids along with solid metadata and authority control are crucial components of discovery in the digital research environment—a topic we will be returning to in Part V (B. Finding Aids). At Columbia, collection-level and website-level metadata are created in Archive-It (Dublin Core). MARC records are created in Columbia’s OPAC, CLIO, and uploaded to Worldcat. This cataloging is done by the Web Resources Collection Coordinator.

Search and data analysis tools are not part of HRWA, though Columbia can be said to have pioneered more robust searching of web archive data through their SOLR indexing of web archive files. But otherwise, there is no DIY toolbox attached to the archive. In the experience of project staff, experienced researchers prefer to get raw data and process it through their own software, rather than have analytical tools “baked in.” Still, for future consideration, HRWA staff believes that less skilled users might appreciate convenient access to some basic analytical functionality, e.g. through the provision of simple visualization tools or the Google Ngram Viewer. In fact, this might help novice users better understand the potential of web archives. (Graham)

5. In-house Use Analysis and User Feedback

Google Analytics data from the HRWA Archive-It account document the level of use of HRWA’s collection access pages and underlying archived material. Since GA tracking began in November 2014, there have been 13,095 sessions, with 49.7% of views from the United States and 50.3% of views from the rest of the world—quite an amazing indication of international impact and area studies relevance. HRWA content in Archive-It has been viewed during this time span by users in

---

**Figure 7:** Example of a website-level description of an Argentinian human rights organization in HRWA. The seed URL is highlighted. Since the actual archival URL is served at the Internet Archive, it is formatted as: https://wayback.archive-it.org/1068/*/http://
166 countries, with 20 countries having over 100 users each; after the United States, the countries with the highest use of the collection are India and the United Kingdom. The Internet Archive’s public collection page statistics for the copy of HRWA archived content added to the general Wayback Machine shows dramatically higher use: 4,482,392 views since 2011, or about 650,000 views per year or 1,850 per day. Views are not citations, of course, but these numbers still document the great attention being paid to HRWA content.

At the same time, impact on published research and scholarship has been difficult for Columbia’s web archiving staff to assess—doing so has also not been the highest priority. As with the Library of Congress Web Archives, questions of documented use in research or public policy publications tend to be put aside for the present in favor of creating and enhancing well-curated, technologically robust archives. It can be said that web archives are to web content what Portico is to commercial journal content: each becomes important in the future as access to current content breaks down—which of course is much more likely with the web than with commercial journals. From the standpoint of scholarly use, “the value of the archive may not be in current awareness and use but in its future potential to restore access when links have broken.” (Graham) As a measure of the importance—and the success—of this work at Columbia, HRWA today includes over 50 websites created by organizations no longer active on the live web, findable thanks to HRWA-created metadata exposed to search engines. This number will only increase over time.

6. Challenges, Future Plans, Blue Sky

There are, of course, still a host of technical challenges at Columbia, including necessary software changes to the locally developed search interface. And there are the legal issues, for example “a lack of clear standards and guidelines related to copyright, i.e. the right to capture sites, host archived sites, and provide for their reuse by researchers . . . ” (Graham and Norsworthy 2019, 229) Staffing needs will grow as the collection grows, recaptures of existing seeds need review, curators need more time to assess and systematically build the collection, and outreach to the research community—research support services, promotion through workshops, et al.—needs to be expanded to entice and enable researchers to use HRWA as a data source for large mining and analysis projects. To expand use, institutional commitment must continue at the high level it has been, building “branded access to web

42 Results of GCI’s own analysis of citations to HRWA content in published research suggest that scholarship is either passing it by or not acknowledging use, at least in any explicit form. Data and possible conclusions to be drawn from this part of our study will follow in IV.
collections to create greater researcher familiarity and use with the Past Web as a research resource.” (Alex Thurman)

Looking at the big picture nationally, there is a keen sense at Columbia that the “primary obstacles to expanding [web-archive-based] activities in libraries are less on the ‘technology’ side and more on the ‘cultural’ side.” (Graham and Norsworthy 2019, 224) For this reason, as we’ve noted above, HRWA is firmly anchored in and closely tied to the more traditional areas of collection development and technical services. Siloed web archives, by contrast, will only lead to reduced quality of curation, isolation from other library resources, and ultimately neglect—with a concomitant loss of research and teaching relevance. Future plans and hopes? Based on the foregoing, these should be clear:

I hope that this kind of collecting can be mainstreamed into the ways libraries identify, select, and preserve information of enduring research value, especially as born-digital collecting becomes more the norm than the exception. I hope that the HRWA can continue to grow in alignment with research needs and increasingly be the product of collaboration between research, advocacy, and library communities. I also hope that we’ll develop highly effective and functional ways of discovering and using the web archives—that some of the high hurdles that currently exist will be diminished or eliminated. (Graham, personal communication)

C. Web Archiving at the University of Texas at Austin: LAGDA and HRDI

1. A Brief History

The history of the two principal active web archiving projects at the University of Texas at Austin—the Latin American Government Documents Archive (LAGDA) and the Human Rights Documentation Initiative (HRDI)—is an integral part of overall library and archive growth at LLILAS Benson Latin American Studies and Collections, one of the world’s most important centers for the study of Latin American history, culture, politics, and society. “LLILAS” is an abbreviation standing for the Teresa Lozano Long Institute of Latin American Studies, an interdisciplinary program integrating more than 30 academic departments and some 160 faculty across the university. “Benson” refers to the Nettie Lee Benson Latin American Collection, one of the world’s premiere repositories of Latin American and U.S. Latina/o materials. The joint organization, formed by the fusion of these two organizations in 2011, comprises what today is

---

43 With thanks to the director of LLILAS Benson, Melissa Guy, as well as to David A. Bliss, AJ Johnson, and the now retired Kent Norsworthy of UT Libraries for providing important background for this section. Unless otherwise indicated, statements attributed to them were contained in personal communications.
called in shorthand simply “LLILAS Benson.” Physical collections number over a million volumes, to which are added a wealth of original manuscripts, photographs, and various media related to Mexico, Central and South America, the Caribbean and Latina/Latino presence in the United States. The creation of LLILAS Benson’s digital collections began in the early 1990s, with a Gopher server. It was the first such information service for Latin America on the internet. A website followed in 1994—also almost prehistoric in the history of the internet, and of course long before the merge of the two organizations. Surely the highest profile digital project of the early years was the engagement of the University of Texas on behalf of the international effort to preserve, digitize, and make accessible the Guatemalan National Police Historical Archive Project (Archivo Histórico de la Policía Nacional, or AHPN), which had been discovered in an abandoned Guatemala City barracks in 2005. The University of Texas Libraries launched the AHPN Digital Archive in December 2011.44

Born-digital archiving at LLILAS Benson did not, however, begin as an extension of document digitization, but out of sheer necessity. As part of its overall collecting mandate, Benson Library had systematically collected Latin American official government documents, including annual State of the Union reports, or Mensajes presidenciales, as well as annual reports that individual government ministries are required by law to produce. (Norsworthy 2016, 22) Beginning in the late 1990s, however, Latin American governments began releasing these documents only in digital form, creating, according to Norsworthy, a “collecting challenge.” Initially, the Benson just collected and organized links to these documents, not anticipating either “link rot” or “content drift,” when a new annual report, for example, replaced the old at the same address. This led to the establishment of LAGDA, started in 2003 with an investigation and planning grant from The Andrew W. Mellon Foundation, becoming operational when the Benson joined Archive-It in 2005.45

Today, LAGDA comprises over a million discrete documents/files from approximately 300 ministries and presidencies in 18 Latin American and Caribbean countries. In terms of preservation, a recent review showed that many thousands of documents and speeches which

44 https://ahpn.lib.utexas.edu/. “With its more than eighty million pages of documents, the AHPN represents the largest single repository of documents ever made available to human rights investigators.” Of these, around ten million pages were publicly accessible at the time of the launch, with more being added now constantly based on collaborative scanning efforts with other international partners in Guatemala. (Norsworthy 2016, 9, for more background, see also Center for Research Libraries 2012)
45 The successful proposal to The Andrew W. Mellon Foundation was submitted by the Center for Research Libraries, four U.S. universities (New York University, Cornell University, Stanford University, and the University of Texas at Austin), and the Internet Archive. Proposal, final report, and other documents related to the “Political Communications Web Archive Project” are archived as https://web.archive.org/web/20140723042928/http://www.crl.edu/archiving-preservation/digital-archives/past-projects/political-communications.
are available through LAGDA no longer exist on the live web, including virtually the entire web presence of the Honduran government under Manuel Zelaya mentioned earlier in this report.46

In addition to LAGDA, LLILAS Benson is also home to several other smaller web archiving projects, including the legacy web archiving projects of LANIC, the Latin American Network Information Center, which, though no longer being actively maintained or updated, remains a serviceable and valuable archive. Among the many components of LANIC is APPELA, the Archive of Political Parties and Elections in Latin America.47

The most significant of the University of Texas's other actively maintained web archiving endeavors is the Human Rights Documentation Initiative, or HRDI, which monitors, crawls, and archives the websites of human rights groups around the world.48 HRDI was founded through a grant from the Bridgeway Foundation in 2008, originally to preserve records documenting the genocide in Rwanda. (Wilmann 2010) Since then, its mandate has grown, especially regarding Latin America. “...HRDI focuses on fragile, ephemeral born-digital resources published on the web by human rights activists and organizations, as well as the extensive audiovisual records those groups create.” (Kelleher et al. 2010, 97) The HRDI has its own account at Archive-It which, as we will see shortly, is visited frequently.

2. Governance and Selection

The existence of the LLILAS Benson partnership means that teaching, research, and collecting take place in a shared institutional context. In theory, this allows new seeds to be nominated by program faculty or curators as old ones expire or otherwise disappear (aka “runaway hosts”). In fact, however, the original seeds list was compiled exclusively by librarians. It also appears that despite the post-custodial paradigm invoked—and now officially and monetarily supported by a new grant from The Andrew W. Mellon Foundation (see next section)—selection remains an on-campus process. HRDI will be archiving the websites of its partners in Brazil, Colombia, and Mexico as part of the new grant from the Mellon Foundation, and it can be hoped, in the spirit of post-custodialism, that these organizations will also be called upon to nominate seeds from human rights organizations they collaborate with.

46 In light of the recent election of radical populist Jair Bolsonaro as president of Brazil, LAGDA’s importance as an archive of vulnerable government publications may once again be highlighted. Yet government documents constitute only a small segment of potentially endangered websites in Brazil’s new political climate.
3. Support and Collaboration

Following the start-up support from The Andrew W. Mellon and Bridgeway foundations, sustaining support for LAGDA and HRDI has come from internal UT sources. As is the case with so many IAS-relevant web archiving projects nationwide, however, the Mellon Foundation continues to provide grant funding for innovation and new programming. Most recently, in spring 2017, the University of Texas at Austin received a grant of $700,000 from Mellon to fund a project through LLILAS Benson titled “Cultivating a Latin American Post-Custodial Archival Praxis.” The project focuses on building local capacity in Latin America to preserve vulnerable human rights documentation and making the resulting documents digitally accessible. (LLILAS Benson Latin American Studies and Collections 2017) This grant builds on earlier projects supporting the digitization of materials from Nicaragua, El Salvador, and Guatemala. The new grant, with a projected length of 33 months, supports similar post-custodial initiatives with partners in Brazil, Colombia, and Mexico, with an emphasis on documenting underrepresented communities. Although not specifically supporting web archiving infrastructure, the grant pays for the position of Digital Processing Archivist—which manages the catalog of seeds crawled by Archive-It—and two other positions.

Of course, LLILAS Benson did not originate the “post-custodial” approach toward partner organizations, but it has embraced this paradigm wholeheartedly, as Melissa Guy made clear in her presentation to the CRL Collections Forum in May 2018. As a policy, “post-custodial archiving” resides somewhere between “governance” and “collaboration,” reflecting a shift in archival theory overall as it relates to area studies. Kent Norsworthy summarizes this partnership-based approach as it applies to Latin America and the Caribbean, forming the basis of the LLILAS Benson collaboration philosophy:

The field of Latin American studies has been changing for some time, requiring an end to the previous paradigm—benevolent study of our “southern neighbors” from an unreflectively northern perspective—and replacing it with the principles of horizontal collaboration among sister institutions across the hemisphere and critical theoretical engagement from a true diversity of perspectives, including those rooted in the south.

(Norsworthy 2016, 3)

49 For example, Archive-It subscriptions are paid for from a UT Libraries digital projects fund source.
51 The terms “non-” and “post-custodial” are most clearly articulated in writings and speeches by Canadian archivist Terry Cook going back more than 30 years, e.g. “What Is Past Is Prologue.” There he writes: “The postcustodial paradigm is an overarching conceptual mindset for the archivist applicable whether the records are transferred to the custodial care of an archives or left for some time in a distributed or non-custodial arrangement with their creator.” (Cook 1997, 62-63)
52 “The ‘Post-Custodial’ Model for Preserving At-Risk Archives in Latin America.” (Guy 2018)
In the field of web archiving, this approach must also involve calling on in-country partners to provide seed nominations and ensuring that access to all born-digital archives is open to those partners, while at the same time protecting individuals in source countries from negative consequences of exposing their data and personal stories. It is therefore no surprise that the “National Forum on Ethics and Archiving the Web” held at the New Museum in New York on March 22–24, 2018, specifically called for contributions on “recognizing and dismantling digital colonialism and white supremacy in web archives,” as well as “strategies for protecting users: from one another, from surveillance, or from commercial interests.” (Documenting the Now Project 2017)

As a final question under the “support and collaboration” rubric, we were interested to learn whether there has ever been collaboration between the HRWA at Columbia and the HRDI at the University of Texas—and perhaps with thematically similar collections at Duke University, the University of Minnesota, and elsewhere. Columbia had told us there was none, and David Bliss at Texas reported only that “HRWA seeds were specifically avoided in the HRDI’s initial seed list.” Similarly, between LAGDA/HRDI and the Library of Congress’s web archiving programs, different collecting scopes take the place of active collaboration, resulting in a minimum of collecting overlap.

4. **Metadata and Search**

Even before the Internet Archive introduced keyword searching—superseding its prior sole reliance on known URLs—LAGDA and HRDI supported it. Cataloging was done consistent with Archive-It standards at the seed level. LAGDA metadata reflected cataloging done in the pre-internet era of international government documents collecting. The seed descriptions at HRDI now used online were created by the small team of subject librarians who also selected the initial seed set, edited by LLILAS Benson staff member T-Kay Sangwand.

5. **In-house Use Analysis and User Feedback**

Like Columbia—and consistent with results from across the United States, as recorded in the 2016 NDSA Survey—the focus of web archiving at LAGDA and HRDI at present is building the archive, i.e. the identification, collection, and preservation of vulnerable materials from Latin America and other world regions. Use analysis is writ small: Google Analytics data on LAGDA/HRDI web archives is not maintained, and there is no indication of student use of these resources, e.g. in student term papers or theses. There is a digital scholarship coordinator on staff, but this position does not distinguish, at this point at least, between commercially-sourced, digitized, and born-digital content when it comes to assisting faculty and students.
The use data page for LAGDA at the Internet Archive, accessed April 4, 2018, records 3,154,453 views since the creation date of August 3, 2011—on average about 470,000 views per year or 1,300 per day.\(^{53}\) The Internet Archive also posts data on views of the HRDI, also accessed on April 4, 2018: there have been 1,806,613 views since July 30, 2011, on average 270,000 views per year or 780 per day.\(^{54}\) As with our consideration of the other two programs, we will be reporting on evidence of the impact of this material on published scholarship in section IV.

6. Challenges, Future Plans, Blue Sky

According to LAGDA and HRDI staff members, the biggest challenges their work faces today are not on the technical side: they have to do instead with the availability of sufficient staff time to properly curate the 300 active seeds—and add new ones as existing seeds go dead. HRDI project archivist David Bliss mentions the demands on staff time placed by the (important) new Mellon grant, and by having to deal with significant staff changes in recent years—among them, of course, the retirement of Kent Norsworthy. “I don’t think we’ve actually added any seeds in the last two years to either of our Archive-It subscriptions,” Bliss says. “We’re hoping to solicit nominations from LLILAS Benson faculty and staff (and maybe the UT community more broadly, for HRDI seeds at least) soon . . .”

At LLILAS Benson, everyone’s hope is that the positions now supported by the Mellon Foundation can be retained as regular funded staff positions when the current project is completed—and that they and other existing staff can begin leveraging the latest “post-custodial” relationships to Latin American partners this project is developing as well as strengthening use analysis, research instruction, and promotion on campus—and, as mentioned earlier, to draw attention to important projects as seeds for future harvesting.

LLILAS Benson staff recognize the importance of strengthening relationships on campus and developing the feedback loop between researchers and web archivists to improve both the quality of LAGDA and HRDI and to encourage their more active use. Graham and Norsworthy touch on at least one important part of this challenge in their 2019 article:

\[
\ldots\text{anecdotal evidence suggests that researchers are creating their own personal archives of information saved, copied, or captured in some manner from the web. How can those scholar-led archiving efforts inform more systematic and comprehensive collection building carried out by libraries? (Graham and Norsworthy 2019, 232)}
\]


\(^{54}\) [https://archive.org/details/ArchiveIt-Collection-1475&tab=about](https://archive.org/details/ArchiveIt-Collection-1475&tab=about).
Ideally, then, such collaboration would make researcher “scrapbooking” largely unnecessary, obviating as well the need for readers of this research to go off in search of content referenced by URLs that no longer link.

IV. An Independent Use Analysis

Use analysis that goes beyond the anecdotal—in other words, empirical study that is based on harvested occurrences of unique and explicit archival URL stems in large databases of scholarly literature such as Google Scholar or ProQuest Global Dissertations—is problematic for several reasons. For one, researchers who cite websites may choose to cite original, live-web URLs even if they have used archived content. Live web URLs are after all shorter, meaning less “clutter [in] the works-cited list” (MLA Handbook, 2016, 48), and authors may assume (or just hope) that readers corroborating research findings will go to the Internet Archive or other known archives if they can't locate/relocate the original content—or worry (as they should) that it might have been altered in some way. Contributing to the reliance on the live web may be that standard style guides (Chicago, MLA, APA, others) are either silent or equivocating on how or even whether to reference archived web content (Davis 2016), reflecting a lack of awareness of the existence of these archives or a lacking community consensus on how to treat them in scholarly work. The most recent (17th) edition of the Chicago Manual of Style, released in fall 2017, even explicitly recommends that authors leave it to readers to look for archived pages if the live web page is gone or appears to have been altered:

If a site ceases to exist before publication, or if the information cited is modified or deleted, this information should be included in the text or note. . . . Such dates, together with the [original?] URL, give interested readers a chance to find the information through the Internet Archive or other means.” (University of Chicago Press 2017, 14.207)

In other words, there is no recommendation that authors have a responsibility to find and cite an authoritative archival location for a “modified or deleted” source page and provide this information for the reader’s convenience. And this, of course, makes it difficult to gauge or even detect use of web archives in published scholarship. (Garrett 2018a, 4)

A further complication stems from the fact that most Archive-It content is also loaded into the public Internet Archive (archive.org) and these URLs are not explicitly traceable to the collecting institutional Archive-It participant. For example, https://wayback.archive-it.org/1068/20120602210549/http://acda.co/, identifiable as part of Columbia’s Human Rights Web Archive, becomes the Wayback Machine’s https://web.archive.org/web/20120602210549/http://acda.co. The new URL deletes the signature “1068” of Columbia’s HRWA as the contributor (“collector”) of the archived content. Attribution/provenance is part of the metadata, but there is currently no way to search large
collections of web.archive.org URLs by collecting institution.\textsuperscript{55} This obscures the provenance of source materials and is inherently problematic for researchers—and of course also for meaningful use analysis.\textsuperscript{56}

There are some additional problems, e.g. the need to search separately for URLs conforming to the encrypted HTTPS protocol, i.e. beginning with https://; and also that URLs cited in Google Books are the result of often very dirty OCR, not born-digital text as with Google Scholar. Not only are transcription errors more likely, but line-ending forward slashes ("/") and other special characters scan as spaces, meaning that a URL such as http://webarchive.loc.gov/lcwa0005/... will not be retrieved by a search for this URL stem: it comes up instead (and only) in the broader search for http://webarchive.loc.gov (see fig. 8).

With all these caveats in mind, it is still useful to see how often and in what contexts explicit references to web archives are found in published scholarship. We therefore conducted searches for the following URL roots in Google Scholar, Google Books, ProQuest Global Dissertations, JSTOR, additionally in several recent scholarly e-books from major academic presses:\textsuperscript{57}

- \url{http://webarchive.loc.gov} for Library of Congress Web Archives
- \url{http://wayback.archive-it.org/1068} for the Human Rights Web Archive at Columbia University
- \url{http://wayback.archive-it.org/176} for the LAGDA and \url{http://wayback.archive-it.org/1475} for the HRDI at the University of Texas Austin

Let us turn now to the actual data.

\begin{notes}
\item[4] For Acevedo-Vila's testimony, see Hearing before the House Committee on Natural Resources, H.R. 856: A Bill to Provide a Process Leading to Full Self-Government for Puerto Rico, 105th
\end{notes}

\footnotesize
\textsuperscript{55} Lori Donovan of the Internet Archive, personal communication, April 2018.
\textsuperscript{56} There has been some early discussion with Internet Archive staff to explore possible workarounds, given how useful this kind of use data would be for Archive-It partners, but this has so far not led to any advances. So at present, for researchers accessing archived content through the Wayback Machine rather than through partner portals, there is no way to get use data traceable to specific Archive-It subscribers.
\textsuperscript{57} Three titles from the Brookings Institution Press (Brands and Suri 2016, Ghanem 2016, Mares and Trinkunas 2016) and one from Stanford University Press (Galemba 2018)
A search for occurrences of the highest level root domain for LC web archiving, 

http://webarchive.loc.gov in full-text, notes, and bibliographies across the Google Scholar database retrieves only 29 citations, mostly to articles about web archiving, not to specific archived content. The same search for www.loc.gov/webarchiving retrieves 48 citations—also not primarily to archived content, but instead to “meta” discussions on the topic of web archiving. Discussions about web archiving certainly have their value, but they tend not to treat the archived content as a primary source.

HRWA and LAGDA: Of about 619 citations pointing to Archive-It content found in Google Scholar, only 1 includes the subdomain for HRWA (/1068/), namely a paper entitled “Visualizing Digital Collections at Archive-It” presented at a 2012 conference—in other words, another “meta”-reference rather than one to content. (Weigle and Nelson 2012) Similarly, filtering for citations to the subdomain for LAGDA (/176/), we find only one, contained (and properly cited) in a student term paper accepted as a senior honors thesis at the University of Maryland. (Shaker 2013). Interestingly, in her acknowledgements, the author thanks a librarian for introducing her to LAGDA:

... UMD librarian Pat Herron was very helpful in this process. Her demonstrations in our . . . library session introduced us to all the relevant databases, which would have materials from research to newspaper articles to how to locate an archive full of government documents from all around Latin America. (Shaker 2014)

---

58 We searched both http://wayback.archive-it.org/ and https://wayback.archive-it.org/ as the URL stem, adding the HRWA subdomain . . . 1068/ to count occurrences of citations to HRWA content.
On the one hand, of course, it’s encouraging to read that bibliographic instruction does lead students to valuable archived web resources they might otherwise overlook. On the other, however, we must ask why we could find no other citations in Google Scholar, this vast database of published literature, to LAGDA material. If we search for citations to HRDI content, we also come up with only a single reference, namely to a recent technical article on developments in web archiving (Fig. 10)—not to actual human-rights-related content. (Fernando, Marenzi, and Nejdl 2018)

Conducting the same searches in Google Books is more problematic because of the systemic scanning inaccuracies already mentioned, but also more productive in terms of numbers. Searching for http://webarchive.loc.gov retrieves 527 citations, of which, however, very few have to do with international and area studies, among them the citation reproduced above (fig. 8) from an actual Google Books results screen (Wasniewski et al. 2014, 623). Just for comparison, http://www.loc.gov/item/ occurs 1,390,000 times in Google Books—this includes, of course, references to hundreds of thousands of photographs archived at the Library of Congress. The URL http[s]://www.loc.gov/item/lcwa occurs many orders of magnitude less frequently: just 21 times.

Also in Google Books, of about 17,450 citations to the URL stem http://wayback.archive-it.org/ or https://wayback.archive-it.org/, only nine refer explicitly to HRWA content. As for LAGDA and HRDI URLs in Google Books, there are a few scattered citations to LAGDA content, such as to a Mexican presidential address of the year 2000, archived by LAGDA in 2006, displayed here (Fig. 11). Recorded citations of HRDI content are no more plentiful.

**Figure 10:** This is the only reference to HRDI content in the Google Scholar database. Highlighting based on a search of the Web version of this article.

**Figure 11:** A link to LAGDA content in a recent monograph (Williams 2012, 140), retrieved by searching for instances of http://wayback.archive-it.org/176/ in Google Books.
The same searches in ProQuest Global Dissertations, which might be expected to give some insight into the research practices of younger scholars, were no more productive than these. The same applies to JSTOR, whereby its embargos on more recent content surely do not help capture current citation trends. Finally, searches in several representative, internet-aware monographs published since 2016 produced no archival URLs.

Despite all the methodological caveats given at the beginning of this section, the evidence strongly suggests that the use of archived web content in formal scholarship is very minimal, and that what use there is occurs not in area studies research, but in scholarship mainly on the topic of web archiving itself. Note that the research described here does not reflect the use of web archives for text and metadata analysis, network analysis, data mining, and other forms of whole-collection study increasingly practiced by researchers in the social sciences and digital humanities (Bailey and Blumenthal 2016, Sandvig and Hargittai 2015, Bode 2018, Maemura et al. 2018)—only (“only”) their use as content (and citation) repositories in traditional research contexts. The new horizons opening up and their potential for new paths of research—and potentially revolutionary new discoveries—will be considered in the following, final section of this report. Looking ahead, and to conclude this section, we can concur with the authors of the 2016 NDSA Survey who assert that: “. . . the lack of clarity on formal research use is understandable, but it does signal an area of activity that merits community attention, knowledge sharing, and success stories.” (Bailey et al. 2017, 27) By drawing attention to the value of web archives, documenting their use by improved and consistent citation guidelines, and sharing best practices, we can hope to bring this resource into the mainstream.

V. Conclusions and Recommendations: Accelerating the Integration and Mainstreaming of Web Archiving

Putting aside for now the need to address technology issues that undermine scholarly acceptance of web archives as evidence in research and publication, e.g. the “time skew” problem discussed above as well as continuing problems harvesting particular digital formats, we are left to discuss in greater detail what is being undertaken—and what could be undertaken—to bring about the integration of web archives into the normal flow of scholarly communication, that is, in the areas of discovery, use, dissemination, and citation. What does “normal flow of scholarly communication” actually mean, though, in the still-early 21st century? The radical shift in researcher behavior over the last two decades marked by the arrival of digital as scholarship’s primary site has been described concisely, at least regarding historians and social scientists more generally, in a recent article by Caribbean and Latin American scholar Lara Putnam in the American Historical Review. There she writes:

The accelerating digitization of primary and secondary sources and the rise of full-text web-based search to access information within them has transformed historians’
research practice, radically diminishing the role of place-specific prior expertise as a prerequisite to discovery. Indeed, we can now find information without knowing where to look. (Putnam 2016, 377. Emphasis added.)

This insight is corroborated for other disciplines in a series of field-specific surveys of contemporary real-world research that have been released by Ithaka S+R since 2012, starting with Supporting the Changing Research Practices of Historians (Rutner and Schonfeld 2012) and extending since then into the hard sciences (e.g., chemists), social sciences (e.g., economists), and humanities (e.g., art historians). First-person comments by actual researchers gathered and analyzed in these reports reinforce the conclusion drawn by Putnam that broad scans of the literature are increasingly taking the place of thumbing bibliographies and undertaking exploratory visits to physical libraries and archives:

“Google is the first port of call. A lot of times I will try to just start with a Google search.”

“[Google Books is] also helpful at the very beginning of a project, when you are not quite sure what sources you are going to use. Or you want to do a massive scan using keywords. I never did that until recently.” (Rutner and Schonfeld 2012, 17-18)

What this means is that in an increasingly digital research environment for both secondary and primary sources, resource discovery is becoming more the result of broad initial sweeps and then subsequent, more targeted and filtered searches—all online. Keyword and keyword phrase searching is preferred, while URL searching, important though it may be for later in the research and documentation process, can rarely serve as a primary discovery strategy. This appears to be the case not only in North America, but also internationally, as documented in a much-referenced New Zealand survey of researcher web archive use, which includes reviews of similar studies of researcher behavior in the UK, Netherlands, Portugal, and elsewhere. Therefore, mainstreaming web archives, putting them where researchers have no choice but to effectively trip over them in the course of their broad online searches, is a far better way to promote their use—and a far better way to increase still lacking awareness of their existence—than telling students and researchers about web archives in the abstract and encouraging them to go out and use them.

59 With thanks to Roger Schonfeld, Ithaka S+R, for drawing my attention to this important article.
60 “The favour of full text search access to web archives is also reflected internationally. The participants in the Netherland’s study preferred full text searches and regarded searches restricted to URL alone as a limitation. While the Portuguese study held that the prevalence of Google has resulted in an expectation for the availability of full text searching on the internet.” (Riley and Crookston 2015, 19)
A. Metadata

In the new research environment, and in importance second only to the preservation of web content itself, sound descriptive metadata created to a consistent standard and its exposure to major search engines and through library catalogs is essential to render a web resource both discoverable and determinably relevant to researcher needs. *Descriptive Metadata for Web Archiving* (Dooley and Bowers 2018), the much-anticipated report of the OCLC Web Archiving Metadata Working Group (WAM)—released in February 2018—emphasizes the important place of descriptive metadata in teaching and research and seeks to reconcile inconsistent approaches, since the “lack of a common approach to creating metadata [is] the most widely shared challenge across the web archiving community.” (Dooley and Bowers 2018, 5) The recommendations of the WAM report urge a “blending” or “bridging” of the differing practices of the library (bibliographic) community on the one hand and the archival community on the other, proposing a simple scalable set of data elements both can employ. The bibliographic/archival dichotomy parallels the two states of web pages, “live” and “past.” The translation of “live to past” in the archiving process is “strongly transformative,” as the report says, changing something which may have originally been intended as timely and exhortative, e.g. “Free Tibet!”, into an object of time-independent study. (Dooley and Bowers 2018, 9, quoting from Association of Research Libraries 2012, 26) Among the differences: the creating agency of a web post is rarely the same as the institution archiving it; and while creation and access dates are relevant for live sites, *date of capture* is key for the content as an archival item. There are numerous other differences, among them the different emphasis placed on item-vs. collection-level records, the former being the traditional bibliographic approach, the latter more typical of archival practice. “WAM’s recommended data elements are equally relevant for description of materials in both libraries and archives, and at both the item and collection levels.” (Dooley and Bowers 2018, 10) They are also scalable, allowing institutions to provide more or less descriptive metadata depending on available staff resources. And finally, the set of recommended data elements owes much to Dublin Core, which is understandable considering Archive-It’s reliance on this standard—Archive-It now being used by an overwhelmingly large and still growing percentage of institutions completing the 2016 NDSA Survey.61

B. Finding Aids

Even if online finding aids are not usually explicitly targeted by researchers looking for primary resources, they still provide some of the most important keyword and keyword phrase material

---

61 87%, up from 71% in 2013. (Bailey et al. 2017, 26)
for the broad scans that researchers, as we have seen, increasingly rely on. For resource discovery, keyword and metadata scanning has replaced the traditional published collection guides and the standard directories of the pre-internet era, consistent with Putnam’s observation that “we can now find information without knowing where to look.” This means that we can put both print and online directories behind us, which has in fact happened. The two principal online web directories, VLIB and DMOZ, no longer exist—DMOZ ceasing publication on March 14, 2017—not to mention artifacts of the 1990s like The United States Government Internet Directory. (Belovari 2017, 67) The subject resource guide of the Latin American Network Information Center (LANIC) based at the University of Texas Austin had been the premiere directory of Latin American content since 1992, but given the changing paths of scholarship, it suffered decline since 2011 or before and was “no longer being actively updated or maintained” as of July 2015. The directory approach complemented the primary URL search capability of the Internet Archive, which, though still available—and still important for locating known or hoped-for archived content—is of little utility as a discovery tool. The Internet Archive has therefore adapted to the new world of resource discovery, introducing, in 2015, a keyword search capability with a limited set of configurable user-determined parameters.

C. Citation Standards & Tools

Further downstream from resource discovery, consistent metadata practices are also key for useful and accurate source citation—citation which distinguishes between live and archived content, providing authors and readers of scholarly publications with the information necessary to replicate results—and thereby also increasing awareness of web archives, which in the minds of researchers are still often indistinguishable from live web content.

The lack of awareness of web archives is, as observed above, both reflected in and reinforced by the style guides consulted by students and researchers and by the citation software many use. EndNote, for example, supports a single reference type for “Web Page” rather than separate formats for live web pages and archived ones (see fig. 12). More serious is the lack of any distinction between live and archived web sources we see in style manuals, which, though being

62 This is comparable to the role played by LC subject headings in discovering library resources. Subject headings are rarely searched explicitly but are demonstrably helpful as providers of keyword material for broad literature scans. (Garrett 2007, Gross, Taylor, and Joudrey 2015, Gross and Taylor 2005)
63 http://lanic.utexas.edu/.
64 In late 2015, the Internet Archive “received a $1.9 million grant from the Laura and John Arnold Foundation to optimize the scope and quality of captured web pages [including] to make some keyword searches possible.” (Belovari 2017, 67, also the announcement of this welcome innovation at Goel 2016)
aware of DOIs and permalinks and even at times making passing reference to the Internet Archive, make little or no mention of any special metadata requirements for archived web content—or the importance of web archives for overcoming source volatility. (Davis 2016) The new 17th edition of the Chicago Manual of Style, though covering the proper use and citation of DOIs and permalinks, misses an opportunity to describe and thereby promote web archives in the appropriate section (“Sources Consulted Online,” 14.6–14.18). An entry for "web archives" is lacking entirely in its otherwise extraordinarily detailed 120-page index. (CMOS 2017, 745-750; 1015-1144)

As Graham and Norsworthy point out (2019, 224 et passim), the problems are not so much about technology anymore, but rather work cultures and the effective advocacy for responsible citation practice. The best way to drive use to archived web content is to require it, and we are at a point where we can just about do that. One example of precisely that—mandating the use of archival URLs—comes from the University of Heidelberg in Germany, where use of the authoritative DACHS “citation repository” is required for papers and dissertations in Chinese studies. It will in fact soon become a requirement for other area studies programs as well, among them the Heidelberg Centre for Transcultural Studies. (Lecher 2006)\(^6\)

Members of the web archiving community should reach out to norm-setting bodies on their campuses as well as to the editors and publishers of style guides—since publications such as the MLA Handbook, APA, and CMOS are widely used and consulted by students, researchers, and publishers alike. Students and scholars deserve guidance from “style guides” when it comes to citing the sources of the information they base their research on.

---

\(^6\) With thanks for additional information provided by Hanno Lecher, a digital librarian at the University of Heidelberg, and Xi Chen, Chinese studies librarian and East Asia collection strategist at UC San Diego Library in April and May 2018. Note: DACHS = Digital Archive for Chinese Studies.
D. Self-archiving

Several style manuals do make reference to self-archiving of consulted websites to help ensure that there is a permanent record of an author’s sources should the original online site disappear. But a page image in an author’s real or virtual filing cabinet is not the same as a reference to a universally accessible archive, and in that regard, not all self-archiving tools are created equal. Zotero’s self-archiving functionality, for example, creates only a local copy of a webpage in html, leaving out many of its components. Taking a “snapshot” (screen grab) of a consulted webpage and saving it as part of one’s own bibliographic record is closer to the practice of “scrapbooking” described above—but as experts have warned, “snapshots” are extraordinarily easy to falsify “so long as you have a rudimentary understanding of web development.” (Feldman 2015)

Even without these concerns, smartphone snapshots of computer screens and the “scrapbooking” of webpages visited are primitive tools for documenting sources. Also, these DIY tools may capture individual webpages, but fail completely as tools to capture entire websites in all their complex hypertextuality.

Far superior from the standpoint of verifiability and secure scholarly communication is the recently introduced self-archiving feature of the Internet Archive, “Save Page Now.” It is easy to use, strongly resists manipulation or falsification, and returns a permanent and citable archival URL, thereby protecting authors from the fate of so many who see their web links disappear after publication—or even before. Other initiatives, such as rhizome.org’s Webrecorder, the recent recipient of a major grant from The Andrew W. Mellon Foundation, promise to lead researchers and authors to a more comprehensive capture of webpages, enabling archiving of data formats such as embedded video which more primitive copying technologies cannot capture. (Rhizome 2018) The goal should be to allow individual researchers and authors to make sure they are citing persistent, complete, and authoritative archived pages (and sites) accessible without having to “contact the author.”

---

66 Thanks to Geoff Morse of Northwestern University Library for his patient explanation and demonstration of Zotero’s capabilities.
E. Credibility Enhancement

Considering the ease of falsifying casually archived web content, and relatedly—as the Joy Reid case documents—the difficulty of defending web archives with unwelcome content against the charge that they have been hacked and their content maliciously altered, it’s not surprising that many researchers harbor a deep mistrust of web archives as a whole. As Belovari writes,

Web archives managers informally recount how scholars leave their sites quickly for Google and other corporate search engines where, as of now, they can still find some older and cached websites, regardless of all the caveats connected to them (e.g. there is no way to verify whether, when and how websites have been altered over time). (Belovari 2017, 65)

Equally disturbing for researchers trying to wrap their minds around this new type of archival resource is that the archived web page or website is not an “original”—or even a “carbon copy” or “snapshot” of an original. In Niels Brügger’s formulation, “the archived website can be considered an actively created and subjective reconstruction of what was once online . . . [I]n this respect it differs significantly from other media types.” (Brügger 2012, 320) A web archive is therefore more like Arthur Evans’ restoration of Knossos or the open-air Plymouth Colony museum than a stored and preserved original artifact. We have already seen above that some media parts of web pages are still not capturable by existing crawler software (though this is rapidly changing), and that on a more basic level, archived web pages are not usually the result of a single crawl but rather of several, spread out (or “skewed”) over hours, days, weeks, months, and even years—resulting in what are called temporal violations. (Nelson 2018a) Archived web pages are, in other words, inherently inauthentic, yet at the same time they are the best—and at times only—evidence of many domains of modern culture that we have, and it would be a great loss not to work with them.

Some of the issues that deter scholarly use can be dealt with through improved technology, although as in so many other areas, there may always be a lag between innovation in one area and the adaptation of other technologies that interface with that innovation—in our case, web archiving technology—as they race to catch up. Other deficiencies of web archives may be more inherent, as Brügger suggests above. Absent a miracle “cure,” these deficiencies will only be dealt with through greater user awareness of them, i.e. better education, remembering that even in the traditional paper archive, records can be incomplete and falsified—Donation of Constantine, anyone?—and that researchers using any archive will always be called upon to fill in blanks, question (or establish) authenticity, and reconstruct the historical record based on knowledge, research, and, yes, educated guesses. Archival research in both the physical and digital worlds has never been an exact science.
In other cases, however, there is more that the web archiving community can do to enhance the credibility of the resources they preserve and make accessible. As pointed out in a recent advocacy piece from the University of Toronto, a “key barrier to the use of web archives for research purposes” is the “current lack of transparency in communicating the processes of web archiving to web archives users.” (Maemura et al. 2018, 3, 6) These authors argue for establishing “mechanisms for transparency in the making of web archives to facilitate the process of evaluating a web archive’s provenance, scoping, and absences.” Some of the most basic questions are perhaps obvious: How are seeds selected? Why are others left out? Other questions are perhaps less so. Maemura and collaborators offer these description points for web archives:

- Appraisal and selection
- Scoping decisions  
- Acquisition or data capture
- Organization and storage
- Quality assurance reviews
- Description and metadata (Maemura et al. 2018, 7-8)

This is useful information for all archives users but is essential for researchers whose “approaches take media as data,” since if certain categories of information are left out of crawls, the quantitative representativeness of the data corpus is affected, slanting the global results.

Ultimately, an independent certification process for web archives may be desirable. It would need to address at least two issues: security measures in place, to avoid the fate of the CIP Americas archive presented earlier in this report; and integrity in light of the description points enumerated above. Researchers and others have an interest in knowing what protective measures are in place at a particular “archive,” up to and including the Internet Archive, to ensure that it is worthy of the name and can be relied on as an authoritative source into the future.

The UK’s Digital Curation Centre (DCC), among other agencies and professional organizations, has produced a list of online tools and methodologies for the audit, assessment, and certification of digital repositories. (Digital Curation Centre [DCC] 2018) What The Andrew W. Mellon Foundation’s Don Waters has described as the goal for the digital environment overall—that it become “a natural place to do scholarship”—also applies within the realm of web archives. (Andrew W. Mellon Foundation 2016) The goal, again, is to advance the integration of archived web content into mainstream scholarly communication.

---

67 “Scoping decisions determine which resources will be captured, through limiting to specific domains or media type, as well as crawl duration, or crawl frequency.”
F. Education, Outreach, Promotion

Without an awareness that “there are few parallels between the traditional archive and the web archive,” and that what students and researchers must do is acquire “a new form of digital literacy,” most web archive novices (including seasoned scholars) will run for the exits—and back to the live web—rather than stay to exploit the riches that web archives hold. (Ankerson 2015, 31) This awareness usually begins by encountering archived web content in a familiar context, e.g. a Google search or in a footnote. It develops further with education and user support for students, faculty, and library/archive staff. Ideally, citations by web archive–acculturated students and scholars will lead wherever possible to publicly accessible and stable archived sites rather than to the ephemeral (rotting), frequently morphing (drifting) sites on the live web:

Libraries and archives should actively engage in outreach to both current and potential web archives users: first, to provide an initial understanding of what web archives are, and how to find and use them; and second, to better understand users’ research challenges in order to find ways to ameliorate them. (Dooley and Bowers 2018, 12)

Institutions and funding sources can accelerate this learning process by hosting panels on campus and at professional meetings showcasing successful uses of archived web material, from strategies of resource discovery to best practices of citing archival URLs. Today’s focus on “best practices” too often highlights huge web-focused projects which surpass the digital competence (and interest) of most students and researchers. Not every meaningful use of web archives needs to be on the scale of Anat Ben-David’s recreation of the entire former Yugoslav top-level domain. (Ben-David 2016) More interesting and potentially more meaningful would be drawing attention to the incremental steps which are within everyone’s reach, which would speak to normal members of the academic community with far more on their minds than the history and idiosyncrasies of the web. As Eszter Hargittai and Christian Sandvig have written, “the actual revolution in digital research instrumentation is going on now, all around us, in smaller, ‘ordinary’ research projects.” (Sandvig and Hargittai 2015, 11)

Let’s draw attention to those ordinary, extraordinary projects, too.

G. Data Mining & Whole-Collection Analysis

Just as we don’t want to overlook the importance of small steps in “ordinary” research, however, we also must not overlook that the amassing of huge amounts of archived information is itself creating a new resource allowing new questions to be posed—and answered—if the right tools are applied. But what are these new questions? And what new tools do we need? This is an iterative process, and we shouldn’t underestimate the difficulty nor the length of time it will take
to make substantive process, not to mention shift long-established paradigms. As Jill Lepore wrote in her 2015 article for the *New Yorker*,

... the tools for doing anything meaningful with web archives are years behind the tools for creating those archives. Doing research in a paper archive is to doing research in a web archive as going to a fish market is to being thrown in the middle of an ocean; the only thing they have in common is that both involve fish. (Lepore 2015)

In the same article, Lepore also quotes Stanford’s web archivist Nicholas Taylor, who describes the quandary thus: “We don’t know what tools to build, because no research has been done, but the research hasn’t been done because we haven’t built any tools.”

Much progress will therefore be accidental and incremental. This is actually the way science often happens. Hargittai and Sandvig, channeling science historian Derek J. de Solla Price, talk about how many of the great discoveries of the last five hundred years were made possible by adapting advances invented for one purpose to serve another:

The first pumps were created for firefighting and mining, but they led to the realization that air is a gas. The cloud chamber was an attempt to create artificial clouds built by a mountaineering buff, but it then became the way to visualize atomic particles. The telescope was first marketed as a novelty, then as a military device, and then Galileo used it to discover that the moon had mountains. (Sandvig and Hargittai 2015, 7)
This was the era of “instrumental revelation” (de Solla Price). We are perhaps standing on the brink of a similar era, with data mining, digital humanities, new concepts of “distant” or “scalable” reading, and then the developing body of speculative thought in the humanities known by cognoscenti as the “Macroscope,”* producing myriad new patterns and constellations of thought, imagery, and metaphor. Some of these will of course be dead ends or visually appealing but otherwise inscrutable images. Creating “heat maps” based on mapped data concentrations, as in an image (previous page, Fig. 14) showing the intensity of Japanese-language business communication in Shanghai, results in intriguing images, but what are such visualizations good for? Eric Fischer’s phenomenal visualizations based on millions of Twitter geotags—like his rendering of O’Hare Airport shown below—are often beautiful and fascinating, evoking as they do microscope images of the human bloodstream or nervous system (or of mold colonies), but standing alone, without fitting into a context of inquiry, they are meaningless.

Using Google’s Ngram viewer, which mines the contents of hundreds of thousands of works in Google Books, we can see the frequency of word occurrences ebb and flow over time, also in comparison with each other. But again, as Sandvig and Hargittai ask: If this is an answer, what is the question?

We could ask ourselves: “What caused the dramatic shift in the popularity of the term ‘ice cream’?” But answering that question is sensible only if a fact can be situated in an intellectual frame and context that makes it evidence for a debate of some importance. We could also ask ourselves: “Who cares?” (Sandvig and Hargittai 2015, 10)

---

* On the origins of the “Macroscope” concept, its variations, and the potential of the paradigm for the humanities, see the overview by Tim Hitchcock in his blog Historyonics and literature referenced there. (Hitchcock 2014)
On the other hand—and in answer to the question “who cares” about heat maps and other visualizations—there is an enormous market for big data, from city planners, election demographers (think: Cambridge Analytica), telecommunications companies, national security agencies, hedge fund managers, also economic historians and political scientists, who trace and study the growth and demise of markets or the spread of Islamist ideologies. The real question is whether web archives data can be served up in a form useful to digital social scientists and humanists and how questions can be formulated in a way the data can respond to. Among the pioneers in this endeavor are Archive-It Research Services, who continuously offer real-life use scenarios and projects to stimulate further inquiry (cf. Bailey and Blumenthal 2016) and the Archives Unleashed Cloud, or AUK, based in Canada at the University of Waterloo and at York University in Toronto, currently the beneficiary of significant support from The Andrew W. Mellon Foundation. In AUK’s own self-presentation, we sense the optimism and enterprising spirit behind the project (whose mascot is, yes, an auk):

. . . we’re convinced that web archives are key to the future of scholarly research. Given the exponential rise of digital-born content, the opportunities for scholarly inquiry using digital content as a primary source of exploration are boundless. Previous work in the web archiving field has demonstrated that the sheer scale of working with web archives can be overwhelming. . . . The Archives Unleashed Cloud (AUK) is accordingly motivated by a desire to lower barriers to working with web archives at scale by providing a convenient, practical, and user-centered tool that takes the stress out of complicated coding scripts and the dreaded command line. (Fritz and Milligan 2018)

But returning for a moment to the value of no-holds-barred speculation and conjecture which has played such a role in the history of science: What is wrong with just looking at entrancing data visualizations, like Fischer’s, and asking what they might be good for? Why not just ponder these images—or even the ups and downs in the frequency of the term “ice cream” in a million books—and see what happens, what insights might dawn? As one data scientist proposes, “Do not assume away discoveries just because you don’t know what you don’t know . . . Let the data take your brain on an unguided journey where discovery is the ultimate destination.” (Smith 2013)

H. Promoting Collaboration

For a single institution with ambitions in the web archiving field, the commitment of staff and the investment in infrastructure may seem too daunting. So in conclusion we should highlight the Cobweb initiative developed by the California Digital Library, UCLA, and Harvard Law School.

69 With thanks to CRL’s Bernie Reilly for engaging me on the question of the utility—or lack of it—of mining the petabytes of data rapidly accumulating in web archives.
Its goal is to promote greater institutional collaboration in web archiving, bringing together the collections, curatorial skills, and technical expertise which are often not all in one place, but can be distributed between different institutions. Cobweb offers an open source platform for finding potential institutional partners and then structuring that collaboration in an efficient way. It was recently the beneficiary of a one-year IMLS grant. (Stine 2018)

In the context of international and area studies, and specifically Latin American and Caribbean studies, the highest priority of all three programs studied is and surely must remain the identification, capture, and preservation of as much content as possible before it goes away—preserving it not randomly, given the sheer overabundance of it, but through engagement of scholars and international partner organizations in identifying materials that are useful, endangered, and unique. Then, as resources allow, this content must be made discoverable, searchable, citable, recoverable, and manipulable by any serious investigator. The technological obstacles to achieving this goal are being overcome far more rapidly than the cultural ones, which as we have seen include the many segments of academic and publishing culture—the whole culture of scholarly communication in fact. These will evolve slowly but surely with attention paid to them by administrators, funding bodies, and, of course, the web archiving community itself. The institutions we have looked at closely in this study are exemplary in the degree to which they have pioneered integration of born-digital archiving expertise into traditional organizational structures, and to which they envision a future which we should together now seek to realize.

VI. Cited Literature


Chavez, Marisela A. 2015. "Não é apenas sobre nós: Food as a Mechanism to Address Social and Environmental Injustices in Mato Grosso, Brazil." M.A., University of Kansas. 
https://pdfs.semanticscholar.org/8cbf/707cae7609163505e9b0c0fa4e7e1243209.pdf.


Documenting the Now Project. 2017. "Forum on Ethics and Archiving the Web." 


[http://www.chrisharrison.net/index.php/Visualizations/InternetMap](http://www.chrisharrison.net/index.php/Visualizations/InternetMap)


[https://doi.org/10.1080/13614576.2010.528342](https://doi.org/10.1080/13614576.2010.528342).

[https://doi.org/10.1371/journal.pone.0115253](https://doi.org/10.1371/journal.pone.0115253).


[https://www.newyorker.com/magazine/2015/01/26/cobweb](https://www.newyorker.com/magazine/2015/01/26/cobweb).


[https://www.loc.gov/webarchiving/preservable.html](https://www.loc.gov/webarchiving/preservable.html).

LLILAS Benson Latin American Studies and Collections. 2017. The University of Texas at Austin Receives Mellon Foundation Grant to Pioneer Archival Projects in Latin America. Austin, TX: University of Texas Libraries. 

Maemura, Emily, Nicholas Worby, Ian Milligan, and Christoph Becker. 2018. "If These Crawls Could Talk: Studying and Documenting Web Archives Provenance." Faculty publications - Faculty of Information Collection. 
[https://tspace.library.utoronto.ca/handle/1807/82840](https://tspace.library.utoronto.ca/handle/1807/82840).


