

## Preserving Web Resources for Research: Latin America as a Microcosm

The Challenges of  
Consulting Web Content  
for International and  
Area Studies: Latin  
America as a Test Case 2

Archiving the Latin  
American & Caribbean  
Web: Three U.S.  
Initiatives 7

Born-digital Primary  
Sources for Area and  
International Studies:  
New Models and  
New Threats 14

Demonstration in Mexico City on the  
disappearance of 43 students in Ayotzi-  
napa, Website photo essay by Clayton  
Conn [https://www.americas.org/  
ayotzinapa-three-years-later/](https://www.americas.org/ayotzinapa-three-years-later/)



### *In This Issue*

In 2016 the Andrew W. Mellon Foundation awarded the Center for Research Libraries funding to develop an “integrated, self-sustaining, international cooperative framework to support area and international studies (AIS) in the humanities and social sciences.” The chief goal of the Global Collections Initiative was to expand electronic access to primary source documentation and data for scholarly research on major world regions like the Middle East, sub-Saharan Africa, and South Asia, 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.

In this issue we report what we learned about the challenges of provisioning scholars in academia and public policy with documentation and data available only in digital form. Jeffrey Garrett shares the findings from his evaluation of the state of web archiving. The third essay discusses persistent threats to the survival and accessibility of digital data and evidence, and suggests some solutions.

—Bernard F. Reilly  
President

# The Challenges of Consulting Web Content for International and Area Studies: Latin America as a Test Case



**Jeffrey Garrett**

*Librarian Emeritus, Northwestern University*

*Consultant, Global Collections Initiative, Center for Research Libraries*

1. Jill Lepore. "The Cobweb: Can the Internet Be Archived?" *New Yorker*, January 26, 2015. <https://www.newyorker.com/magazine/2015/01/26/cobweb>.

2. See for example Rebecca B. Galembo. *Contraband Corridor: Making a Living at the Mexico-Guatemala Border*. Stanford, CA: Stanford University Press, 2018. About 10% of the 102 links to open web content in this bibliography examined as part of the CRL study were no longer functioning by the time of its late 2017 release, including two to the now defunct [cipamericas.org](http://cipamericas.org) address (p. 265).

3. According to the monitoring site Internet World Stats, of 4.2 billion internet users in the world (as of June 30, 2018, 10.4% reside in Latin America and the Caribbean, while 8.2% are in North America. Enrique de Argaez. "Internet World Stats: Usage and Population Statistics." <https://www.internetworldstats.com/stats.htm>.

4. Images by Chris Harrison of the Human-Computer Interaction Institute at Carnegie Mellon University. Chris Harrison. "Internet Maps." <http://www.chrisharrison.net/index.php/Visualizations/InternetMap> and personal communication.

5. Pamela M. Graham and Kent Norsworthy. "Archiving the Latin American Web: A Call to Action." In *Latin American Collection Concepts: Essays on Libraries, Collaborations and New Approaches*, edited by Gayle Williams and Jana Krentz, 224–236. Jefferson, N.C.: McFarland, 2019.

Former president of Honduras Manuel Zelaya. When his government was overthrown in 2009, the entire content of its web presence was deleted. Detail of photo from Wikimedia Commons, by Ricardo Stuckert/PR.

In 2018 CRL commissioned a [report](#) assessing the effectiveness of current efforts to archive open web content as source materials for international and area studies (IAS) research, focusing specifically on the Caribbean and Latin America. While the web is now a key delivery medium for news, data, and contemporary discourse, 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 communications. Conducting, sharing, and reading research based on open web sources is “like trying to stand on quicksand,” Harvard historian Jill Lepore noted in 2015.<sup>1</sup> While issues of preserving and citing content that is published in online scholarly journals have largely been resolved, the accessibility and citability of fugitive source materials existing on the open web remain a problem.<sup>2</sup>

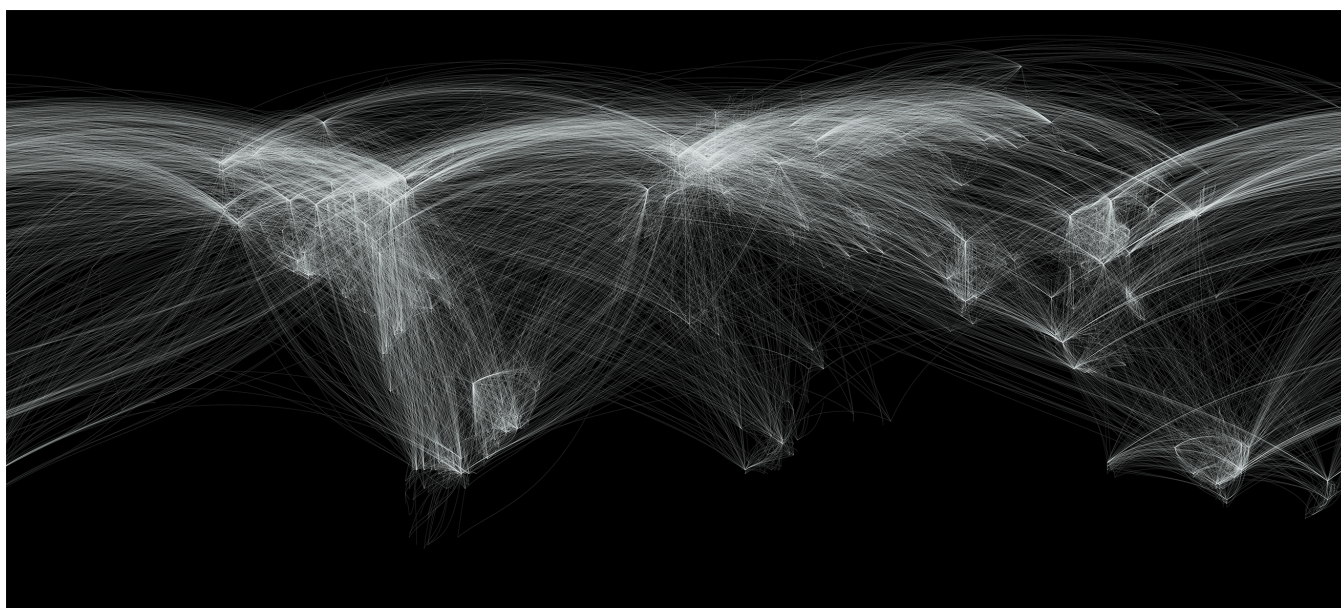
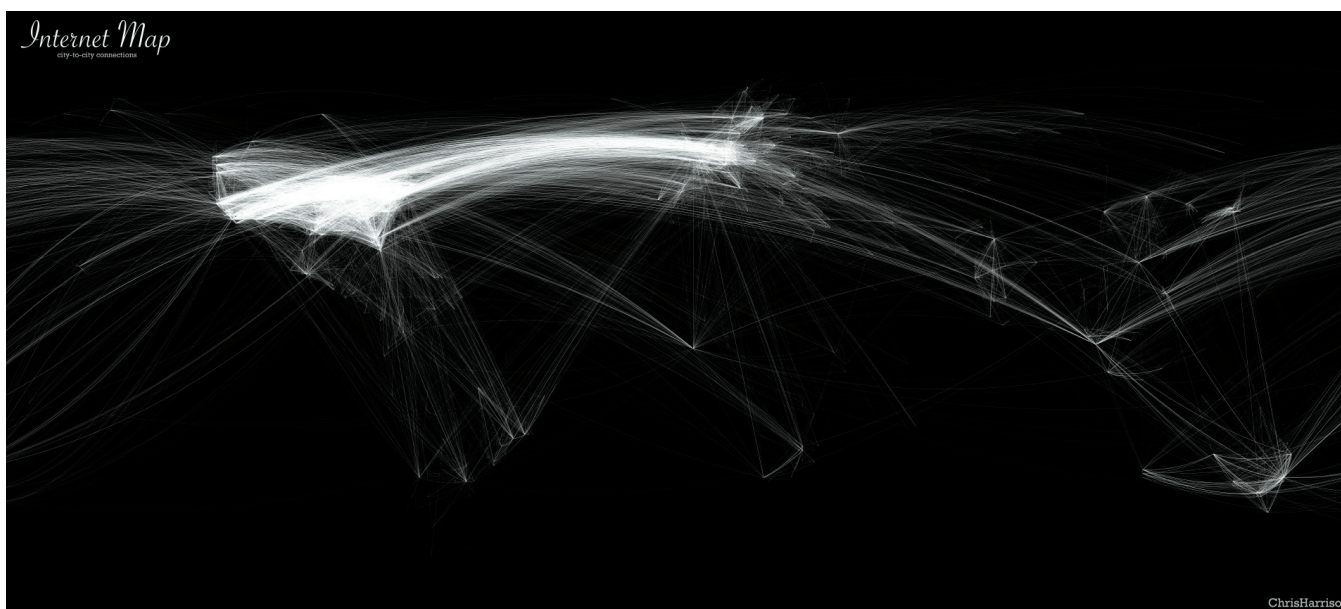
This article—the first of two highlighting aspects of the CRL report—examines systemic and region-specific issues arising from proliferating web content in Latin America, demonstrating challenges encountered in both real and hypothetical research examples.

## 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, and now exceeds that of the United States and Canada.<sup>3</sup> The first image on page three shows city-to-city internet connections between North America, Europe, Africa, and South America in 2007. The graphic beneath it, created in 2011 using the same mapping algorithm, reflects the density of internet connections just four years later. 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.<sup>4</sup>

Use of open web content, consisting of commercial, political, cultural, and scholarly websites, as well as the accessible subset of social media interchange, 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, as Pamela Graham and Kent Norsworthy point out, “much digital publishing [in Latin America] is not channeled or distributed through traditional publishers but is instead only taking place on the freely accessible web.”<sup>5</sup> 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





Worldwide City-to-City Internet Connections in 2007 and 2011. Courtesy of Chris Harrison, Carnegie Mellon University.

6. Tim Berners-Lee, who developed the Hypertext Transfer Protocol (HTTP), the core DNA of the modern-day internet, 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)

has traditionally provided the basis for the historical record. There can, therefore, be no doubt that the capture and archival preservation of web content from 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 faces particular challenges.

Unfortunately, preserving web content in Latin America has been especially slow getting off the ground. Some of the reasons are inherent to the web itself, while others are specific to Latin America and its history. On the systemic side, first, there is the inherent ephemerality of the medium: new content constantly overwrites the old, leaving not a trace of what had been there before.<sup>6</sup> Second, there is still a prejudice held by many that content on the web has no heft, that it is more akin to idle conversation than content that merits preservation. For centuries this perceived lack of archival “worthiness” has made ephemeral formats—pamphlets, posters,

playbills, newspapers—a lower priority for library preservation, despite the role ephemera have played in documenting, even precipitating, momentous events of history.

Third and finally, on the systemic side ethical concerns stand in the way of the preservation of much currently relevant web content, especially social media. These concerns become even more acute for web content in the human rights domain, where personal data regarding victims, informants, and perpetrators may become exposed to public view in a way paper archives cannot be. The attendant moral, legal, and political issues are aggravated by the globally perceived sense, underscored by news almost every day, that big American tech firms are predatory data gatherers, unconcerned with personal privacy and safety. Such concerns could easily scare off many smaller institutional players both here and abroad whose collective efforts to collect web-based text and data in Latin America and the Caribbean are essential, but are becoming more complicated—legally, ethically, and logistically—than ever before.

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. 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, consolidated in the 18th century in Seville, at the Archivo General de Indias.<sup>7</sup> 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).<sup>8</sup>

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 recent past is so easy on the web is a gift to rulers seeking to eradicate the memory of their predecessors or, perhaps, information reflecting poorly on their own regimes. To offer just one example: 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.<sup>9</sup>

In Latin America, as elsewhere, valuable websites can also be hijacked by hostile commercial or political actors. For example, beginning sometime in the last few years and lasting until recently, visitors to [cipamericas.org](http://cipamericas.org), the website of the Center for International Policy's Americas Program based in Mexico City, found themselves redirected to a website offering cannabis derivative products. Faced with this problem, the parent organization, based in Washington, DC, decided to migrate to a new domain, [americas.org](http://americas.org), where all their "archived" content can once again be found. The move to [americas.org](http://americas.org) solved one problem but has caused others, since until 2007 [americas.org](http://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*. Live web links to their content are now broken, too: the only access is through the Internet Archive.<sup>10</sup>

An inability or disinclination to rely on durable web infrastructure can also affect sustainable access for Latin American and Caribbean studies. Research on Cuban literature, for example, cannot overlook that much new work is circulated online, via blogs and webzines, and is backed up by individual readers only on "flash drives, kindles, etc. (and even, sometimes, in hard copy) before, during, and after

7. <http://www.mecd.gob.es/cultura/areas/archivos/mc/archivos/agi/presentacion/historia.html>.

8. See <http://netpreserve.org/about-us/members/> for a map showing the worldwide distribution of IIPC members.

9. For more on politically motivated website disappearances in Latin America—including the fate of the Zelaya government pages and the Zelaya administration's post-coup afterlife as a website—see Kent Norsworthy. "Web Archiving and Mainstreaming Special Collections: The Case of the Latin American Government Documents Archive." In *The Signal*, interviewed by Trevor Owens. Washington, D.C.: Library of Congress, 2012.

10. Graham Stinnett. "Rebel Collectors: Human Rights and Archives in Central America and the Human Rights Commission of El Salvador and the Resource Center of the Americas, 1978–2007." Thesis, University of Manitoba/University of Winnipeg, 2010. A photo illustration from the current site of [americas.org](http://americas.org) introduces this issue.

circulation online.”<sup>11</sup> Recognizing how fragile this distribution infrastructure is, individual scholars in the United States, as well as other countries, have used personal websites to store some of Cuba’s literary production and to share it on the web.<sup>12</sup> These are by no means comprehensive, much less durable “archives.”

It is, however, both inaccurate and unfair to single out the Latin American web as somehow unique for link rot, content drift and loss, hijacking, and other forms of website abuse and manipulation. 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, has not occurred in the region as it did in multiple cases in Eastern Europe during the 1990s.<sup>13</sup>

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

To further illustrate the challenges facing researchers using web archives for this region, one might 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 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. She begins with broad scans in omnibus databases, such as Google, Google Scholar, Google Books, JSTOR, ProQuest Global Dissertations, and others. She discovers 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.” 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](http://programadegoverno.dilma.com.br), no longer exists on the live web—an instance of *link rot*. What to do? She does succeed in finding the document—or at least a document bearing the same title—on the live web, 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 some point after the election or after Dilma Rousseff was impeached in 2016 (*content drift*).<sup>14</sup> 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 is not currently an accessible collection created for the 2014 Brazilian election.<sup>15</sup>

Finally, 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 Internet Archive, she can hope that it will also have a permanent location findable by later scholars. The only discomfiting fact is that the time-date-stamp of the 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 disparity is often called a *time skew*. The problem derives from the fact that an archived

11. Emily A. Maguire. “Islands in the Slipstream: Diasporic Allegories in Cuban Science Fiction since the Special Period.” In *Latin American Science Fiction: Theory and Practice*, edited by M. Elizabeth Ginway and J. Andrew Brown, 19–34. New York: Palgrave Macmillan, 2012. Also personal communication.

12. For example, there is an archive of the Cuban SF magazine *Disparo en Red* at the University of South Florida, see <http://digital.lib.usf.edu/disparo>. Upper levels of this site have been crawled (and are preserved) by the Internet Archive, but individual issues of this magazine (active between 2004 and 2008) are not.

13. Anat Ben-David. “What Does the Web Remember of Its Deleted Past? An Archival Reconstruction of the Former Yugoslav Top-Level Domain.” *New Media & Society* 18, no. 7 (2016): 1103–19.

14. 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.

15. According to Library of Congress, LC did indeed crawl and archive the 2014 election in Brazil: as of this writing it was planned to be mounted as a collection in the near future.

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”: a composite reconstruction using crawls of different elements of a live web page undertaken at different times.<sup>16</sup>

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 many web archives in the United States and Canada. This is, of course, a significant flaw in web archiving technology, at least from the perspective of researching historians.<sup>17</sup> Perhaps this helps explain why so few researchers use archival versions of websites in published research, instead preferring to footnote a live website that may no longer exist or whose content may have 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.

We live in a new documentary age when it comes to studying and reporting on the societies, politics, economies, and cultures of Latin America and the Caribbean. The fixity of past research material formats is gone, superseded by electronically produced and distributed source materials that morph or even disappear entirely before or after research based on them is shared. The following article looks closely at several large-scale preservation efforts in the United States to see how these programs approach the monumental task of capturing and preserving the evanescent forms information today often takes—and to see whether scholars are using these stable and preserved resources to document their work. ❖

16. Gordon Mohr. “Wayback Machine & Web Archiving Open Thread, September 2010.” In *Web Archiving at archive.org*, edited by Internet Archive Web Team: Internet Archive, 2010.

17. Susanne Belovari. “Historians and Web Archives.” *Archivaria: The Journal of the Association of Canadian Archivists*, no. 83 (Spring 2017, 2017): 59–79.



# Archiving the Latin American & Caribbean Web: Three U.S. Initiatives

**Jeffrey Garrett**

*Librarian Emeritus, Northwestern  
University*

*Consultant, Global Collections Initiative,  
Center for Research Libraries*

In this article, we consider three programs in the U.S. which are key library-based initiatives for archiving ephemeral, freely accessible web content from Latin America and the Caribbean. The programs we review are: The Library of Congress's Web Archives (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. This article describes how these programs approach the sustainable capture of open web content, and the extent to which they succeed in providing archived content useful in the teaching, research, and publication mainstream. More detailed descriptions of each initiative can be found in the [CRL report, An Evaluation of Web Archiving Programs in the US Relevant to International and Area Studies 2019](#).

## Web Archiving at the Library of Congress<sup>1</sup>

### History

National libraries around the world have recognized since the 1980s that their collecting responsibilities need to encompass the digital realm, especially where materials historically submitted to them in print form for legal deposit now exist only in digital form. Starting in 1996 some countries, among them the UK, Australia, Sweden, and Denmark, passed legislation to allow or mandate the collecting 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). As its first major project, MINERVA worked with the Internet Archive (IA) to archive the 2000 presidential election. Then, in the wake of the 9/11 terrorist attacks on the Pentagon and the World Trade Center, LC harvested domestic and foreign websites reflecting world reaction to these events, preserving this content before it disappeared. Over 30,000 websites were captured at that time: the September 11, 2001 Web Archive is today LC's single most visited web archive collection.<sup>2</sup>

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 its Archive-It interface or the Wayback Machine; instead, the archived

1. Discussion of web archiving activities at the Library of Congress is based primarily on publications by the LC Web archiving team led by Abigail Grotke; 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.

2. <https://www.loc.gov/collections/september-11-2001-web-archive/about-this-collection/>.

content is loaded on LC servers and is available only through the LC portal. Moreover, much content is available only on the LC premises, analogous to the treatment of deposited print publications. Unlike many other countries, however, LC has never been granted a legal mandate requiring publishing entities and individuals to deposit their digital output, and conversely, it is not legally required to archive websites. This has led to a complex system of permission requests.

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](http://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<sup>3</sup>, and the archives of the two Brazilian presidential elections of 2010 and 2014, the latter of which was as of this writing not yet officially posted.

LCWA's most recent annual report indicates the total size of the archive is currently 1.3 petabytes.

### Governance and Selection

Policies governing the selection and archiving of the foreign sites (of importance to Latin Americanists and other area studies researchers) are given special attention in a set of “Supplementary Guidelines”<sup>4</sup> to LC's Collections Policy Statements:

Foreign websites 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 be made if there are concerns over the long-term accessibility of a foreign website.

Proposals are being actively encouraged since LC 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.).”<sup>5</sup>

On the topic of Twitter and social media, much has been made of the LC Twitter archive, first acquired in 2010 with tweets going back to 2006 and with the charge to include all public tweets going forward.<sup>6</sup> The collecting mandate is no longer comprehensive, reflecting many concerns on the part of LC, among them the need to honor deleted requests, but especially the widely varying needs of researchers who want to use the vast amounts of data collected for projects in a multitude of fields and disciplines. Permission must be granted by both LC 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 licensed by Twitter like Dataminr and Gnip, typically to mine feeds for certain topics, opinion research, trends, and other patterns.

### Support and Collaboration

Not having the sweeping digital depository mandate of national libraries in other countries, such as Britain, France, Denmark, among others, LC 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 messages 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.

3. <https://www.loc.gov/collections/brazil-cordel-literature-web-archive/>.

4. <http://www.loc.gov/acq/devpol/webarchive.pdf>.

5. Michael Matos, Library of Congress collection development analyst, personal communication.

6. It should be noted that this project is organizationally entirely separate from LCWA.



Matters are much more complicated with hosts of websites in foreign countries. 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. And yet, ultimately a remarkable level of coverage has been achieved for LC’s archives of foreign elections and of the international response to 9/11.

### Use Analysis and User Feedback

Use data for LCWA is collected using the Adobe Marketing Cloud. The number of total visits in 2017 for the archive interface at [loc.gov/websites](http://loc.gov/websites) was 180,238, or roughly 500 a day. Recognizing that the value of archived content will grow over time as live websites now being archived disappear, the LCWA Team is primarily focused on building the archive rather than on performing downstream use analysis, at least at this time. This is not unusual across the country.<sup>7</sup>

Not much is known about specific uses of archived content in the LC Reading Room or elsewhere in the country or world. As an independent investigation suggests, LCWA is only infrequently cited in published research. LC is beginning to experiment with creating data sets that will allow researchers to use the archives in new ways. It will continue to collect expansively, working with partners across the globe, especially through organizations such as the International Internet Preservation Consortium (IIPC).

## Web Archiving at Columbia University Libraries: The Human Rights Web Archive<sup>8</sup>

### 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. By 2013, Columbia was funding its own Web Resources Collection Program, which includes large thematic web collections in areas such as 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 documentation held at Columbia. Organizationally, it is an initiative of the University Libraries’ Center for Human Rights Documentation and Research (CHRD).<sup>9</sup> 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. Project work on HRWA transitioned to programmatic work in 2010. As of early 2018, the project had collected 15 terabytes of data and has active harvests of about 700 seeds.

### Governance and Selection

HRWA is the largest of four thematic web collections being built by CUL’s Web Resources Collection Program (WRCP). 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. There is frequent interaction, both formal and informal, between HRWA staff and the university’s faculty and

7. 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.” Jefferson Bailey, Abigail Grotke, *et al.* “Web Archiving in the United States: A 2016 Survey.” (February, 2017). [http://ndsa.org/documents/WebArchivingintheUnitedStates\\_A2016Survey.pdf](http://ndsa.org/documents/WebArchivingintheUnitedStates_A2016Survey.pdf).

8. Sources for this description include an initial phone interview with project coordinator Pamela Graham on December 15, 2017; then an onsite meeting on January 26, 2018.

9. For a full description of the project, refer to <http://hrwa.cul.columbia.edu/about>.

students: an important source of useful intelligence about existing as well as new potential seeds for the web archive.

The value of conjoining traditional collection expertise with the selection and management of seeds at HRWA is made clear in a publication co-authored by Pamela Graham of CHRDR:

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 . . . Examples include websites of marginalized social groups or movements, or emerging writers or artists who only disseminate their work online.<sup>10</sup>

### Use Analysis and User Feedback

Google Analytics data from the HRWA Archive-It account show that since 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. 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 great attention paid to HRWA content. The actual impact of the archiving activity on published research and scholarship has been difficult for Columbia's web archiving staff to assess—it has also not been the highest priority to do so.<sup>11</sup> As with the Library of Congress Web Archives, questions of documented use tend to be put aside for the present in favor of creating and enhancing well-curated, technologically robust archives. Use will inevitably rise as live-web versions of archived sites go offline or their content “drifts.”

### Challenges and Future Plans

There are, of course, still a host of challenges for Columbia's HRWA, including technical issues involving the locally developed search interface, and legal issues related to copyright. To expand use, institutional commitment must continue at the current high level. Looking at the big picture nationally, as Graham and Norsworthy do in their book chapter, there is a keen sense at Columbia that the “primary obstacles to expanding these activities in libraries are less on the ‘technology’ side and more on the ‘cultural’ side.”

## Web Archiving at the University of Texas at Austin: LAGDA and HRDI<sup>12</sup>

### 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 the Teresa Lozano Long Institute of Latin American Studies (LLILAS) Benson Latin American Studies and Collections. LLILAS Benson is one of the world's most important centers for the study of Latin American history, culture, politics, and society. LLILAS's interdisciplinary program integrated more than 30 academic departments across the university. The Nettie Lee Benson Latin American Collection is one of the world's premier repositories of Latin American and U.S. Latina/o materials.

The Benson's physical collections number over a million volumes, to which are added a wealth of original manuscripts, photographs, and various media related to

10. Pamela M. Graham and Kent Norsworthy. “Archiving the Latin American Web: A Call to Action.” In *Latin American Collection Concepts: Essays on Libraries, Collaborations and New Approaches*, edited by Gayle Williams and Jana Krentz. Jefferson, N.C.: McFarland, 2019 [forthcoming]. Quotations are based on a pre-publication version of this chapter provided by the authors.

11. Results of the author'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.

12. 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.

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. A website followed in 1994—also almost prehistoric in the history of the Internet. Surely the highest profile digital project of the early years was the engagement of the University of Texas on an international effort to preserve, digitize, and make accessible the [Guatemalan National Police Historical Archive Project](#) (Archivo Histórico de la Policía Nacional, or AHPN), launched in 2011, based on more than eighty million pages of documents discovered in an abandoned Guatemala City barracks in 2005.<sup>13</sup>

Archiving born-digital web content, rather than digitized materials, at LLILAS Benson did not, however, begin as an extension of document digitization, but out of sheer necessity. Benson Library had systematically collected Latin American official government documents, including annual State of the Union reports as well as annual reports from individual government ministries. Beginning in the late 1990s, however, Latin American governments began releasing these documents only in digital form. 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, and becoming operational when LLILAS Benson enlisted Archive-It in 2005.<sup>14</sup>

Today, LAGDA comprises over a million discrete documents/files from approximately 300 ministries and presidencies in 18 Latin American and Caribbean countries. From a preservation perspective, a recent review showed that thousands of documents and speeches, which are available through LAGDA, no longer exist on the live web, including virtually the entire web presence of the Honduran government under Manuel Zelaya. In light of the recent election of Jair Bolsonaro as president of Brazil, LAGDA's importance as an archive of vulnerable government publications may once again be highlighted.

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.

The most significant of the other web archiving endeavors relevant to Latin America and the Caribbean actively maintained at the University of Texas is the Human Rights Documentation Initiative, or HRDI, which monitors, crawls, and archives ephemeral materials from the websites of human rights groups around the world. HRDI was founded in 2008, originally to preserve records documenting the genocide in Rwanda. Since then, its mandate has grown, especially regarding Latin America. As to whether there has ever been collaboration between the HRWA at Columbia and the HRDI at the University of Texas, David Bliss at Texas reported only that “an effort was made when compiling the initial seed list to avoid overlap with the HRWA.”

### “Post-Custodial” Archiving

In spring 2017, the University of Texas at Austin received a grant from The Andrew W. Mellon Foundation to fund a project titled “Cultivating a Latin American Post-Custodial Archival Praxis.”<sup>15</sup> The project focuses on building local capacity in Latin America to preserve vulnerable historical documentation, making the resulting documents digitally accessible. Building on earlier projects supporting the digitization of materials from Nicaragua, El Salvador, and Guatemala, the new grant

13. The AHPN represents “the largest single repository of documents ever made available to human rights investigators.” Around ten million pages were publicly accessible at the time of the launch. Kent Norsworthy, “Digital Resources: LLILAS Benson Latin American Studies and Collections, University of Texas at Austin,” in *Oxford Research Encyclopedia of Latin American History*, 2016, p. 9. <http://oxfordre.com/latinamericanhistory/view/10.1093/acrefore/9780199366439.001.0001/acrefore-9780199366439-e-81>. See also “The Archivo Historico de la Policia Nacional de Guatemala at the University of Texas,” *Focus on Global Resources*, Winter 2012, 31 (2) <https://www.crl.edu/focus/article/7499>

14. The successful proposal to The Andrew W. Mellon Foundation was submitted by CRL, 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” can be found at <http://www.crl.edu/reports/political-communications-web-archive>.

15. <https://legacy.lib.utexas.edu/benson/announcements/university-texas-austin-receives-mellon-foundation-grant-pioneer-archival>.

will support similar post-custodial initiatives with partners in Brazil, Colombia, and Mexico, with an emphasis on documenting underrepresented communities.

LLILAS Benson did not originate the “post-custodial” approach toward partner organizations, but it has embraced this paradigm wholeheartedly.<sup>16</sup> 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 . . .<sup>17</sup>

The post-custodial paradigm seeks to break through the colonial and post-colonial approach based on the acquisition or copying of cultural resources from their source communities—another form of resource extraction, in other words. The new paradigm was pioneered by HRDI with digitization projects done jointly with the Kigali Genocide Memorial Centre in Rwanda and the aforementioned Archivo Histórico de la Policía Nacional in Guatemala.<sup>18</sup>

In the field of web archiving, this approach involves 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.”

### In-house Use Analysis and User Feedback

As at Columbia University, the focus of web archiving at LAGDA and HRDI at present is building the archive. Use analysis is writ small: Google Analytics data is not maintained, and there is no archive of student use of the archived resources.

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.<sup>19</sup> 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.<sup>20</sup>

### Challenges and Future Plans

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 and resources to properly curate the 300 active seeds—and add new ones as existing seeds go dead. 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:

. . . 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

16. 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. See also Melissa Guy, “The ‘Post-Custodial’ Model for Preserving At-Risk Archives in Latin America,” presentation at CRL Global Collections Forum, May 18, 2018. <https://www.crl.edu/events/crl-global-resources-collections-forum-2018>.

17. Former director of the UT Libraries, 2016.

18. UTL’s Fred Heath, referring to the collaboration which brought the AHPN Digital Archive to Austin, noted: “. . . the cultural heritage of the [Guatemalan] nation will remain in country—a reversal of a century or more of ‘tail lights going north’ with national patrimony and a total volte-face in the way U.S. research universities are viewed by nations to our south. Now we just have to prove ourselves worthy of their trust.” Quoted in *Focus on Global Resources*, 2012.

19. <https://archive.org/details/ArchiveIt-Collection-176&tab=about>.

20. <https://archive.org/details/ArchiveIt-Collection-1475&tab=about>.



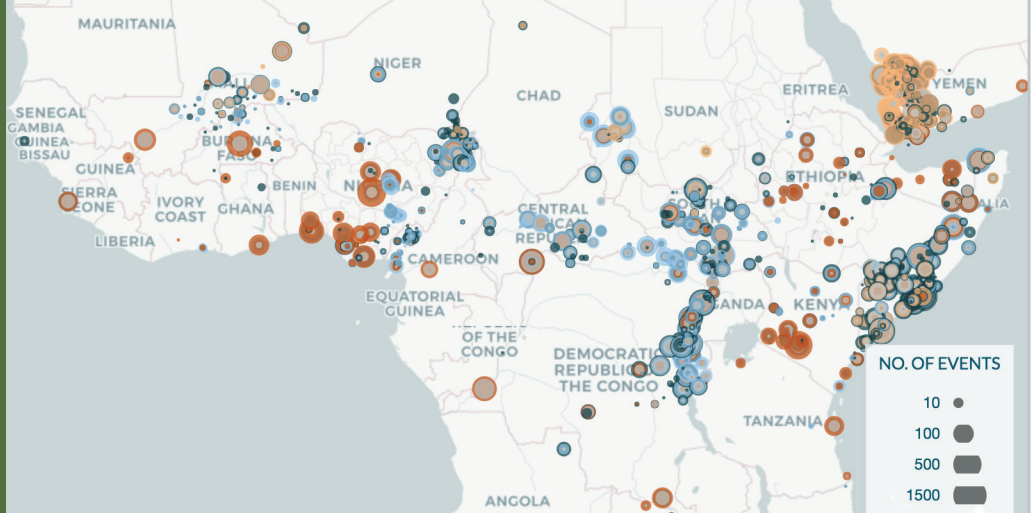
scholar-led archiving efforts inform more systematic and comprehensive collection building carried out by libraries?

Ideally, then, such collaboration would make researcher “scrapbooking” largely unnecessary.

## **Conclusion**

The 2018 CRL report identified measures that archiving efforts can take to become more useful to researchers and therefore more sustainable. These include standardizing metadata across the library/archives divide; developing better finding aids and exposing them to web crawlers; and introducing certification standards to enhance the credibility of archived web content among skeptical scholars. Education and outreach at discipline-specific professional meetings will be useful. Ultimately inter-institutional—and international—collaboration will be necessary, to leverage the strengths of multiple library, archival, and publishing partners in validating and preserving information distributed on the web. ❖

# Born-digital Primary Sources for Area and International Studies: New Models and New Threats



From ACLED (Armed Conflict Location & Event Data Project): [dashboard](#) mapping the number of violent events in Africa and elsewhere.

**Bernard F. Reilly**

*President*

*Center for Research Libraries*

## The Global Collections Initiative

The chief goal of the Global Collections Initiative is to expand electronic access to primary source documentation and data for scholarly research on parts of the world like the Middle East, sub-Saharan Africa, and South Asia, areas where the information landscape differs from that in the U.S. and Western Europe. A major focus of the initiative is access to materials existing only in digital form. Many of these materials are available openly on the web and on social media platforms; other digital-only materials are available through commercial databases and services. While established mechanisms and practices are in place for preserving tangible materials like print books and journals, for which relatively uniform and stable publishing and distribution models exist, production and supply chain realities in the digital realm present new challenges for libraries.

Jeffrey Garrett's article, "Archiving the Latin American and Caribbean Web," outlines the prevailing approach taken by major libraries to preserving open source web based materials. That approach involves harvesting html files and other digital content and metadata from websites, and storing that data independently in separate, controlled digital environments. In his article Garrett points out some of the limitations of that approach. The efforts are particularly daunting when considering preserving complex content like databases and video and highly dynamic news and social media sites and materials maintained behind paywalls and other barriers by commercial producers. If the primary goal is to ensure the long-term accessibility and integrity of web materials, approaches other than web harvesting might serve libraries and scholars better.

Given the overwhelming amount of open web content possibly relevant to area and international studies research, where does one begin? Historically research libraries have given priority to documentation and data useful to the widest range of humanities and social sciences: information essential to the understanding of major actors and forces in society, like news, the documents and archives of governments and NGOs, economic, financial and geospatial data, public opinion and population information. In their preservation triage libraries have also prioritized materials most likely to disappear absent their efforts.

## New Threats and Challenges

Open source materials are particularly susceptible to corruption or loss through reliance upon highly fluid and complex technologies, loss or suppression by hostile political interests, and the growing privatization of data.

The technology threats are well known. The use of non-standard or proprietary formats and applications, unstable platforms and the potentially obsolescent media undermine the persistence and integrity of much open web content. Sites hosted by individuals and small, non-governmental organizations with limited means or technical sophistication are particularly vulnerable.

Recent political trends also pose threats. Censorship and suppression of news and other third-party documentation of illicit government and criminal activity have become more common with the rise of authoritarian regimes. As powers like China, Russia and Saudi Arabia gain economic and political influence, liberal norms for access to public interest information are likely to erode. Moreover, information that was once widely available to the public is rapidly being privatized. The new value of data to the corporate and national security sectors is driving monetization of global news as well as financial, legal, population, and environmental data, eroding the public information domain. According to the *Journalism in the Americas* blog, paywalls are becoming the norm for the major newspapers in Mexico, Brazil, Argentina and elsewhere in Latin America, as publishers turn to digital subscription rather than advertising as their primary source of revenue. Even data traditionally provided by governments, like census and statistical information and legislative proceedings, is now packaged and productized by commercial vendors, especially where poorly resourced governments are unable to equip their open platforms with the robust functionality modern researchers expect.

An important and particularly vulnerable category of web content is the product of a new phenomenon that has emerged in the information landscape: large web-based repositories of digital documentation. In 2010, with its release of Afghanistan and Iraq War video and text documents leaked by the U.S. military, WikiLeaks created a template that has since become widely adopted by grass-roots activists, journalists, policy researchers, and other civil society actors: collecting and curating critical and often highly sensitive evidence of human rights violations, government corruption, and environmental crime, and exposing that documentation on the web. The most widely known of such groups is the [International Consortium of Investigative Journalists \(ICIJ\)](#), which in 2016 broke the Panama Papers scandal using a trove of over 11 million pages of leaked legal and financial documents to reveal the offshore and illegal banking activities of thousands of world leaders, public figures and corporations.

Even more endangered is online documentation hosted by small news organizations operating in conflict zones and transitional societies. [Zaman al-Wasl](#), an online news agency established in Homs, Syria, in 2005, has collected and exposed documents regarding the operations of ISIS, the Syrian government, and Western powers. In 2018 it published a leaked archive of 1.7 million documents on disappeared and victims of arbitrary arrest by the Assad regime, obtained from the Syrian intelligence service.

Other types of organizations harvest data on global conflicts, epidemics, trade, and other subjects from social media and the open web, and generate databases and visualizations that serve a variety of research communities, such as legal scholars, jurists, and policy researchers. Since 1997 the *Armed Conflict Location & Event Data Project (ACLED)*, based at the University of Sussex, has been aggregating and collating data from online news sources on violence and conflict in sixty countries in Africa, South Asia, South East Asia, and the Middle East. ACLED provides informative visualization and analysis of the conflicts on an almost real-time basis.

Relevant to public policy in the present; such documentation will be vital to humanities and social science researchers in the long term. The organizations that preserve such data are not libraries or archives per se, but in effect have taken responsibility for stewardship of important evidence. While they have a vested interest in the survival of the data they gather, at least in the near term, many of them operate on marginal funding, deploy obsolescent technologies, and struggle to survive.

### Some Alternative Approaches

All told, online documentation is being created and exposed in ways that elude capture through conventional web-archiving means. Harvesters cannot penetrate the commercial paywalls and other obstacles erected by activists to protect proprietary and sensitive content. And web-crawling using the standard harvesters proceeds too slowly to capture the dynamic visualizations and other resources like ACLED that pull data from social media and the live web in real time.

This suggests a need for libraries to work further upstream in the information “supply chain”. Support for new stewardship efforts like ICIJ and ACLED could go a long way toward ensuring the survival of critical web content. Rather than independently capturing and archiving their web content, it may be better to engage the producers in shaping a more serviceable and durable product. It is not clear how and where best to apply such support, as more research is necessary. Some common needs evident, however, include guidance on secure, non-proprietary platforms and hosting arrangements; best practices to promote discovery, interoperability, and scholarly mining of their content; and legal risk assessment and indemnification for dealing with sensitive materials.

Similarly, collective dealings with the commercial sources of digital data and documentation might mitigate some of the worst effects of paywalls and other barriers to scholarly access. Acting in unison, the academic library community could exploit the robust capabilities the news industry has built to manage and mine the enormous archives of electronic text, audio, moving image, and datasets they acquire and produce. The digital asset management systems (DAMs) used enterprise-wide by organizations like Associated Press and El Pais have powerful repository features. Perhaps through the instrument of national site licenses, the news media organizations could be persuaded to harness those capabilities to provide enduring electronic access tailored to scholarly practice.

The web offers an almost infinite source of global data and documentation relevant to humanities and social science research. This apparent surfeit makes the task of finding “the signal in the noise” at once difficult and imperative. Therefore libraries should invest locally and collectively in building the analytical capabilities needed to locate value and redundancy in commercial databases and open access resources alike. Landscape analysis, and auditing of the major data producers and repositories could provide intelligence for decision-making. As the onus of preservation falls on more and more organizations outside the traditional library orbit, transparency will also become more crucial. To identify the high-value targets and priorities, libraries will need a stronger base of knowledge than exists currently. ❖



**FOCUS on Global**

**Resources**, published semi-annually, is compiled by CRL's Communications Department. Virginia Kerr and Gloria Johnson, Editors. Graphic design services provided by Molly O'Halloran, Inc.

**ISSN #: 0275-4924**

**Center for Research Libraries Staff Contacts  
(800) 621-6044**

President  
Bernard F. Reilly x 334  
breilly@crl.edu

Administrative Services Specialist  
Yvonne Jefferson x 319  
yjefferso@crl.edu

Member Liaison and Outreach  
Services Director  
Mary Wilke x 351  
mwilke@crl.edu

Vice President, Collections  
and Services  
James Simon x 324  
jsimon@crl.edu

Director of Technical Services  
Amy Wood x 327  
awood@crl.edu

Director of Information Systems  
Patricia Xia x 341  
pxia@crl.edu

Head, Access Services  
Kevin Wilks x 314  
kwilks@crl.edu

Head, Stack Management  
Bethany Bates x 339  
bbates@crl.edu

Head of Communications and  
Development  
Virginia Kerr x 265  
vkerr@crl.edu

Communications Coordinator  
Gloria Johnson x 289  
gjohnson@crl.edu

**Global Resources Program Contacts  
(800) 621-6044**

Director  
James Simon x 324  
jsimon@crl.edu

Global Resources Network and  
AMPs Program Manager  
Judy Alspach x 323  
jalspach@crl.edu

News Database Analyst  
Maria Smith x 322  
msmith@crl.edu