May 18, 2018

Harvesting Web Content and Data from Emerging Regions: a Status Report

Jeffrey Garrett
Independent Consultant
Expansion of the Web in Latin America during 2007–2011

Why the Web Matters—especially in International and Area Studies, even more especially in Latin American and Caribbean Studies

- Harrison’s mapping “only reflects density of connections, and not usage—hundreds of people may utilize a single connection in an internet cafe, often the only form of connectivity people have access to in developing nations.”

- Of 3.9 billion Internet users in the world (as of June 30, 2017), 10.4% reside in Latin America and the Caribbean, while 8.2% are in North America. Almost half—49.7%—are in Asia.

- “. . . in Latin American Studies, . . . much digital publishing is not channeled or distributed through commercial publishers but is instead only taking place on the freely accessible web . . .”

Sources: Chris Harrison (2017); Internet World Stats: Usage and Population Statistics (2017); Graham & Norsworthy (2018)
Research & Citation Practice 20 Years Ago


Contraband Corridor: Making a Living at the Mexico-Guatemala Border. Galemba, Rebecca B. Stanford, Stanford University Press, 2018

Research & Citation Practice Today


New Free Trade Agreement With Mexico Takes Effect in Guatemala

Guatemala has, as of yesterday, adopted all terms of its Free Trade Agreement (FTA) with Mexico. The new FTA includes Honduras and Costa Rica, which have already been trading under these terms with the northern neighbor. El Salvador and Nicaragua are also parties to the agreement but have yet to receive full implementation from their legislatures.

The FTA negotiations ended on October 20, 2011, and those involved gave contingent signatures on November 22, 2011, in San Salvador, El Salvador. The Senate of Mexico then gave approval on December 15, 2011, and since then the congresses of all the Central American countries have been
Contraband Corridor: Making a Living at the Mexico-Guatemala Border. Galemba, Rebecca B. Stanford, Stanford University Press, 2018

Research & Citation Practice Today


17 of the 120 Links to Web Content in *Contraband Corridor*
(Stanford UP, 2018)

http://foreignpolicy.com/2016/02/23/obama-pena-nieto-mexico-corruption/
http://nofrackingmexico.org/nueva-ley-para-criminalizar-la-protesta-social-y-limitar-el-libre-flujo-de-informacion-en-el-marco-de-las-reformas-estructurales/
www.animalpolitico.com/2013/11/hacienda-cierra-12-garitas-aduanales-en-4-estados-fronterizos/
http://rightsaction.org/sites/default/files/Rpt_130220_Aguan_Final.pdf
www.prensalibre.com/economia/Conflicto-ingreso-maiz-blanco-Mexico_0_230976910.html
www.cipamericas.org/es/archives/15407
http://fpif.org/mexicos-oil-privatization-risky-business/
www.cipamericas.org/archives/1834
www.migrationpolicy.org/article/mexico-caught-between-united-states-and-central-america
www.prensalibre.com/pl/2006/septiembre/07/lectura_dept.html#151072  (link no longer available at time of publication)
https://loschapincitos.wordpress.com/2013/07/21/en-guatemala-cae-hermano-de-jefe-narco-que-ordeno-la-matanza-de-9-policias/
www.conasami.gob.mx/formatestimonios.aspx  (link no longer available at time of publication)
Advice from Style Manuals I: MLA Handbook

- “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.”

*MLA Handbook, 8th edition (2016), p. 48*
MLA Handbook, 8th ed.: Criticism

• This 8th ed. of the *MLA Handbook*, published in 2016, shows no awareness of the existence of archival Web content.

• Its advice is to depart from an author’s responsibility to the reader: to research and document one’s sources.

• Jill Lepore: “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)
Advice from Style Manuals II: Chicago 17th ed.

▪ “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 URL, give interested readers a chance to find the information through the Internet Archive or other means.” (2017, 14.207)

▪ I.e, no recommendation to cite the archive as the actual source or as a backup and more persistent source.
17 of the 120 Links to Web Content in *Contraband Corridor*
(Stanford UP, 2018)

http://foreignpolicy.com/2016/02/23/obama-pena-nieto-mexico-corruption/
http://nofrackingmexico.org/nueva-ley-para-criminalizar-la-protesta-social-y-limitar-el-libre-flujo-de-informacion-en-el-marco-de-las-reformas-estructurales/
www.animalpolitico.com/2013/11/hacienda-cierra-12-garitas-duanales-en-4-estados-fronterizos/
http://rightsaction.org/sites/default/files/Rpt_130220_Aguan_Final.pdf
www.prensalibre.com/economia/Conflicto-ingreso-maiz-blanco-Mexico_0_230976910.html
www.cipamericas.org/es/archives/15407
http://fpif.org/mexicos-oil-privatization-risky-business/
www.cipamericas.org/archives/1834
www.migrationpolicy.org/article/mexico-caught-between-united-states-and-central-america
www.prensalibre.com/pl/2006/septiembre/07/lectura_dept.html#151072  (link no longer available at time of publication)
https://loschapincitos.wordpress.com/2013/07/21/en-guatemala-cae-hermano-de-jefe-narco-que-ordeno-la-matanza-de-9-policias/
www.conasami.gob.mx/formateestimaciones.aspx  (link no longer available at time of publication)
Your connection is not private

Attackers might be trying to steal your information from www.cipamericas.org (for example, passwords, messages, or credit cards). Learn more.
NET::ERR_CERT_COMMON_NAME_INVALID

Automatically send some system information and page content to Google to help detect dangerous apps and sites. Privacy policy.
Surprise! “cipamericas.org/es/archives/” is now a site advertising cannabis oil products . . .
So I called the CIP (Center for International Policy) in Washington, D.C., and asked where their content had gone and what they were doing to get it (and their “archival” website) back . . .
Content Wrangling in the Wild West of the Web

- CIP said that after their site had been hijacked/corrupted, they had abandoned “cipamericas.org” and moved their archive along with other compromised subdomains to “americas.org”.

- As documented by a review of crawls of this domain performed by the Internet Archive, “americas.org” had been given up in 2007 by the Resource Center of the Americas in Minneapolis, then used briefly by La Conexión de las Américas before taken over by CIP.

- I found the referenced article at the new location.

- I notified the author at the University of Denver. She was not happy—but what could she do? The book had already been published.

- I decided to archive the content at its new location . . .
Self-Archiving Your Sources Is Fun and Easy!

Hrm.

Wayback Machine doesn't have that page archived.
Want to search for all archived pages under
http://www.americas.org/es/archives/15407?

This page is available on the web!
Help make the Wayback Machine more complete!
Save this url in the Wayback Machine
INTERNET ARCHIVE
WayBack Machine
Explore more than 327 billion web pages saved over time
http://
BROWSE HISTORY

Is the Wayback Machine useful to you? Then keep it growing. Your gift matched today! DONATE

Tools
Wayback Machine Availability API
Build your own tools.
WordPress Broken Link Checker
Banish broken links from your blog.
404 Handler for Webmasters

Subscription Service
Archive-It enables you to capture, manage and search collections of digital content without any technical expertise or hosting facilities. Visit Archive-It to build and browse the collections.

Save Page Now
Capture a web page as it appears now for use as a trusted citation in the future.
Only available for sites that allow crawlers.
Deportación, detención ilegal y abusos en la frontera de EEUU con Guatemala
By Laura Carlsen | 7/julio/2015

La frontera sur de México se ha vuelto la línea de contención del país más poderoso del mundo. Las víctimas de esta política extraterritorial son las personas migrantes de Centroamérica que cruzan cada día, buscando salvar su vida de la violencia y el hambre que azotan sus países.

Desde hace años, y de manera intensificada en el último año, el gobierno de los Estados Unidos ha reclamado la frontera sur de
And good that I did, because when I returned to the live site on May 3, 2018, this is what I found . . .
So . . .

• . . . as an author or a student . . . why not just cite the stable, archived URL in the first place??

• And if there isn’t one, why not make one yourself?? (Go ahead, be your own archivist!)

• It could save your readers inconvenience . . . and yourself embarrassment!

• Why aren’t style manuals, why aren’t we giving this advice to our communities??
The Way It Could Be Done (I)

An explicit link to LAGDA content (University of Texas at Austin) in a recent monograph (Williams 2012, 140), retrieved by searching for instances of http://wayback.archive-it.org/176/ in Google Books.
Fraser-Pryce, Shelly-Ann (1986–), track and field athlete, world champion, was born in Kingston, Jamaica, on 27 December 1986. Shelly-Ann and her brother were raised in a one-bedroom tenement house in Waterhouse, an impoverished, overcrowded neighborhood. Her mother, Maxine Simpson, worked as a street vendor, while her father, a master builder, who lived in the States, rarely came home. Her extended family, had very little to do with them.

First participating in track and field at George Headley Primary School, Fraser-Pryce won the Primary School Championships at the age of 10. After completing primary school, she qualified scholastically to

Bibliography


Reliably archived, but note missing content!

New tools allow the capture of more content types.
The Way It Could Be Done (III)

Senior Honors Thesis, University of Maryland (2013), citing content from the Latin American Government Documents Archive (LAGDA) at the University of Texas at Austin.

scholar Yaakov Ariel says that Latin American Christian Zionism is part of a global movement that is growing in new places like Korea and joining places traditionally supportive of Christian Zionism like the United States.36

• “UMD librarian Pat Herron was very helpful in this process. Her demonstrations in our LASC 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.”

• “At the end of this process, I very much feel like a historian who is able to carry out independent research, and feeling like I really have those skills is a great end to my undergraduate career at UMD.”

Sandra A. Shaker, Student at UMD (2013)
AN EVALUATION OF WEB ARCHIVING PROGRAMS IN THE US RELEVANT TO INTERNATIONAL AND AREA STUDIES

The Example of Latin America and the Caribbean
Parts I, II, and III:

An introduction to web archiving as it relates to IAS—especially to Latin America & the Caribbean . . .

. . . and as practiced at 3 leading US programs:

Library of Congress
Columbia University
University of Texas
Part IV: An Independent Use Analysis

IV. An Independent Use Analysis

Use analysis that goes beyond the anecdotal—in other words, that is based on harvested occurrences of unique and explicit URL stems in large databases—is problematic for several reasons. For one, researchers who cite websites may choose to cite original, live-web URLs even if they used the archived content. Live web URLs are 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 to specific institutional archives if they can’t locate/relocate the original content. Contributing to this practice may be that the 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 either a lack of awareness of the existence of these archives or lacking community consensus on how to treat them in scholarly work. 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,
Part V: Accelerating the Integration & Mainstreaming of IAS-Relevant Web Archives

1. Greater Interinstitutional Collaboration
2. “Postcustodial” Collaboration with Partners Abroad
3. “Desiloization” within Libraries, Archives, & Communities Served
4. Metadata Standardization
5. Improved Citation Standards & Tools
6. Development and Promotion of Self-Archiving Tools
7. Credibility Enhancement through Clearer Scoping & Certification
8. Outreach to Publishers, Associations, Funding Bodies
9. Ramp Up Outreach on Campus through Public Services
10. Promoting Opportunities for Whole-Collection Analyses & Datamining
In Closing: The Problems Are Not So Much About Technology Anymore

- “We can assert that the primary obstacles to expanding [web archiving] activities in libraries are less on the “technology” side and more on the ‘cultural’ side.”  
  (Graham & Norsworthy, 2018)

- “I . . . 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 . . .”  
  (Pamela Graham, personal communication, 2018)