Long-Term Preservation for Fragile Web Content: LANIC's Web Archiving Program

by Kent Norsworthy

THERE'S A GROWING PROBLEM FACED by many Latin Americanist scholars: you have bookmarked a number of key Web sites in Latin America that include valuable data—it could be text, statistics, a news item, video, or comments on a blog entry—that is related to your research. You visit the sites periodically and begin to cite some of the material in a paper you are writing; some of the sources are just references, but others have content that is integral to your analysis. You send a draft of the paper to a colleague for comments, but she writes back: "I couldn't really follow your analysis because several of the URLs you cite give me a '404: Page not Found' error message." You attempt to relocate the content by poking around on the original Web site or googling the material itself, but to no avail. The content appears to have been completely removed from the Web.

At root, the problem here is a simple one: the estimated life expectancy of a Web page is between forty-four days and two years. Content can disappear from the Web for any number of reasons, some benign, others quite malicious, but at root, the fact is most organizations that publish and host Web content have no interest in long-term preservation. Compounding the problem, the Latin American Web is typically much more volatile than the Web at large, due to technical and resource issues, as well as to the prevailing culture of information. The ephemeral nature of content on the Web represents a problem in all domains of knowledge: from biology to art, classics to statistics, government documents to rich descriptions of daily life, researchers from any area or discipline can be confounded by this problem.

Within Latin American studies, LLILAS has taken a leading role in working on solutions to the vexing problem of disappearing Web content. Since 2003 the institute's Latin American Network Information Center (LANIC) has been involved in efforts to shape the emerging field of Web archiving, which aims to capture or take a snapshot of a live Web site and then provide for long-term preservation and public access for this captured version, regardless of subsequent changes to, or removal of, the original Web site.

LANIC's efforts in this area essentially have attempted to marshal the resources of a global leader in Web archiving, the San Francisco–based Internet Archive, to the field of Latin American studies. Since 1966 the Internet Archive—which has the lofty mission of providing universal access to human knowledge—has been building the world's largest Internet library, designed to provide permanent access for researchers, historians, and scholars to historical collections that exist in digital format. In a nutshell, the Archive seeks to transform the content of the Internet from ephemera to enduring artifact. While the Internet

Archive collects various types of digital content, including text, audio, moving images, and software, it is perhaps best known for its collection of archived Web content, with over 150 billion Web pages collected to date. The collection can be consulted on the Web through the Internet Archive's Wayback Machine at http://www.archive.org/

While the Wayback Machine is an invaluable tool for consulting archived Web content, and indeed it should be every researcher's first stop when confronted with the problem of Web content that has vanished, it does have some limitations as a tool for research. For one, while the collection is massive in scope, there are still many Web sites that are not included, nor is there any provision for suggesting or ordering inclusion of a site that is currently not in the collection. In addition, the Wayback Machine does not allow for full-text searching: in order to consult an archived site, you must type in the original URL.

The Internet Archive sought to address these and other limitations with the launch of their Archive-It service in 2005. Archive-It is designed to meet the needs of memory institutions—state archives and libraries, universities, historical societies, etc.—that are interested in creating and archiving their own Web collections. Archive-It allows subscribing institutions to determine exactly which Web sites they want to archive and at what frequency. The resulting collections are full-text searchable and free and open to the public via the Web. Archive-It currently has more than 850 collections built by over 70 subscribing institutions, including the Library of Congress, the Biblioteca Nacional de Chile, the U.S. Department of Energy, and numerous universities and state archives and libraries.

LANIC has been affiliated with Archive-It since its inception four years ago. LANIC's Archive-It project, which is conducted as a partnership with the UT Libraries and the Benson Latin American Collection, is focused on the Latin American Government Documents Archive (LAGDA). The LAGDA collection seeks to apply Web archiving technology to a specific collecting area: born-digital Latin American government documents.

The project grew out of a collecting challenge faced by the Benson Collection. Historically, the 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. Traditionally, such reports were published and collected in print format. But beginning in the late 1990s, increasing numbers of Latin American government entities ceased paper publication of these official documents and reports, opting instead to publish them in digital format directly to the Web.

Initially, the Benson Library hoped to be able to link directly to these born-digital versions of the reports on the Web and to integrate those links to the existing library catalog records for these serial publications. However, they soon ran into the problem described above: namely, while creating the initial set of links was fairly straightforward, over time the number of reports remaining online at the original address declined significantly as part of the process known as "link rot." Typically, when a new annual report or State of the Union is produced and uploaded, the publishing entity deletes the previous year's version from the Web site. Critical gaps in the coverage of these reports historically provided by the Benson, in some cases stretching back to the nineteenth century, began to appear.

LAGDA was launched with a view toward plugging these gaps by providing for systematic capture of Latin American government Web sites. Benson staff identified close to three hundred key sites from eighteen countries in Latin America and the Caribbean, primarily government ministries and presidential sites, where such documents were published to the Web. Since 2005, the LAGDA project has used Archive-It to gather, four times per year, the entire contents of these Web sites. The resulting collection today totals nearly 44 million URLs, or discrete documents/files, amounting to almost four terabytes of data. Users can consult the archived sites in two different ways, both available through the LAGDA Web site at http://lanic.utexas.edu/project/lagda/ One, they can conduct a full-text search across all contents of the archive, where the search result list consists of direct links to the archived content. Two, they can browse the archived content starting at a list of links, ordered by country, of the nearly three hundred ministries and presidencies targeted by the Benson Collection.

LANIC plays a key role in the LAGDA partnership, both coordinating overall project activities and taking lead responsibility for managing the Archive-It subscription. This includes precrawl tasks, using the application to manage the seed list and configure the Web crawl settings, as well as postcrawl tasks, such as reviewing crawl reports and applying a quality control protocol to the archived sites. LANIC staff also use the Archive-It application to manage metadata associated with each archived Web site.

A systematic review conducted by LANIC has confirmed that LAGDA contains thousands of official documents and speeches from Latin American governments that have long since disappeared from the live Web, including not only text documents, but also audio and video files. In addition to the annual reports and state of the union addresses mentioned above, these include large numbers of speeches delivered by Latin American presidents and their cabinet ministers, sectoral reports, economic indicators, survey results, and other data gathered by government entities. Another advantage LAGDA provides for researchers is that all documents and other types of Web content are preserved in their full original context, that is, the entire Web site where such documents were originally housed.

In addition to the LAGDA effort, LANIC and the UT Libraries also are using Archive-It to capture and preserve other types of Latin American Web content. Two earlier collections cover Venezuelan political discourse and Latin American political parties and elections, while a current and ongoing collection covers "Mexico 2010," a collection of Web sites launched in conjunction with Mexico's commemoration of the two hundredth anniversary of independence and one hundredth anniversary of the Mexican Revolution in the year 2010. The UT Libraries has a separate project underway using Archive-It and other tools to archive the Web sites of human rights groups from around the world, including Latin America.

We invite you to take a step back in time by visiting some of the sites archived through LAGDA at the address above. We also encourage you to help us preserve these valuable resources by sending us your feedback, including suggestions regarding other types of Latin American Web content that you think would be important to archive. *http://lanic.utexas.edu*

Kent Norsworthy is LANIC Content Director.