

FOCUS *On Global Resources*

Winter 2008–09 • Volume 28 • Number 2

NEWS PRESERVATION

**“On the Record”—
A Forum on Electronic
Media and the
Preservation of News** 3

**World Newspaper
Archive Update** 14



In This Issue

This issue of *FOCUS* reports on the recent Global Resources Forum “On the Record”—A Forum on Electronic Media and the Preservation of News. The Forum was held in October at the New York Public Library Humanities and Social Sciences Library. It brought together individuals from many fields—journalism and the news media, academia, business, and policy research—with those who map library strategies for collecting and preservation, to explore the ways in which today’s news content is created, distributed, and consumed. Attendees heard presentations from distinguished experts and engaged in focused discussions to formulate a library agenda for news preservation and access.

The Global Resources forums are important agenda-building events for CRL. “On the Record” helped us identify and understand the new challenges that CRL and the libraries in its community face in the collecting and preservation

On page 1: The November 16, 1889 cover of *A Provincia de Sao Paulo* announces the founding of the Republic of Brazil.

continued from page 1

of news reporting. We here summarize many of the ideas raised in the Forum, and outline an agenda for meeting the challenges identified.

—*Bernard F. Reilly, Jr.*
President

“On the Record”— A Forum on Electronic Media and the Preservation of News



The death of Zapata, Mexico's revolutionary hero, is covered in the April 11, 1919 edition of *Excelsior* (Mexico City).

Keynote: The Future of Newspapers: A Conversation

Alex Jones

Laurence M. Lombard Lecturer in the Press and Public Policy and Director of the Shorenstein Center, Harvard University

and

John Carroll

Former Editor, *Los Angeles Times*

Summary: The recent and ongoing contraction in the newspaper industry and the emergence of the Internet as a venue for real-time news suggest that preserving newspapers will become less effective as a strategy for capturing and preserving the news record as time goes on.

Some of the observations made were:

- *More versioning, but less reporting.* The prevailing newspaper business model has been shattered by the Internet, with the sharp decline of newspaper classified advertising revenue. Yet the digital media outlets that are purported to be replacing newspapers tend not to “put reporters on the street” or support sustained investigative reporting of the kind that broke stories like Watergate, Whitewater, and others in the past. Original news gathering and investigative reporting are expensive, and newspaper organizations have been scaling back their investment in same, while new media, such as blogs, feeds, and news Web sites, largely draw on the traditional news media sources. As a result, while there are more sources from which people, particularly young people, get their news, the amount of substantive news coverage has been decreasing.
- *Pressures on bottom-line performance.* Newspapers, and news people, are themselves committed to a “public service” mission, of promoting an informed citizenry, but today are finding it difficult to reconcile that mission with the relentless pressure from owners to make money. This pressure is particularly intense at newspapers owned by stockholders. Online ad revenue alone is not yet adequate to sustain a paper, but because most expect Web content to be free, it will be difficult to succeed with a business model under which readers must pay for content.
- *Declining consumption.* Newspapers are also facing the reality that few people are interested in serious news and even fewer are willing to wade through lengthy text and analysis. Demand for instantaneous news outweighs the demand for authoritative reporting, and the former demand is largely met by cable and the Web.

- *Investigative reporting* is still alive, albeit on a limited basis, at non-profit centers like the investigative journalism nonprofit, [ProPublica](#), based at the University of California at Berkeley, or supported by grants from foundations like the Fund for Investigative Journalism. On another independent news site, Josh Marshall's [Talking Points Memo](#), “. . . even if he doesn't report it—and he often does—he also keeps big media on the story, and it's a great new pressure.”

Session I—Archiving and Access to News

Challenges in Making the World's Newspapers Accessible: A Status Report on the Digitization of News Back Files

Denise Hibay

Interim Director for Collections Strategy

The New York Public Library

Summary: In the past, to ensure the long-term availability of news back files for scholars and the general public, the major research libraries have collected and preserved newspapers in hard copy and microform. Economic pressures and the ready availability of news on the Web made support for this kind of pro bono activity by libraries hard to justify.

In recent years many libraries have embraced digitization as a way to provide access to the newspapers they preserve, but their efforts are hampered by a scarcity of resources, standards, and holdings information and by copyright restrictions. Commercial vendors also provide valuable content in digital form, but the cost of those digital resources is often an issue for libraries. Greater transparency of digitization efforts and access to granular, reliable information about the major research libraries' holdings of newspapers in microform and paper are necessary to prevent the duplication of efforts, unnecessary spending, and the loss of unique materials.

Some points raised in the presentation:

- *Diminishing means and high costs.* As libraries experience budget reductions, microfilming of newspapers is often eliminated. With more online content and the ability to print from electronic resources on demand, a decreasing number of libraries are continuing to subscribe to newspapers in hard copy. Those libraries still subscribing to and collecting current print newspapers often must retire back files to remote storage, where they are largely unavailable to researchers.
- *Microform is still useful.* At the same time, the New York Public Library still sees significant use of newspapers on microfilm, although the number of researchers willing to use materials in micro-format is far lower today than it was a decade ago.
- *Rising user expectations.* The current needs for news content are rapidly changing and becoming increasingly user focused. Users want more content delivered electronically. (NYPL sees heavy use of digital back file collections like Readex's *America's Historic Newspapers*.)
- *High costs of digitization.* A large amount of funding is necessary—to maintain equipment, software, hire staff with the appropriate technical skills or train existing staff. Licensing issues are also prevalent, raising the questions: Who has control of the content? Should choices be made between access and ownership?
- *Some solutions are showing promise.* Examples include the National Digital Newspaper Program (NDNP) and [Utah Digital Newspapers \(UDN\)](#). Individual institutional strategies are also showing early success rates, such as the [California Digital Newspaper Collection](#) at the University of California Riverside.
- *Library partnerships with the commercial sector may provide solutions.* A num-

ber of libraries are working with commercial electronic publishers like iArchives, ProQuest, and Cengage Gale to digitize newspaper back files. For example, the British Library collaborated with Cengage Gale to digitize British Nineteenth-Century Newspapers. These commercial collaborations, however, can be compromised by differences in the respective missions of the partners, and a consequent lack of transparency. Many commercial contracts prohibit disclosure of terms of the collaborations, which is difficult for libraries to reconcile with their obligations to their communities.

If You Build It Will They Come? One Library's Experience with Converging News Content

Debora Cheney

*Larry and Ellen Foster Communications Librarian, and Head, News and Microforms Library
The University Libraries, Penn State University*

Summary: News collections are migrating from physical to digital formats, reflecting the current state of the news industry and new user behaviors. While libraries are not the largest players in the market for electronic news, they can impact this market favorably. They can control their costs and improve and expand the newspaper content available to their researchers in electronic form by being discerning in their acquisition of digital resources, and by collaborating closely with other organizations that have an interest in maintaining news content, such as other libraries, the news industry, and database vendors.

Therefore, again, informed library collection development and preservation require access to granular, reliable metadata about the contents of electronic news collections and digital libraries.

Some points raised in the presentation:

- *The traditional place of libraries in the “supply chain” for current news is changing.* One of the roles that North American academic and other research libraries have traditionally played is to provide access to recent issues of major newspapers from Europe and various world regions. As many of these news organizations now provide Web access to their current news reporting, the value of this library service is diminished.
- *Hard copy newspaper back file collections are disappearing.* The “just in case” scenario has become unaffordable for libraries. With the reduced demand for paper copies for current news awareness, and relying almost entirely on electronic products for back files, libraries have little incentive for continuing to subscribe to newspapers in print and microfilm. Many librarians assume that what they cannot preserve will be available from other institutions or obtained from other sources when and if needed. Microfilm is perceived as an archaic medium.
- *Investment in news content is nearly unsustainable with current models.* There is currently too much duplication of content at too high a cost in the electronic news packages. Unfortunately, libraries are not collecting or preserving as complete a collection as they think, leaving many librarians unaware of how incomplete content is. Much of the same content is purchased in different formats.
- *Yesterday's news is fodder for a wider range of disciplines than ever.* Changes in teaching and research methods have created a greater demand for newspaper back files and other news content. History, political science, sociology, gender studies. One Penn State researcher analyzed the placement of articles about modern Argentine women politicians in domestic newspapers. Advertising content in newspapers is heavily used by students and researchers.

- *User expectations are high and growing.* The user motto is, “Bigger, better, more.” Researchers expect access to non-current news in a wider range of formats, including radio and television broadcasts, Web sites, and blogs. Libraries need to find ways to provide access to digital and broadcast content, and news is no exception.

Google: Search Discovery and Web Access to the World’s News

Andrew Madden

Director of Strategic Partnerships

Google

Summary: In 2008 Google created *Google News Archive Search* to enable users to search news items in “historical archives,” *i.e.*, Web news items more than 30 days old, and the back issues of newspapers published by Google’s News Archive Partners. Search results link directly to the news publishers’ sites, or to content hosted for those partners by Google. The goal is to generate activity through the Google search engine and drive traffic to publishers’ content.

Some points raised in the presentation:

Google News Archive Search provides an ostensibly viable model for leveraging rights to make proprietary content available on the Web, and will probably digitize extensive back runs of newspapers. However, relatively few newspaper publishers are yet on board. Because of this Google will provide access to only a small portion of the world’s newspapers for the time being.

To guide library investments in their own digitizing activities and their purchase of digital back files from vendors like Readex, Cengage Gale, Factiva, and others, it will be important to monitor Google content-building to know what newspaper content is actually available.

- *Previously Google news searches led to the Web sites of various media*, including television, radio, and newspaper sites. Those search results included news that was as much as 30 days old. (Older materials were rolled into regular Google searches.)
- *For News Archive Search, Google is now digitizing newspaper back files from microfilm held by publishers and commercial vendors.* Many newspaper publishers had microfilm in their archives and asked Google to provide digital conversion and hosting of this content. Google is also digitizing from titles held by ProQuest and Heritage Microfilm.
- *Google digitizes, indexes, and hosts this content without charge.* AdSense revenue for advertising on the results page is shared with publisher. Google gets no revenue either from pay-per-use access fees charged by publishers for content, or for ads on the publisher’s home site.
- *Google shares use data with publishers.* There are plans for a “dashboard” that will enable publishers themselves to track how their content is viewed, integrating results into the publisher’s home site, powered by Google.
- *Preservation is not Google’s goal, per se.* Google’s mission is to organize the world’s information and make it universally accessible and useful. Google has been quite clear that their mission does *not* include preservation or archiving.

Session 2—Getting the News Out

Effective strategies for preserving news require understanding the entire lifecycle of news content. “Getting the News Out” examined how news is being gathered and reported today, and how news content is created and managed by actors inside the news industry itself.

Archiving the 'Daily Miracle': Preservation in the Digital Newsroom

Victoria McCargar

New Media Preservation Consultant

Summary: The newspaper industry itself is doing a poor job of preserving back issues and editorial files in digital form. At a time when libraries are collecting and archiving fewer newspaper titles than in the past, it is still not possible to rely upon the newspaper publishers and news aggregators to maintain editorial files and back issues. The newspapers themselves are faced with aging databases, dwindling quality control, legal constraints on archiving of copyrighted works, and inadequate or suspect metadata. These problems are only continuing to accumulate.

Some points raised in the presentation:

- *Electronic content files and related metadata are particularly prone to being lost, corrupted, or otherwise compromised in the content management systems used by news organizations and the electronic aggregations maintained by vendors like LexisNexis, Factiva, and others.*
- *As with the creation of each issue, newspaper preservation requires careful, systematic measures procedures.* There are three general forms digital newspaper content takes:
 1. Editorial files, where texts, photos, tables, and other digital content which appeared in newspapers are stored in content management systems maintained by the news organization;
 2. Press files, usually in PDF format, are output by publishers for printing and thus preserve the appearance of the printed version of a newspaper.
 3. Commercial databases like LexisNexis, Factiva, and others contain the texts of news articles and headlines appearing in newspapers.
- *While the industry maintains a massive amount of content for its own purposes in each of these forms, each has limitations as a preservation solution.*
 1. *Editorial files:* In newspaper editorial and digital asset management systems, articles and headlines are typically stored in one database, and photos in another. Such systems are therefore designed to enable reporters to find and retrieve information quickly and easily. As with earlier clippings files, articles are stored individually and information on the context of the article within the edition where it originally appeared is often lost. Moreover, such archives are often maintained in proprietary systems or using proprietary formats, complicating the task of migration. Finally there is evidence of extensive file corruption and loss of both photographs and electronic texts in publisher archives through inherent defects in software and incomplete migrations.
 2. *Press files:* Archiving the PDF files for a given newspaper edition preserves such contextual information as page layout, illustrations, and advertisements. However, in this format, the underlying bibliographic metadata for individual stories, such as article titles, authors, rights, etc., are lost.
 3. *Commercial databases:* Not all content published in a newspaper is transferred to the aggregator databases. For example, vendor restrictions do not allow the import of Associated Press and other wire service text, or much syndicated material. This means that not all content from a given issue of a newspaper is archived electronically. Moreover, often the transfer of data from and to databases is faulty, and data can be lost or incompletely rendered.

- *Much newspaper content today is also exported to the Web.* However, not everything in the Web version will be included and maintained in the publisher's database.

How News is Distributed: Repurposing News at the Associated Press for Different Audiences

Valerie Komor

Director of the Corporate Archives

The Associated Press

Summary: The Associated Press is the oldest and largest news gathering organization in the world. Its content appears in thousands of newspapers, Web sites, and radio and television broadcasts. By taking an enterprise-wide approach to developing a system for authoring, management, and distribution of its digital news content, AP has created a repository for preserving international news on a global basis and an unprecedented scale. AP's pioneering role in deploying advanced technologies, and its ability to set standards industry-wide may give the organization a formative role in the preservation of electronic news.

Some points raised in the presentation:

- *With the development of the "eAP" or "Electronic AP system," the Associated Press is centralizing the management of all of its content for the first time.* In 2003 AP's first corporate archives was established. In an effort to promote the importance of preservation within the organization, an archivist, not the bureau chief, is now in charge of all older records at a bureau. eAP (created in 2003) is a database that holds all AP content and delivers news in all formats: text, audio, photos, video, etc. There is access to eAP from multiple channels. Eventually eAP will replace the current text archive. It currently contains more than 65 million documents.
- *Digital content has not changed AP's mission, only its method of delivery.* Since 2003, changes have been made in AP's business systems to make AP more competitive. More of the tools are self-service—content can now be accessed independently by many throughout the AP organization and its membership. Additionally, AP packages stories in multiple versions and formats and distributes them through multiple channels in order to reach as wide an audience as possible.
- *The current challenge is what to do with born-digital content.* In the digital realm it is currently easier and less costly to save such content than to delete it, and AP is now finding revenue streams for its older content. There is a tradition of entrepreneurship and technological innovation at AP. If AP believes that it is worth preserving, AP will find the resources to preserve it.
- *With the disappearance of the archives and morgues of individual newspapers in the face of spending cuts and falling newspaper revenues, the AP archive grows in importance.* AP has in place a network of providers of news reporting and incentives to ensure it a steady supply of news content from thousands of local sources. The value of local content and the incentives offered by AP for local news organizations to contribute their news reporting. For example, AP tagging of news text enables search engines to support contextual advertising.

How Citizen Journalism and Blogging are Decentralizing News Gathering and Reporting

Sree Sreenivasan

Dean of Student Affairs and New Media Professor

Columbia University Graduate School of Journalism

Summary: In contrast to the highly centralized activities of the Associated Press,

blogs, citizen journalists, and Web-enabled social networks represent a different model of news reporting and aggregation. Social networking creates “real-time value” in the world of communications. Web 2.0/blogs/sites in news areas are growing and are now almost as important as mainstream media. Rather than dismissing them as insignificant or as a passing trend, libraries must acknowledge them as integral to the larger news record and worthy of preservation, albeit on a selective basis.

Some points raised in the presentation:

- *There is a role for blogs, citizen journalism, and other non-traditional forms of news reporting as valuable adjuncts to the mainstream or old media channels.* These new applications are exploiting Web technology and social trends (such as “the power of BAW,” *i.e.*, the large population of those “bored at work”) to assemble virtual communities around shared interests.
- *These sources have proved adept at fact-checking, and at exposing inconsistencies and bias in traditional reporting.* While most bloggers do not aspire to the extended, original investigative reporting associated with traditional news media, they often create new value by bringing together and synthesizing information from disparate sources. An example is paidContent.org, a site that provides commentary and information on the economics of digital content.
- *Citizen journalists provide valuable on-the-scene reporting on rapidly developing and geographically dispersed events, like the July 2005 bombings in the London Underground.* Other citizen journalists report from areas too remote or dangerous for media journalists. The site [Alive in Baghdad](http://AliveinBaghdad.com) gathered and distributed video footage of testimonies by civilians and daily life outside the Green Zone in Iraq.
- *Traditional news organizations now also recognize the value in social networks as a means of direct access to public opinion and commentary.*
- *Hence, scholars should remain open to the possibilities that these sources create for recording events.* The preservation challenge, however, is to determine what is worth saving, since not everything can or should be archived. When journalists operate under the umbrella of newspaper organizations, those organizations provide vetting and validation of the important materials. Libraries could be instrumental in such validation today.

Session 3—New Uses for Yesterday’s News

The audience for retrospective news content is growing. It includes genealogists, members of the financial community, public policy researchers, and scholars in a wide spectrum of disciplines. Increasingly these users are deploying automated tools to tag and analyze news texts, for purposes such as sentiment mining, identification of bias in political reporting, and “reputation management” in the business world. These tools often include “home-grown” programs and software, but aggregators like Factiva often incorporate their own proprietary tools in their news databases to add value. Such uses raise the ante for electronic access to news, in terms of the functionality expected to be provided by libraries and aggregators along with the news content itself.

Text Mining of Electronic News Content for Economic Research

Panagiotis G. Ipeirotis

Assistant Professor, Leonard N. Stern School of Business

New York University

Summary: Recent business school research at New York University is using

sophisticated text-mining techniques to explore how online content affects real-life economic transactions. The presenter identified obstacles to applying text and data-mining technology to proprietary news back file collections and asserted the importance of making electronic news back files accessible for machine processing by academic researchers. The presenter also raised the question of whether new rights will need to be obtained by libraries from publishers and other holders of intellectual property in news content, to enable such uses by scholars.

Some points raised in the presentation:

- *NYU researchers are learning how textual feedback and comments posted online influence purchases from online vendors.* By tagging electronic text in posted reviews of products, and correlating that text with pricing and sales information, economic value can be systematically assigned to particular words describing product qualities.
- *Similarly, sentiment analysis using techniques for electronic mining of online news can provide “leading indicators,”* tracking the impact of public comments on markets, and may even predict outcomes of elections and other events.
- *This kind of research, however, is stymied by restrictions imposed by commercial aggregators of electronic news and by the lack of standardization in Open Access digitized news collections.* Most commercial aggregators erect barriers to the use of third-party text and data-mining software on their databases. This is to protect their market for costly proprietary tools like *Factiva Insight* and *LexisNexis Analytics*. Such tools are normally priced for the business and legal markets, and are consequently unaffordable to all but the largest research libraries.
- *Such restrictions, in fact, resulted in the shutting down of a NYU research project* when the provider of news content “prohibited the mining of ‘yesterday’s news’ to provide ‘today’s intelligence.’”
- *Many Open Access news collections, on the other hand, are not extensive enough in themselves, or sufficiently interoperable with related Open Access collections to support text-mining to a meaningful extent.*

On Becoming a Librarian (Inadvertently): Mining News Sources for the Policy Research Community

Gary Sick

Senior Research Scholar at the Middle East Institute and Adjunct Professor of International Affairs, the School for International and Public Affairs

Columbia University

Summary: Specialized, community-generated news aggregations, like the *Gulf/2000* Web site and listserv established by Gary Sick at Columbia University, now wield considerable influence in the fields of public policy and national security. But most *Gulf/2000* content is available only to a community of carefully vetted invitees, limited to those with a demonstrated professional interest in the Persian Gulf.

The contents of the site are not archived, per se. The site will present a number of challenges to archiving, such as the restrictions placed on the use of the materials on the basis of privilege, confidentiality, and copyrights.

Some points raised in the presentation:

- *The Gulf/2000 site now holds more than a gigabyte of text,* available for online searching and download by authenticated members. These include policy memoranda, draft policies and analysts’ reports, and other types of unpublished documents, much of it obscure and fugitive information not to be found elsewhere.

(The Open Access portion of the Web site includes links to published materials from a variety of open sources.)

- *Much of the content is posted in real-time*, and many topics discussed are ahead of the news.
- *The listserv allows members to maintain continuous contact with journalists, academics, government officials, etc.*, on issues pertaining to the Persian Gulf, and the membership list currently reads as a Who's Who of Persian Gulf studies. Members include public policy and defense analysts in the government, policy research, and academic worlds, as well as the Middle East correspondents of nearly every major newspaper in the world.
- *Both the listserv and network are moderated, providing a measure of quality control for the resources.* Use of the materials is restricted. Journalists, for example, are not permitted to use documents directly from the system, but may contact those posting particular items.

What Old News Tells Us That Data Does Not: The Uses of News Reports in Monetary Policy Research

Katrina L Stierholz

*Research Information Officer and Director of Library and Research Information
Federal Reserve Bank of St. Louis*

Summary: The presenter surveyed the myriad services the St. Louis Fed's library provides to economists and staff, collecting and distributing a range of types of documentation including news reports. Despite the proliferation of electronic databases of financial information and reports and digital tools for mining and manipulating those sources, back issues of newspapers remain important as sources of time-stamped information—and opinion. Access to these sources can help researchers reconstruct the information and public opinion environments in which monetary policies of the past were created, and can explain and, conversely, reflect the impact of the policies on the same environments.

Some points raised in the presentation:

- *Old news is important to economists.* Much of the data pertaining to past economic performance available in today's electronic databases differs from what was originally reported. What was initially reported was frequently incorrect or incomplete, and has since been revised and updated based on new information. In order to understand and create models of how economies and financial markets behave, however, economists often need to know what data was available to the public and to policy makers at a given point in time.
- *Contemporary newspapers are useful in helping economists reconstruct the knowledge bases of earlier periods.* Economists can find the original data and track subsequent revisions and updates in contemporary news reports. Reports in such media as the *Wall Street Journal* announce not only the release of new data but how those new data confirmed or contradicted earlier reports and expectations.
- *Contemporaneous news reports also help economists create timelines of market events.* News reports provide a sense of the pace and sequence of events, such as the collapse of Long Term Capital Management and Asian financial crisis of the 1990s.
- *Newspaper reports also capture the mood and perceptions—and sometimes confusion—of the public, the media, and the financial communities as they react to conditions at a particular moment.* “Time-stamped” news reports from spring and summer 2008, for instance, show how the slowly emerging awareness of the gravity of the current economic crisis affected the national policy response.

- *Economists use newspapers to find, re-create, and clarify data that is no longer available.* Old news is also a unique source of policy makers' comments and contemporaries' reactions to them. St Louis Fed economist Ed Nelson used U.S. and British newspapers, for example, to trace the thinking and intent of policy makers at the U.K. central bank during the Great Inflation of the 1970s.

The Challenges Identified

The presentations and subsequent discussion among attendees identified several serious challenges libraries face in their efforts to continue to preserve world news.

- As newspapers continue to lose readers and, consequently, their place in the supply chain for the world's news, research libraries must revise their preservation strategies accordingly.
- The Internet is becoming the preferred means of access to current news. This development, coupled with economic pressures, is causing libraries to collect and preserve fewer current newspaper titles than they did in the past.
- Microfilm may soon become obsolete as a technology for capturing newspaper content and no longer meets the needs of most researchers. A wide and growing array of new research methodologies and lines of enquiry draw upon news in both print and broadcast media. And an increasing number of researchers now employ sophisticated search and text-mining tools and applications that require news content in digital, machine readable form.
- While microfilm continues to be the preferred (and in some cases legally mandated) medium of deposit for large collecting institutions, access to microfilm is constrained by increasingly conservative copyright stances, such as that implemented by the Library of Congress to prohibit distribution of foreign newspaper microfilm except by written permission of the newspaper publishers.
- Some libraries are creating collections of digitized back files of newspapers and mounting them on the Web in a variety of platforms. Others are archiving current news directly from the Web either independently or in cooperation with LOCKSS and the Internet Archive. However, there is as yet little uniformity or interoperability among such collections.
- Commercial aggregators like Google, Cengage Gale, Readex, and ProQuest are also digitizing newspaper back files at a rapid pace. But the benefits of those efforts are limited by gaps in coverage and by inadequate or undefined publisher provisions for archiving the digital files.
- Because standardized metadata for digital files created through these commercial efforts is not available, determining what content is actually present in the various resulting products is difficult. This leads to content redundancy and raises the cost of access for libraries and, indirectly, for users.
- News publishers like the *Los Angeles Times* and Associated Press are now investing heavily in robust systems that archive their own born-digital news content. But because these efforts are driven by the publishers' own uses and markets, and often involve proprietary systems and software, the content is preserved in ways that may ultimately render it less useful to historians than the newspaper collections preserved in the past by libraries.

CRL Global Resources Agenda

Discussion at the Forum suggested an ambitious set of measures to enable preservation of today's newspapers and digital news content from the world's regions for

researchers in the future. Those measures, listed below, comprise an ambitious agenda for concerted action, by libraries independently and in cooperation with aggregators and publishers:

- 1. Collectivize the library market for news back files, databases, and tools.** Their newspaper holdings and collective purchasing power provide libraries potential leverage in persuading aggregators and publishers to make digital collections (such as *America's Historical Newspapers*, *Factiva*, *LexisNexis Academic*) and tools (such as *Factiva Insight*) available to libraries on affordable terms. This will require the cooperation of independent and academic research libraries whose collections serve as sources of content for the aggregators.
- 2. Work with publishers, aggregators, and other repositories to ensure adequate archiving of digitized legacy content.** Establish realistic requirements for print and digital archiving of newspapers digitized by publishers and aggregators, and achieve compliance with those requirements. Such requirements will include the creation and disclosure of standardized, page-level metadata.
- 3. Increase the amount and granularity of available information on newspapers held and digitized by libraries, aggregators, and publishers.** Expand the [ICON Web site](#) and database to provide comprehensive information on the scope and completeness of world newspaper titles held by libraries in print and microform; and on newspapers digitized by libraries, publishers and aggregators. Gather and aggregate metadata from such resources as Google's News Archive Search, the National Digital Newspaper Project (NDNP), and the World Newspaper Archive.
- 4. Secure acceptable terms for access to Library of Congress microfilm and paper holdings for researchers in the CRL Global Resources community.** Obtain exemption from the current LC prohibition on duplication to permit a shared copy, for interlibrary loan and limited digital delivery to research libraries, of copyright-era newspapers microfilmed by LC Overseas Operations.
- 5. Press the Library of Congress and other national libraries to implement electronic copyright deposit to acquire and archive current newspapers in digital format.** Achieve CRL Global Resources community consensus on a standard document type definition (DTD) for electronic copyright deposit.
- 6. Press for uniform, persistent archiving by producers of electronic news.** Work with the Associated Press and other news reporting organizations to ensure that their systems for archiving digital news content serves not only their own immediate needs but the needs of future scholars and researchers as well. ❖

World Newspaper Archive Update

James Simon

Director

International Resources



A May 11, 1915 *El Mercurio* (Santiago, Chile) article on a meeting of the Liberal Alliance.

The World Newspaper Archive is a collaborative effort of the Center for Research Libraries and its partner institutions to preserve and provide persistent electronic access to historical newspapers from around the globe. The program was launched in 2008 with the financial and in-kind support of CRL's member institutions.

Background

Through a century of major, sustained investment in acquisition, documentation, and preservation, North American research libraries have amassed a large and valuable corpus of newspapers from all regions of the world. Those libraries' aggregate holdings of newspapers in paper and micro-formats constitute a body of historical and cultural evidence spanning four centuries, which is not, and could not be, replicated elsewhere.

While preservation of global newspapers on microform has enjoyed sustained support in the United States and elsewhere, there is comparably little being done to capture or convert these resources into electronic format. The National Digital Newspaper Program, the flagship effort in the United States, addresses only U.S. newspaper content and is still early in its implementation. Institutions in Europe and other developed countries have begun to launch efforts to convert their own historical contents, but the scale of these efforts is daunting.

The Center for Research Libraries has traditionally focused its cooperative preservation efforts on areas with less robust infrastructure in terms of preservation and digitization capacity. CRL has microfilmed news titles from Africa, Eastern Europe, Latin America, Middle East, and South and Southeast Asia since 1956. Under the World Newspaper Archive effort, the Center and its affiliated member institutions intend to undertake the systematic, large scale digitization of this corpus of world newspapers and news-related materials.

The World Newspaper Archive is envisioned as an ongoing, multi-year and multi-stage endeavor wherein CRL and affiliates will combine expertise and resources to digitize and make available for scholarly use their holdings of newspapers from several world regions. The first phase of the effort involves content from Latin America, starting initially with material in the public domain and extending the effort over a number of phases. Details of the effort are currently found on the [CRL Web site](#).

Goals

With the World Newspaper Archive, we hope to accomplish three things:

1. **Community Access**—NewsBank, through its Readex Division, will provide electronic access to back files of newspapers from microform and

paper holdings of CRL and several major newspaper repositories. Participating repositories include Harvard University, the New York Public Library, the University of Washington, University of Texas at Austin, University of Illinois-Urbana/Champaign, University of Florida, Princeton University, University of California, Berkeley, and University of California, Los Angeles. The World Newspaper Archive will employ the robust and reliable platform for search and discovery that supports Readex Early American Newspapers and Hispanic American Newspapers collections. The annual access fee will be nominal for CRL members and partner libraries that have contributed support and/or content to create the collection.

- 2. Persistence**—CRL will guarantee the long-term persistence and continued functionality of the news content for the CRL community. We aim to ensure not only persistent access, but CRL member control over the future costs and quality of that access.

Moreover, the process of locating and preparing these materials for digitization is generating valuable preservation metadata and information about the existent holdings of these rare materials. This “last-copy” information will be made available by CRL through the International Coalition on Newspaper (ICON) database, to guide librarians in their decisions about preserving and digitizing their local holdings. The project is also subsidizing the cost of replacing lost, damaged, or deteriorated microform copies of the newspapers.

- 3. Growth**—NewsBank shares with CRL a strong commitment to identifying and preserving primary source materials for international studies and research worldwide. This new working relationship will give the CRL library community a voice in the digital conversion of the news collections that they have preserved for the past century in print and microform.

We also hope, with NewsBank, to make available to the CRL community a growing corpus of electronic news content on favorable terms.

Sample of the results screen of the World Newspaper Archive browser.



Preliminary Release

The preliminary release of Latin American Newspapers—a World Newspaper Archive collection—was issued on December 8, 2008. This preliminary release provides more than 60,000 pages of *El Mercurio*, an important Spanish-language paper published in Santiago, Chile.

The WNA Latin American collection will be a rolling release, with completion of the estimated 900,000 pages scheduled to be released by fall 2009. Approximately

70,000 pages will be processed monthly, with periodic releases of new titles. On completion, Latin American Newspapers will include approximately 35 fully searchable newspapers printed throughout this region in the 19th and 20th centuries. CRL will guarantee the long-term persistence and continued functionality of the news content for the CRL community.

Future Phases

Such an ambitious initiative, featuring a massive amount of data from several world regions, will require an ongoing, multi-stage commitment by CRL and affiliated libraries to combine expertise and resources, to digitize and make available holdings of newspapers. Nonetheless, CRL is committed to extending the project to different world regions and is in the planning stage of prioritizing next phases.

It is likely that the program will continue under the present model of focusing the first release of a world region on a larger body of material in the public domain (pre-1923), with subsequent modules adding new titles and/or extending existing titles into the copyright era. This might include orphaned works and/or involve entering into discussions with existing publishers on a limited basis.

CRL member institutions with holdings of news titles in microform are encouraged to contact the author (jsimon@crl.edu) to provide information on complementary collections of material as well as interest in collaboration. ♦

World Newspaper Archive display of the February 12, 1918 *El Mercurio* (Santiago, Chile) article on the founding of the city of Santiago in 1541; the Battle of Chacabuco in 1817; and the republic's declaration of independence in 1818.



FOCUS on Global Resources, published quarterly, is compiled by CRL's Communications Department. Don Dyer, Editor. 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

Assistant to the President
Yvonne Jefferson x 319
yjefferso@crl.edu

Vice President/Director of Programs
and Services
Melissa Trevvett x 316
mtrevvett@crl.edu

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

Director of International Resources
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
Patricia Finney x 328
pfinney@crl.edu

Digital Program Manager
Virginia Kerr x 265
vkerr@crl.edu

Global Resources Program Contacts (800) 621-6044

Director
James Simon x 324
jsimon@crl.edu

Project Manager—DSAL Project
Gerald Hall x 318
ghall@crl.edu

Project Coordinator—Global
Resources Network
Judy Alspach x 323
jalspach@crl.edu

Project Coordinator—ICON
Carolyn Ciesla x 315
cciesla@crl.edu

Contact for Information Regarding: (800) 621-6044

Billing/Invoicing/Deposit Accounts
Accounting Department x 346

Membership/Communications/
Public Relations/Marketing
Don Dyer x 317
ddyer@crl.edu

CRL Publications Orders (*non-FOCUS*)
Meeting/Schedules
Yvonne Jefferson X 319

Loans/Photocopies/Demand
Purchases
Access Services Department x 314

Microfilm/Microfiche Sales
Lorraine Farley x 336
lfarley@crl.edu

Purchase Proposals
Mary Wilke x 351

Collection Deposits
Mary Wilke x 351