Submitted on: 8/16/2014

A New Template for the Preservation of Electronic News

Bernard F. Reilly, Jr.

Center for Research Libraries, Chicago, USA
reilly@crl.edu

Abstract:

The digital revolution and the emergence of the Web have transformed the production and distribution of news, rendering traditional library approaches to news preservation obsolete. Recent CRL analysis of the electronic news lifecycle suggests that the new preservation templates will involve a fundamentally different model of stewardship for libraries, focusing less on collecting and archiving, than on systems analysis, partnership building and new attention to the modern practices and methods of news-based research.

Keywords: Electronic news, preservation, digital humanities, digital media.

The Challenge of Preserving Electronic News

Over the last two decades, the production of news has been radically altered by the digital revolution and the emergence of the Web and mobile communications platforms. The advent of new digital content production and management technologies and practices has fundamentally transformed all phases in the news life cycle -- gathering, writing, editing, managing, publishing, and disseminating news.

This “sea change” has shattered the existing templates used by research libraries in preserving news, making it necessary to devise new approaches that are better suited to promoting the longevity and integrity of news that is born-digital. New templates for preserving news must take into account not only the dynamic technologies now in use in the field of news publishing, such as social media platforms and enterprise-scale content management systems, but also the highly developed infrastructure of standards and standardized practice that the telecommunications and journalism industries have put in place.

I shall propose here some new strategies, suggested by recent research and analysis undertaken by the Center for Research Libraries (CRL). Specifically, these strategies are suggested by the findings of a study undertaken by CRL for the Library of Congress in 2011, [1] and by subsequent discussions at a 2013 CRL forum on electronic news.[2] The approaches we propose involve cooperative action by research libraries, some of which has already been initiated by North American libraries through the agency of CRL. We believe that taking such actions is not just a matter of practicality or expediency -- it is necessary to prevent the ongoing, wholesale loss of important historical and cultural evidence. We are now nearing the end of the second decade in which the Web has been a major venue for news distribution. Without a coherent strategy for the systematic preservation of web-based news, we face a widening gap in the historical record.

CRL's Analysis

In September 2009 the Library of Congress National Digital Information Infrastructure and Preservation Program (NDIIPP) convened an invitational conference to explore possible strategies for collecting and preserving digital news on a national basis. For purposes of discussion at that forum, LC defined digital news to include, at minimum, "digital newspaper Web sites, television and radio broadcasts distributed via the Internet, blogs, podcasts, digital photographs, and videos that document current events and cultural trends."

The Library of Congress has a vested interest in addressing this challenge. The Library has been a critical link in the newspaper supply chain for the major academic and independent research libraries in the United States. LC is the repository for legal deposit in the U.S; was partner with the National Endowment for the Humanities on the long-running and successful United States Newspaper Project; and maintains overseas acquisition and microfilming operations that secure newspapers from several world regions for U.S. libraries. CRL, a consortium of over 200 North American academic and independent research libraries, is also a major repository of foreign and U.S. newspapers in print and microform, and has coordinated its own preservation activities with LC preservation efforts since 1949. Both the LC and CRL programs, however, were designed to preserve news published in print format. It was evident that new solutions are needed for electronic news.

Prompted by the LC discussions, CRL undertook an in-depth study of the flow of news distributed in print and online form by four U.S. newspaper publishers. In a 2011 report CRL documented and mapped the "lifecycle" of news content and information from production and sourcing, through editing and processing, to distribution to end users. The mapping was to provide specific data on which to formulate a strategy for preserving news published online by the major U.S. newspaper publishers.

Four U.S. newspapers served as the test bed for the study: The Arizona Republic, Seattle Post-Intelligencer (since 2008, seattlepi.com), Wisconsin State Journal, and The Chicago

[1] See Preserving News in the Digital Environment: Mapping the Newspaper Industry in Transition, Center for Research Libraries for the Library of Congress National Digital Information Infrastructure and Preservation Program, 2011.
http://www.digitalpreservation.gov/documents/CRL_digiNews_report_110502.pdf. Contributors to the report were: Jessica Alverson, Kalev Leetaru, Victoria McCargar, Kayla Ondracek, James Simon, and Bernard F. Reilly.

[2] For the report on the 2013 CRL forum, see "Access to News in the Digital Era" in Focus on Global Resources, 32:4 (Summer 2003). http://www.crl.edu/sites/default/files/focus/pdf/FocusSummer2013_1.pdf.

Tribune. These titles were chosen to represent a broad segment of the domestic newspaper industry and a range of publisher types. To provide further context, CRL also examined the workings of three other news organizations: The New York Times, Investor’s Business Daily, and The Associated Press.

CRL’s analysis examined the workflows and systems employed in the production of web-based, electronic news by the four organizations, with reference to analogous stages in the production of print newspapers. For purposes of the analysis, CRL identified the three major stages in the life cycle as:

A. Sourcing: the gathering of news information and content by the news organization from those who create, report, and/or own that information and content;

B. Editing and Production: the editing, processing, tagging, and other enhancement of news content and information, and its preparation for distribution through various media. CRL also examined a phenomenon that has become a common feature of web news: the sourcing of content from highly distributed and varied third-party sources, including data feeds from financial markets, multimedia advertising from ad servers, and comments from users on social media.

C. Distribution: dissemination and exposure of news content, and products and derivatives thereof, through print and online media. This included direct publishing to the web and mobile networks as well as electronic facsimiles and text feeds to aggregators like LexisNexis and Factiva.

The Findings

CRL found that news sourcing, production and distribution in the electronic era differ so fundamentally from those same processes in the print era as to render traditional library preservation modes, like the archiving of back files, harvesting of published content, and other custodial approaches, essentially ineffective. Three developments were factors here:

1) The Print-Digital Shift: Print editions are gradually but inexorably becoming a smaller, if not secondary, portion of the total output of news publishers. Given the amount of “web-only” news reporting and the frequency of updates to newspaper websites, the printed product can no longer be considered the “edition of record.” The most comprehensive record of news in the developed world is now the electronic record.

2) New e-Content Management Capabilities: In recent years the news media have developed robust capabilities for managing, processing and mining vast amounts of complex digital content. Large-scale enterprise-wide editorial and content management systems allow publishers to create and maintain extremely rich text and multi-media content and interactive datasets. Using highly sophisticated, and often proprietary, applications news organizations today routinely annotate and enhance their content with structured metadata about authorship, rights, place of origin, and subject. This metadata is stripped from that content on publication, but is maintained within the publisher’s systems. News industry consolidation in the wake of widespread deregulation of media ownership, and the complex demands of handling digital text, still image, moving image, audio, and statistical content, are driving the major media

organizations to centralize their content management operations and even move those operations to "the cloud." As a result, news "morgues" are now larger than ever, and uniformity and standardization of practice have been growing across the industry.

3) Changing Research Practices: Researchers today are using the news record in myriad, innovative new ways. Historians, economists, public policy experts, sociologists, and other researchers are employing sophisticated new tools and applications for discovery, analysis, topic modelling, and visualization of digital data and content. Electronic news offers a particularly useful source of raw material for computer-assisted research, and this source is being actively mined by scholars, often working directly with the publishers. Therefore, long-held assumptions about researcher needs, upon which library models of preservation are based, need to be updated to acknowledge these trends.

These three developments have implications for library preservation. The robust digital content management systems employed in the news industry, and the inherently dynamic nature and ever increasing technical complexity of digital content, make library preservation of news content outside of the native production systems impractical and unlikely. Web crawlers are unable to keep pace with the constant flood of rapidly changing content on the major news sites, a flood that only grows with each passing year.[3] Moreover, the pay walls recently erected by the News Corporation, New York Times, Wall Street Journal, and others present a new obstacle to the harvesting of news from the websites of publishers.

For similar reasons, the nascent electronic legal deposit programs of many national libraries are either not scaled to ingest significant amounts of electronic news content, or are not designed to capture online news content in formats that support current research practices. The products of these efforts cannot be easily searched or mined electronically, and are often off-limits to most users because of copyright restrictions.

Other library practices are no better suited to the current realities of news production and consumption. Many libraries, for example, now archive the page-image files, usually in PDF format, generated by newspaper publishers in the course of producing the printed edition. These static files capture none of the functionality of web-based news and so archiving in this format, however stable, may prove to provide little of the functionality that future researchers will require.

All things considered, this "asymmetry" of digital content management capabilities between publishers and libraries, combined with the high expectations of modern news researchers, suggest that the news industry itself is likely to be the primary locus of the preservation of today's news record.

The Outlines of a New Preservation Template

If this is so, what then is the role of libraries in this new reality? How do we reconcile our responsibility to ensure the continued integrity and accessibility of the journalistic record of

[3] For example, a recent examination of the Internet Archive's harvest of the New York Times online (landing page http://www.nytimes.com), between December 12, 1998 and September 05, 2007, as represented in the Wayback Machine, shows that only about 1,008 "issues" were captured over the course of 3,000 days. In the captured materials, important non-textual content (photos, graphics) is missing, and many links (like ad server feeds and third-party-provided financial data) no longer functioned.

human events, individuals, nations, ideas, and ideologies for the researchers, current and future, whom we represent with this tectonic shift in technology and economics? In a report on its 2013 forum, CRL outlined a number of strategies that libraries might adopt to fulfill this important historical mandate. The report sketches in broad strokes what an effective approach to preserving digital news might "look like." Specifically, three kinds of action at the national level are called for.

1) Monitor and Document Electronic News Production and Distribution: Evaluating and documenting the technologies and processes of contemporary news production and distribution could enlarge our understanding of these processes and help preserve for future researchers the ability to reconstruct and excavate how news was produced today after today's news production infrastructure has become "legacy." This may be particularly important in the arenas of law and government, where stringent requirements exist for judging the admissibility and credibility of evidence. In the legal world, newspapers are commonly cited as reliable sources of information about contested events and actions. In the past, precedent and the laws of evidence established what kinds of documentation could be brought to bear in civil and criminal proceedings. Historians and other scholars apply similar principles in their evaluation and use of documentary evidence. Libraries could, as curators have in the past, play an important role in preserving the "chain of custody" and integrity of digital content in the context of the new modes of news production and exchange.

Particularly challenging is documenting source and provenance of content published in online news but held in third-party, proprietary databases, such as financial data held by Bloomberg, and public opinion data held by Pew, Nielsen and others. Understanding how these data providers manage their digital assets will be important in validating electronic news as evidence for future historians and jurists.

For its part CRL now endeavors, as resources permit, to monitor the technologies used in the production and distribution of news content. We plan to continue to refine and update the maps and life cycle information contained in the 2011 report, adding details, particularly about technical processes that, because of their proprietary nature, we were not able to obtain within the timeframe of the original study.

2) Identify the Needs of our Consumers: It is probably time to re-examine the underlying goals of our preservation activities. Libraries cannot afford to base resource decisions on outdated assumptions about the practices and needs of contemporary and future scholars and researchers. Most preservation and acquisition policies are based upon premises that were valid during the print era, before the radical changes in information consumption brought about by digital technologies. These premises need to be reexamined in the light of today's technologies.

Libraries need to better understand the kinds of tools and analytical practices today's researchers bring to bear on their use of electronic news, and the kinds of scholarly products and outputs those researchers are creating. We know little, for example, about the sophisticated text and data mining applications and practices now widely used in economics, finance, development studies, and other fields. Nor do we know how researchers determine which of the multiple versions of a given "story" or news report is the “authoritative” version, and whether the “snapshots” of news web sites, in whose archiving we invest so much, actually capture that version.

One goal commonly cited by library curators and archivists is to preserve today’s news in a form in which future researchers not only can recover the content of the news but can understand how contemporary readers “experienced” that content. In a world of dynamic media, attempting this is probably an unachievable goal. The proliferation of devices for accessing the same news content in multiple forms (mobile phone, tablet, PC, Kindle, etc.) and the variety of applications used to present that news (RSS, news readers, mobile apps) atomizes the user’s "experience" of electronic news into a million possible variations. In addition, the online transaction between the producers and consumers of today’s news involves customizing the content delivered to suit an individual user’s profile. As the “real-time analytics" built into the technologies for news distribution increasingly shape the content of news, it is unrealistic to expect to be able to reconstruct all types of news consumption experiences in the future.

3) Reframe the Relationship between Libraries and News Organizations: National libraries may need to reimagine legal deposit requirements along the lines of a mutually beneficial relationship between the library and its domestic publishing community. They might leverage their historical rapport with their domestic news industry to find ways to ensure the long-term integrity of the rich news content produced and managed by those organizations in forms that support the libraries’ constituents over time. This could involve advocating for new incentives to be embedded in legislation, to reward publishers for grant new usage rights or provide other benefits that better support research. Or perhaps such grants might be achieved in the context of negotiating national site licenses for electronic access to news.

Major academic libraries, instead of harvesting and maintain electronic news apart from the publishers’ native, host systems, might be able to persuade publishers to implement measures that support persistence and other protections within those native systems, through negotiated licenses of digital news databases. Such licenses could well be structured to guarantee uninterrupted, long-term access to the news content in ways that better support the work of researchers in the humanities and social sciences. Publisher concessions could range from ensuring stable URLs for historical content, to providing standardized tools for text and data mining, to escrowing proprietary interface programming code to provide critical database functionality in the event of a publisher failure.

Libraries represent a significant sector of the media organizations’ customer base. As such they are positioned to demand, at minimum, greater standardization and uniformity. This would reduce the potential costs of their “taking custody” of a publisher’s news the content and the necessary enabling systems in the future. Under such an arrangement, the research library community would not actually “own” the content, but could exercise a measure of control over it.

Limits on library resources require that we concentrate on measures that are most likely to produce tangible benefits. North American libraries, through the agency of CRL, have begun to take some steps to address the new realities of electronic news. CRL uses its webinars and Primary Source Awards program to identify and publicize the innovative practices of researchers who are using large bodies of news text and content. CRL has also begun to try to leverage the collective influence of its constituent libraries on a number of news publishers and aggregators. Working closely with the North East Research Libraries consortium and other U.S. partner consortia, CRL now negotiates terms for purchase of and subscription to major electronic news databases. One promising recent initiative involves an academic site license for the New York Times online for North American academic libraries.

In the struggle to preserve the world's important data, the current generation of curators and archivists will have to reimagine the roles that libraries will play. Clearly more analysis, work and investment are needed. The approaches that I suggest here do not resemble those we traditionally associate with library preservation. Yet they are necessary to prevent the ongoing, wholesale loss of historical evidence and documentation critical to civil society.

As we near the end of the second decade in which the Web has served as a major venue for news publication and consumption, we lack a viable strategy for the systematic preservation of electronic news. This lapse on our part, if not corrected, will cause the widening gap in the historical record to become an unbridgeable chasm.