PRESERVING NEWS IN THE DIGITAL ENVIRONMENT: OUTLINE FOR AN AGENDA FOR NORTH AMERICAN LIBRARIES

An outline for a libraries agenda, based on the report,"Mapping the Newspaper Industry in Transition" from the Center for Research Libraries

June 20, 2013

Observations and Conclusions

Included here are seven summary observations on strategies that libraries might adopt to ensure the long-term preservation of the journalistic record and more comprehensive researcher access to news in electronic form. Our aim is to describe in broad terms what an effective effort to preserve digital news might "look like," based on the findings of this study Preserving News in the Digital Environment: Mapping the Newspaper Industry in Transition. The ideas offered here are relatively undeveloped, but may bear exploring in further depth in the months ahead.

There is some urgency to developing such a strategy, however. We are now nearing the end of the second decade in which the Web has been a major venue for news "publication," without a coherent strategy for systematic capture of Web-based news. This lapse in coverage is creating a widening gap in the historical record.

Bernard F. Reilly
Center for Research Libraries

1. On preserving the electronic facsimile

A number of national libraries have begun to acquire, or are exploring acquiring for their collections through copyright deposit of PDF files produced by newspaper publishers. Indeed copyright deposit has been a major source of U.S. newspapers for the Library of Congress. We believe that this approach is limited in effectiveness, and is an interim strategy at best.

Nearly all major publishers produce the "e-facsimile" page image files for printing purposes. They are, moreover, a fairly complete record of the printed newspaper: they include not only the articles, features, news reporting and accompanying visual content in relatively high resolution, but much of the advertising and syndicated material that appears with those items as well.

The newspaper e-facsimile page-image files, output in Adobe PDF or PostScript formats, lend themselves to ingest and to processing for archiving. While the metadata routinely embedded in or accompanying those materials is not uniform, the aggregators and plate-setters have succeeded in normalizing same and processing them into their own workflows. (See report Sections C.1 and C.2.) Given the ability of these service providers to deal with the newspaper PDFs and codes through the use of profiles, it should be possible for libraries to come up with a standard metadata requirement that enables an electronic deposit/ingest system to recognize and configure the PDFs for a given issue for processing and viewing. Libraries might even specify certain minimum metadata to be included in the PDF files using the Adobe XMP schema as a requirement of copyright registration. This should be no more burdensome for publishers than completing the current copyright registration form.

However, we see two disadvantages to investing much in that approach. First, the contents of the print editions are gradually but inexorably becoming a smaller portion of the total daily news output of the publishers. There is a great deal of Web-only news reporting and writing coming from the newspaper publishers, and because of the frequency of updates to newspaper websites, the number of electronically "published" versions of a given article or feature has multiplied exponentially.

Moreover, demand for the e-facsimiles may well decline as consumers become more accustomed to reading on the Web and on personal devices. Tablet devices, in particular, seem to be rapidly gaining favor as a format for news consumption. Together with the decline of print circulation, this could lead to discontinuation of PDF production by the publishers, and render the effectiveness of a PDF acquisition program short-lived.

Further, because the page image files, as output by the publisher, have only minimal metadata attached, by archiving them a library does not reap the benefits of the extensive annotation and coding of the content files that takes place within the editorial and digital asset management systems of the publishers, and which provides useful information about authorship, rights, provenance, and subject matter of the content. Such information, if harvested, could provide the basis for library management of the news content and for human and automated analysis of large news corpora by researchers and semantic interaction with other relevant bodies of research materials. Perhaps the publishers, for instance, could export a uniform XML package at the issue or article level, perhaps captured on output from the pagination or editorial system. This is the moment in the lifecycle of the news item when the annotation is richest and the data most highly structured.

In general, we think that a preservation/acquisitions model built around ingest of PDFs would, in effect, "leave money on the table."

2. The limitations of Web archiving by current methods for news

On the Web there is essentially nothing comparable to the print "edition" as such. The collection of pages that makes up a newspaper Web site at any given time has no integrity beyond the instantaneous view by a single individual using a particular browser or device. Moreover, even this collection of pages cannot be captured effectively because of the time lapse involved in crawling from one page of a Web site to the next.

Many, if not all, of the existing Web harvesters are ineffective in capturing the full contents and functionality of newspaper web sites. Such harvesters are unable to keep up with the rapid pace of content updates on the major news sites like the New York Times and Chicago Tribune, a pace that only accelerates with each passing year. Nor are they able to mirror the customization such sites provide to individual users, with content delivered to a given user's device varying according to the operator's preferences and browsing history. For example, the Internet Archive's harvest of the New York Times between December 12, 1998 and September 05, 2007, as represented in the Wayback Machine, holds only about 352 "issues" of http://www.nytimes.com, the landing page of *The New York Times*. In these, much of the non-text content (images, graphics) is missing and many links (like AP and Reuters feed) are not functional. The Library of Congress's recent crawls of the Detroit Free Press improves on the Internet Archive's results, but still is hampered by similar limitations.

Moreover, the pay walls now being erected by the News Corporation, New York Times, Wall Street Journal, and others present a new obstacle to harvesting news from the Web sites of publishers. And the subscription services planned by Apple (per the App Store) and Google (One Pass) may put another layer of defense between the harvester and targeted news content. This latter development could put more control of content and distribution in the hands of the device makers and the search engines.

It may be a better strategy to work with major news organizations and concentrate on capturing articles and other categories of discrete news objects, rather than entire sites. The article remains the fundamental unit of content around which the news production and discovery systems, including Web production systems and news search engines, are built. (Google News Search, for example, returns search results in the form of article references.) This gives the article some enduring integrity as a news "object" worthy of preservation.

3. A preservation role for the large media organizations

With the convergence of the "vertical" media (see David Pogue's January 2, 2011 NY Times column) libraries might consider organizing their news collecting and preservation efforts around some of the major media organizations, such as the New York Times, Associated Press, and News Corporation; or sectors (financial information, advertising, news cooperatives), rather than around formats (newspapers, serials, broadcast).

Content production and management activities are increasingly centralized, as a result of the growing cost of content management and consolidation in the news industry in the wake of deregulation of media ownership. The same factors are driving the merging of editorial and content management systems and reliance upon data centers to process and store content shared across a given media organization’s several platforms. This consolidation promotes uniformity of practice, and might thus present an opportunity for libraries. One strategy might be to attempt to influence the major media organizations to manage their content in ways that would better serve current academic and policy researchers and future historians, (like the use of persistent digital object identifiers).

This approach would involve working not with local news organizations but with the large parent companies that increasingly control the content of their local newspaper properties. These include media titans like Gannett, the News Corporation, Tribune Company, and McClatchy. As some of these also control and create content for radio and television outlets, dealing with them might yield benefits for LC collections in the broadcast arena as well.

A variation on this strategy might be to work with the major producers of the production and content management systems like *NewsGate* and *SAVE.* Those producers might provide information about their systems and systems engineering that enable a fuller understanding of the content management processes. It might then be possible to employ such knowledge to create a repository or repositories that “mirror” selected content produced by multiple news organizations that employ the same systems.

4. Understanding new and emerging research needs

It is time to examine the underlying goals of our news preservation activities. Most library preservation and acquisition policies are based on certain premises that held up during the print era, before the radical changes in information production and consumption brought about by digital technologies. These premises need to be reexamined in the light of today’s technologies. We may be making false assumptions about the practices and needs of contemporary and future scholars and researchers.

For example, a commonly stated library goal 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 citizens perceived and experienced the news. This has become a tall order in the age of dynamic media. The proliferation of devices for accessing the same news content (mobile phones, tablets, PCs, Kindle, etc.) and variety of applications used for presenting that news (RSS, news readers, news apps) atomizes the user's "experience" of electronic news into a million variations. In addition, the online transaction between the producers and consumers of today’s news involves customization of the content delivered to individual user traits. As these “real-time analytics" built into the technologies for news distribution increasingly shape the content of advertising and news, it is probably unrealistic to expect to be able to reconstruct all types of news consumption experiences in the future.

The needs of today’s commercial, public policy and academic researchers have also changed radically in recent years. These users are increasingly relying on computers to locate – and in some cases even to interpret -- information in large bodies of news content. Witness the hedge funds’ large recent investment in news mining engines -- and the price of Factiva *Insight*. The value of metadata for these researchers might rival that of the content itself, e.g., for generating new information and findings about people, organizations, subjects and places.   Rich metadata added by publishers and aggregators enables computer-assisted quantitative and qualitative analysis across large bodies of text and media content that is not possible using the published content alone.

Because of the growing gulf between breaking news and studied analysis, the Web and broadcast channels today are becoming the primary venues for reporting new developments, while many newspapers are becoming platforms for detailed, reflective analysis. Whereas in the past both types of information had a place in the newspaper, user demand for real time, actionable information is shortening the time lapse between the sourcing and capture of news information and the dissemination of that information. (Real-time stock market information is continually fed to the news Web server from independent data providers like Bloomberg and Dow Jones, updating several times a minute.) Therefore it is safe to assume that researchers will be mining Web news for different types of information than they sought from newsprint. To construct effective preservation strategies we simply need to know more about these uses.

5. The value of documenting news production and distribution systems and processes

Detailed documentation and mapping of the processes involved in the production and exchange of news content has a forensic value. Enlarging our understanding of these processes might help preserve for future researchers the ability to analyze evidence and excavate information produced by discarded systems. This may be particularly important in the arena of law and government, which have distinct needs regarding the preservation and presentation of evidence, and where newspapers have always played an important role in documenting contested events and actions.

Not enough is known about how well the current methods of producing and distributing digital news content serve legal needs. In the past, precedent and the laws of evidence established what forms of documentation could be brought to bear in civil and criminal court actions. It might be useful to examine how the chain of custody and other legal principles are being affected by the production and exchange of electronic news in the new systems. This could then inform development of LC's news preservation strategy.

We also need to look closely at the relationship between content provided by newspapers and that held in major proprietary databases, such as weather data gathered and preserved by the National Center for Atmospheric Research and National Weather Service; financial data held by Bloomberg; and public opinion data held by Pew, Nielsen and others. Understanding how these databases and the digital asset management systems of the news giants like Associated Press and Gannett will be important in evaluating and interpreting the electronic news that survives. Libraries can play a role in enabling this understanding.

6. The prospective benefits of collective bargaining

LC and CRL have together formed a key link in the newspaper supply chain for the major US academic and independent research libraries. This role has been supported by the U.S. copyright deposit system, CRL and LC microfilming operations, and by collecting through LC overseas field offices. Those activities, unfortunately, have been undermined by economic, technical and policy changes: funding shortfalls, copyright deposit restrictions, the rapidly rising costs--and risks -- of maintaining foreign offices, and so forth

Instead of attempting to acquire and maintain electronic news apart from the original host systems, North American libraries working together, through the agency of CRL, might consider securing preservation services and rights through negotiated licenses of digital news databases structured to guarantee uninterrupted, long-term access to the news content for research purposes, provided by the publishers and/or the major aggregators. Such licenses could build upon current library subscriptions to paid news services like the Wall Street Journal and to text aggregator services such as LexisNexis, Factiva, and Access World News. Such licenses would have to be iron-clad and ensure specific rights and privileges with regard to use of the news content over the long term. Concrete assurances for long-term access to the content could be provided either through a trusted third-party dark archive (along the lines of Portico or CLOCKSS), or through rigorous data maintenance measures taken by the publishers and aggregators themselves and periodically audited. (CRL is currently assessing the self-archiving systems of both ProQuest and Readex.)

Such an arrangement might be negotiated on behalf of the U.S. research libraries community with aggregators like NewsBank, Dow Jones, and LexisNexis, or with the media organizations themselves. The last are yearly becoming fewer and fewer in number. And the largest organizations, like the News Corporation, Bloomberg, and Associated Press, are producing news for multiple delivery platforms: print, Web and broadcast. Therefore dealing with those organizations could conceivably satisfy research libraries' needs for news from all three types of platforms.

7. A broader working alliance between libraries and the media industry

The scope of the license could conceivably be enlarged to provide electronic access for the major US universities and colleges as well as for public policy institutes such as the Hoover Institution, Brookings, RAND, and others. Those organizations might be expected to bear a significant share of the cost of this investment. Under such an arrangement CRL might serve as agent of the larger research community -- not as a repository, but as the guarantor of accessibility and integrity of the record of US news production. CRL now fulfills an analogous role in the world of print newspaper collections.

This would indeed require a substantial investment by CRL and its member libraries. But that investment would pale in comparison with the cost of replicating separately the functionality of the many systems required to store and manage electronic news as it is produced today. A side benefit of such an arrangement might be the exchange of technology, preservation, and subject matter expertise and knowledge between the media/publishing world and the CRL community.

The scale of such an investment, moreover, might well purchase for research libraries some influence on the production and content management practices of the news organizations. Representing a significant sector of the media organizations' customer base could position libraries to force standardization and uniformity that would reduce the future costs of its "taking custody" of the content and necessary enabling systems in the future. Or libraries could require an escrow of the program code and design for the key news interfaces and content management systems. Under this arrangement the research library community would not "own" the content, but would exercise a measure of control over it.