Breaking the Mold: How Digital News Production Changes the Preservation Paradigm

Bernard F. Reilly, Jr.
Center for Research Libraries, Chicago, U.S.A.
reilly@crl.edu

Abstract:

In the last decade the continuing digital revolution and the emergence of mobile and other new communications platforms have radically altered life cycle of news. News gathering, reporting, editing, and distribution in the digital environment increasingly rely upon robust, multi-purpose content management capabilities deployed by large media organizations like the BBC in the U.K., The New York Times in the U.S., and Grupo Clarín in Argentina. In addition, most news published and broadcast today appears in digital formats that depend upon interaction and linkages to other networked content that may reside on entirely separate platforms and cloud repositories. This technical "sea change" challenges the mechanisms upon which research libraries have traditionally relied to preserve news for future generations of scholars and researchers, mechanisms such as legal deposit, print subscriptions, and microform and digital reformatting.

The author of this presentation examines the dynamic technologies used in news production today, and the implications those technologies have for library preservation strategies. The author also reflects upon the broader societal functions that libraries fulfil in preserving news, and suggests how library roles must evolve new strategies to remain relevant in the electronic news environment.

Keywords: News; digital news; preservation; databases; licensing.

The Radically Altered Landscape of News Production

During the last two decades the continuing digital revolution and the emergence of mobile and other new communications platforms have radically altered the lifecycle of news. In the digital environment the stages in that life cycle, i.e., news gathering, reporting, editing and distribution, now normally involve robust, multi-media content management technologies and capabilities. Major media organizations like the BBC in the U.K., Gannett and The New York Times in the U.S., and Grupo Clarín in Argentina now deploy sophisticated digital content management systems (CMS's) to support the production and versioning of news content for websites, smart phone apps, broadcasts, and other distribution platforms, transforming news content itself in fundamental ways in the process.

Some the systems now in use are home-grown, like The New York Times's *Scoop* and The Guardian's *Composer*.[1] Other systems, like CCI Europe's *Newsgate* and Acquia's *Content Hub,* are produced and marketed to media companies by large technology firms, often along with related cloud services like data storage and maintenance. Large-scale content management has now become essential to keeping pace with a radically accelerated news cycle, where updates are no longer tied to morning and evening deadlines but constitute a continuous feed of "real-time" information delivered to consumers almost instantaneously. Reporters now compose their articles directly into the CMS systems, where writing and editing are facilitated by powerful version-control and change-tracking functionality, enabling the enrichment of stories and features with contextual information and standardized subject tags.[2]

Because many journalists now work across media platforms (print, web, mobile, broadcast) and file accompanying photographs and video along with their stories, the systems also provide tools for composing and editing still and moving images. These tools facilitate the framing, cropping and manipulation of photographs, and sophisticated video editing that previously would have required the services of a layout artist and post-production studio. The systems themselves enable publishers to capture and maintain the rich metadata that comes with the stories, photographs and video uploaded by reporters and photographers: information on authorship, usage rights, time-stamping, and GIS coordinates of subjects, and so forth. This metadata persists within the CMS and has utility for republication and other subsequent phases of the digital news lifecycle.

The systems themselves can be multi-platform, capable of handling multiple media and all stages of the news production process, or can be used in tandem with third-party applications like WordPress for blogging and "recommendation engines," for generating those "Most Emailed" and "Trending Stories" features that appear in the margins of many news site web pages.

Interoperability of news content is now pervasive. "Digital first" production workflows generate news output, websites, mobile app editions, electronic text feeds for aggregators and, increasingly, the versions of articles, videos, and other materials to be embedded by APIs in social media platforms like FaceBook and Twitter. In fact most news published or broadcast today appears in formats that depend upon interaction and linkages to other content often maintained centrally in third-party platforms and utilities like *YouTube*, *Document Cloud,* and *Wikipedia*.

The new content management systems are not mere add-ons to corporate technical assets, but involve sizable financial investments and are now at the core of the news enterprise. They are essential components of the infrastructure that the news industry has put in place to maintain content for the long term. As such they are analogous to the morgues maintained by newspaper publishers in the print era and the analog video and sound archives amassed in the broadcast era. While publishers and broadcasters readily jettisoned those materials once they became too expensive to support, the "long tail" economics of digital content now reward media companies for maintaining their older content. The plummeting cost of digital storage and new possibilities for repurposing and monetizing electronic content offer new incentives for retaining "historical assets."

These developments have implications for the role of libraries in preserving news. The digital production processes challenge the mechanisms libraries have traditionally deployed to preserve "the first rough draft of history" for future generations of scholars and researchers, mechanisms such as acquisition of newspapers through legal deposit and subscription, and conversion of paper media to microform and digital format and analog broadcast to digital media.

New Content and Old Templates

The templates libraries have used for centuries to preserve news are based on the idea of fixed and discrete units of content. Newspapers appeared in daily or weekly editions, and although multiple editions published on a given day have often bedeviled catalogers and bibliographers, libraries were able to identify and harvest news reporting fairly effectively. News content was normally image- or text-based, delivered periodically in geospatial environments, and read by humans. Within an edition the composite units, i.e., articles, advertisements, photographs, and data tables, were all complete in themselves. The preservation model devised to accommodate content packaged in that way involved the idea of a self-contained repository or “collection”, where fixed, discrete content could reside for the long-term.

That model is poorly matched to the digital content paradigm, where content is dynamic and in formats that are constantly evolving and that may function only in proprietary, native digital environments. News content in the digital world can be delivered instantaneously, and is as often read and interpreted by text-mining algorithms and aggregation engines as by people.

There have been efforts by libraries to address the digital news problem, but they are hampered and rendered ineffective by adherence to the old templates. A number of national libraries and research libraries in the U.S. and Europe are building the preservation efforts around the capture of PDF files generated by news publishers. PDFs are essentially the high-resolution digital image files produced to serve as source files for the pages of printed newspapers. As static images, however, they fail to capture the dynamic characteristics, and thus the full content, of digital news. Nor do they preserve the rich metadata and subject tags that are embedded in born-digital text, image and data files maintained in their original production environments, the content management systems.

Another widespread approach taken by libraries to preserve web-based news, the use of web crawlers to harvest news sites, tends to preserve only partial and disjointed sections of active news sites, and these are captured on an erratic basis. Because of the considerable amount of time needed for the harvesters to crawl a news site of any scale, such as *The New York Times* or *Chicago Tribune*, the web archive efforts provide little useful content.[3]

Both approaches are products of outdated notions of preservation, based on the premise that content always resides in “objects,” which are subject to harvesting and capture, as opposed to being fluid, interconnected, and dependent for their functionality and even their meaning upon native digital environments and technology platforms.

Economic as well as technological factors also come into play. In the past, libraries acquired newspapers primarily to satisfy the constituents’ need for current awareness, i.e., intelligence on business and current affairs. Different considerations motivated the long-term maintenance of the works thereby collected, and resulted in the building and stewardship of massive and comprehensive collections of newspaper back files and broadcast recordings at national libraries and research institutions. Much of that activity catered to the interests of future generations, and their historian interlocutors of the past, in the survival of the public record, or “first rough draft of history.”

Today the flow of current news content, be it reporting on foreign and local affairs, market and economic data, sporting results, or celebrity gossip, has been reengineered to meet the expectations of news consumers for immediate and even real time access to information. The commercial providers now fulfill that need better than libraries: much breaking news is free on line, thus reducing user reliance upon libraries to subsidize current awareness. Other news is behind the pay walls of news sites like nytimes.com, faz.net, and Bloomberg. Libraries may subscribe to these databases on behalf of their local constituents, but have no say in long-term persistence of the content they’ve purchased.

Other changes in the nature of the news business are also having an effect. Local newspapers are now likely to be owned by a distant parent company; and that parent company may even be an international corporation, like Rupert Murdoch's News Corp. Therefore today local and even national libraries have less leverage in their dealings with news organizations than they did when publishers and broadcasters were smaller in size.

A More Viable Approach

It is clear then that the times require a new template for preserving news. The overwhelming volume, variety and velocity of "big news data" are swamping our preservation efforts, and new solutions are therefore likely to look very different from the trusted methods of the past. But while change comes with risks it is possible that if we do not radically alter course soon, our future constituents will be deprived of the information and documentation essential not only to the writing of history but to the very health of civil society itself.

Due diligence today requires a fresh approach. Such an approach could, aggressively and on an international basis, exploit the "power of the purse." Rather than trying to build the ever more complex and costly systems needed to capture, contain and maintain the digital output of the news organizations it would make more sense for libraries to actively engage the major media organizations in the work of preservation. National libraries or national or international library consortia might broker, on behalf of their constituents, something akin to national site licenses for certain key news databases and services. This would make important news databases available to scholars and other researchers, the proprietors of small businesses and entrepreneurs, and the general public.

While some national libraries have negotiated site licenses for use by patrons on the library's own premises, this author is not aware of an existing nationwide public license for a major news database. Currently negotiating individually or through local consortia, libraries have little leverage with the large publishing conglomerates and multinational corporations that now control news publishing. Greater scale through collective bargaining could provide leverage: a nation's public and academic libraries speaking with one voice through the national library or national-level consortium could not only win favorable access arrangements for constituents but might even influence the ways in which the media organizations maintain their content for the future.[4]

The funding needed for such licenses would be substantial, but not disproportionate to the public investment in bricks-and-mortar library storage facilities made in the past. An investment on that scale would have to be protected, because media organizations often fail. "Future-proofing" would have to be designed to ensure that the applications and technologies needed to support access to orphaned digital content, as well as the content itself, could survive the business collapse of the providers or the obsolescence of the enabling technology for content or platform brought about by owner neglect or abandonment. Perhaps the terms of legal deposit could be refashioned to require, as well as transfer of digital news content, an escrow of the code for the enabling platforms. This would involve a new and different relationship between national libraries and media organizations, and enable legal deposit to once again function as an engine for creating public assets. The licensing of databases at the national level would therefore function as a critical part of the digital preservation apparatus. That would be a gift to future historians.

New Societal Benefits

The radical changes in the news information lifecycle that digital technologies and business models have brought about also have implications for the broader societal functions that libraries fulfill in preserving news. News is an important part of the public record, and independent media play a vital role in providing that record. By facilitating broad public access to that record, libraries have always helped level the playing field for citizens. In an age when much information is privatized to a degree unprecedented since the eighteenth century and income inequality is rampant, that role is more important than ever. In negotiating terms for broad public access, national libraries and consortia would be well positioned to ensure the privacy of users. In the past, U.S. libraries closely held the data they collected on the books and newspapers citizens read and the videos they viewed. Now that so much news content resides in The Cloud, publishers and platform providers are custodians of that data. The premium placed by web business models on user information means that such information is now aggregated, sold and traded in myriad ways. Libraries might negotiate curbs on such uses.

In short, libraries, particularly national libraries, have a civic duty to explore new and alternative ways to preserve electronic news. Without a radical change in approach, the vital public record that an independent news industry and libraries worked together in the past to provide will inevitably erode. By not investing collectively, strategically and boldly in alternative strategies now, libraries will be ceding a critical societal role to the private sector. And the industry consolidation and growth of private equity holdings in the media sector will make that arena a place of dangerously little transparency.

References

[1] For a useful description of the role content management systems play in the electronic news editing and production process, see Luke Vnenchak's June 17, 2014 "Scoop: A Glimpse into the NYTimes CMS." http://mobile.nytimes.com/blogs/open/2014/06/17/scoop-a-glimpse-into-the-nytimes-cms/. Accessed May 16, 2016.

[2] The features of the text-editing tools in *The Guardian's* editorial system are described in Oliver Joseph Ash's March 20, 2014 blog post, "Inside the Guardian's CMS: Meet Scribe, an extensible rich text editor." https://www.theguardian.com/info/developer-blog/2014/mar/20/inside-the-guardians-cms-meet-scribe-an-extensible-rich-text-editor. Accessed April 20, 2016.

[3] For an analysis of the Internet Archive's capture of the New York Times online, see CRL's review of the Wayback Machine in eDesiderata. https://edesiderata.crl.edu/resources/wayback-machine. For a deeper look at the Internet Archive harvesting process, see Kalev Leetaru's 2015 blog post "How Much of the Internet Does The Wayback Machine Really Archive?" *Forbe,s* November 16, 2015, http://www.forbes.com/sites/kalevleetaru/2015/11/16/how-much-of-the-internet-does-the-wayback-machine-really-archive/#5e63e16e88d4, Accessed May 12, 2016.

[4] CRL, an international consortium serving academic and independent research libraries, ventured down this path recently with some limited success. In 2014 CRL negotiated the terms of an academic site license for U.S. institutions for online access to www.nytimes.com. The negotiations were difficult and terms were far from ideal, but to date approximately 174 libraries have subscribed under those terms. Since those terms were negotiated The Times has added to the site its entire digitized archive of articles back to its founding in 1851, and academic users now have unrestricted access to The Times online content.