Investing in the Persistence of News: an eDesiderata Forum

October 4, 2017

Summary Report

As the web and digital media transform the ways news is sourced, distributed and consumed, research libraries must find new ways to ensure long-term scholarly access to important journalism. CRL's longstanding preservation model is built around news in print and microform, and its access model around interlibrary loan. At the 2017 eDesiderata Forum presentations by news media insiders, scholars, publishers, and librarians examined the media and publishing landscape, digital news marketplace, innovative research uses of news. The presentations, and the conversations that followed, provided useful insights on the challenges digital news poses for libraries and suggested a number of ways CRL and its community can address those challenges.

Session I: The Digital News Environment and Marketplace

Conversation 1: Inside the Online News Machine

The news media have built formidable capabilities for managing digital news content throughout its lifecycle. Major news organizations today maintain massive repositories of digital content as a routine part of their business activities. It is difficult to imagine this capability being replicated in libraries. Because that content is not preserved elsewhere, it is important for libraries to understand what goes on inside the systems that enable the management of digital news assets. Unfortunately those systems are often carefully guarded proprietary assets, and therefore "black boxes." This conversation is designed to shed some useful light on them.

Evan Sandhaus, Executive Director, Knowledge & Metadata Management, *The New York Times.*
Administering a 166-Year-Old Database: Tagging, Taxonomy and Better Living through Metadata

Sandhaus reported on the role of tagging in *New York Times* content management. He explained that today tagging is a way of making NYT current and historical content retrievable and reusable in-house as well as discoverable within the Times online archive. Applying tags to articles text, identifying people, places, organizations, titles, and subjects is a continuation of the indexing of the NYT print edition since 1913 and early experiments with information retrieval such as the "*New York Times* Information Service" an ill-fated effort to leverage Times information-gathering and the content of its archives.

The Times archive represents a unique challenge with a total of 60,000 published issues, 3.5 million pages and 15 million articles to date, , The article archive digitized by the Times is currently segmented into three content types: a deep archive of images of articles plus abstracts; a transcribed archive presently covering 1960-1979; and the born digital archive (1980 to present). Therefore tagging facilitates searching across all three content types in the archives.

Aside from its utility for the archives, tagging is incorporated into the news production workflow at the Times, occurring as soon as content is created and entered into the content management system. It also enables automated recommendations for users and news alerts on developments on specific topics, promotes discovery of Times current content by search engines and is used internally for advertising and analytics. The substantial investment in tagging and content enrichment suggests that Times management values its current and historic digital content highly.

Philip Spiegel, Senior Director for Content Management Operations, LAC Group. *Managing and Preserving Broadcast News Assets.*

Digital asset management systems (DAMs) have become critical enterprise infrastructure for broadcasters like Univision, CNN, and NBC. The systems replace the vaults and video libraries of the analog era, which were often not well maintained. Because of the scale and speed at which the news industry works today the DAMs are now indispensable. Media organization operations involve 24/7 content ingest (often averaging 100 hours of video intake daily); continuous creation of new records for video content and updating existing ones; and management of media server space and traffic to and from the digital archive. This is driven by the editorial and production needs of media organizations, and for business needs to curate and repurpose daily content and highlights into easily accessible packages and other derivatives. Metadata applied to the content on source, authorship, ownership, and rights enables intellectual property management and can also preserve the important "chain of custody" of video footage.

In tagging and enriching content with metadata, the emphasis is on the needs of media company internal users, and often systems are customized to reflect an organization's culture, workflows, and internal practices. However, media organizations recognize that such information is potentially useful for researchers as well. Moreover standardization of metadata and digital file formats, critical for preservation, is being driven by consolidation in the news media industry and by centralization of digital asset management at the parent company level ("mother ship location").

Conversation 2: Digital News and Its Scholarly Users

As context for this discussion moderator Mary Feeney, of the University of Arizona Libraries, provided an overview of the scholarly importance of news. News has research value not just for historians: it is also raw material for research in other fields such as sociology, communications, political science, and public policy. Scholarly practice continues to evolve, with new approaches to mining and computer-assisted analysis of large aggregations of news, digitized and born-digital.

Much new research involves real time interaction with web- and social media-based content, with active web sites and other open source content harvested from the *live* web, and with structured data in large databases of aggregated news, like *LexisNexis* and *Factiva*. Increasingly, researchers download content to analyze text for trends and biases, to map social networks¸ and to produce visualizations that are not possible with print and microform, and are limited even with commercial databases. It appears that in these kinds of large-scale analyses relatively little use is being made of *archived* web content, like that harvested by the Wayback Machine and by the national libraries.

Nick Adams, Berkeley Institute for Data Science. *Mining Online News and News Data for Insights on Political Affairs and Public Awareness.*

Adams reported on three projects in which computer-assisted analysis of large bodies of news text on events was used to detect patterns and trends in the behavior of police and protesters, and bias in news reporting about those events. His analysis of news coverage of the 2011 Occupy Movement used Google searches of online articles on open source national, regional and local news sites. Adams observed that researchers like him are building their own archives, creating large local databases of digital content that can be navigated and queried for their own purposes and can interact with new tools, like CapitolQuery and Text-Thresher, created at the Berkeley Institute for Data Science.

Adams pointed out that this work is not fully automated: the projects adopted a hybrid approach, employing both humans and machines, but he observed that machines are "more reliable, reusable, and scalable" for projects that involve mining thousands of news articles and applying hundreds of variables.

James Danowski, Department of Communication, University of Illinois at Chicago (retired). *Mining Electronic News for Political and Social Science Research*.

Danowski discussed three studies of his own employing extraction of news documents for analysis of political communications:

1. "Presidential Communications Network Centrality and Job Approval in the Obama Administration,"
2. "Changing Semantic Networks for Jihad among Majority-Muslim Nations Before and After the Early Arab Spring Uprisings"
3. Ivory Coast Community Resilience Study after Muslim-Christian Civil War: Examining the Effects of News Content on Recovery from Conflicts

The first study involved automated extraction of data on a network of political actors--i.e., the Obama cabinet--from news reports. Using text extracted from the *LexisNexis Academic* database, Danowski analyzed reporting in *The New York Times* and *The Washington Post* over a period of 204 weeks on two events: the Debt Crisis and the Libyan Civil War. The other studies used translated transcripts of web pages, broadcasts, newspaper stories, and other textual material from the *LexisNexis World Publications* and *BBC International Monitoring* databases and mobile phone data provided by an international telecommunications company, to study social networks among communities following major social upheavals.

Problems encountered with the commercial databases included limitations on the number of search results, often to 1,000 documents (which required "slicing up" searches into smaller units); download limits; and the need to strip out extraneous metadata words and multiple copies of the same reports in search results, using home-made software.

Issues encountered in analyzing live web news content included the need to rely upon third-party analytical tools like Twitter APIs to extract data, which lack clarity regarding the selection of datasets and algorithm biases.

Conversation 3: The Changing Library Supply Chain for News

Patrick Reakes, Associate Dean of Scholarly Resources & Services, University of Florida. *Acquisition of Print and Microform: Shifts in Library Collection Strategies*

The experience of the University of Florida in adapting to changes in newspaper production and researcher practices is typical of research libraries in general. There is the reduced acceptance of microform. Even researchers doing historical work in print sources now want source materials in digital format, and have little tolerance for the interlibrary loan delivery cycle. Many users are now doing content analysis and in- depth research, requiring tools that interact with digital but not analog content. Moreover, current awareness, formerly a justification for library newspaper acquisitions, is now satisfied by access to real-time news on the web. Meanwhile, commercial databases can have a high cost per use, which undermines their viability as substitutes for collecting the news titles in print.

In response, Florida libraries are making reductions in film and print subscriptions. The reductions at individual libraries are not entirely compensated by statewide cooperative arrangements that distribute the responsibility for collecting film, which come with their own risks.

Moreover due to a lack of transparency, there is considerable confusion as to what is in the full-text databases (like *Factiva*, *LexisNexis*) versus the page-image aggregations (like *ProQuest Historical Newspapers*). There is no consistency, moreover, across print copy, web copy, commercial database, and archives on newspaper websites, let alone clarity about how much news comes from social media and non-traditional news sources, or on how libraries might address “fake news”.

In a time when advertising and subscription revenues are dwindling, publishers are especially covetous of their content and resist giving libraries the right to ingest and digitize content for academic users. The University of Florida’s experiments with the *Florida Alligator*, a local paper with a large circulation, has had some success: the University once collected and microfilmed print copies; then turned to digitization, and recently began ingesting born-digital ePub (PDF) copies. Yet the web-only content is not being captured.

Mark Sweeney, Associate Librarian for Library Services, Library of Congress. *Digital Developments in News Preservation at LC*

LC’s partnership with the National Endowment for the Humanities on the United States Newspaper Project was a sustained cataloging and microfilming effort that identified over 140,000 domestic newspaper titles. Of those, the Library holds 10%, unlike European national libraries which tend to hold all newspapers. Today only 400 U.S. titles are still being collected by LC, most of which are received in microform.

LC’s foreign newspaper preservation efforts reformat more than four million pages a year on microfilm, many through LC’s overseas field offices. Altogether LC acquires 874 newspapers from 155 countries or geographic areas in 81 languages.

LC is seeking to transition from a microfilm-based acquisition program to a digital one. This will involve having vendors and publishers provide newspapers in ePrint or PDF format. Lack of uniformity in formats and metadata standards in the industry will make this challenging.

Guided by LC curators and, to a lesser extent, Congressional Research Service interests the Library has also been actively harvesting and archiving news websites since 2000, covering mainstream media and partisan publications, including "fake news" sites as well. However, the major, dynamic news sites (*New York Times*, CNN) are problematic for web crawlers. LC is archiving mainly smaller U.S. newspaper sites, such as *Huffington Post* and Breitbart. The harvested content is embargoed from public display for at least one year, and thereafter displayed only when written permission is granted by the publisher.

Dorothy Carner**,** Head of Journalism Libraries and Adjunct Professor, Missouri School of Journalism. *The State of Legal Deposit Abroad*

Carner reported the findings of a recent survey of legal deposit laws, conducted in 2014 and updated for a presentation at the IFLA conference in 2017. Responses were received from 40 respondents in 26 countries, 23 of which have legal deposit laws. Findings include:

- While most responding countries' legal deposit laws now include digital deposit , only nine survey respondents indicated that they require publishers to deposit digital works, many of whom limit such deposits to specific types of materials like online-only journals, and do not necessarily include newspapers.

- Eleven libraries indicated that they harvest news websites/webpages, with frequency of capture varying from daily to hourly to whenever content is updated. Methods of capture include crawling and scraping, taking submissions through physical delivery on hard drives or CDs, and capturing article content from RSS feeds.

- Use of the archived news content is minimal. Most libraries provide only onsite access to deposit content due to copyright and/or the terms of publisher agreements; a few provide offsite access after an embargo period; and very few provide immediate offsite access. At the time of the survey Sweden, the most advanced in terms of technical archiving and metadata systems, had no provisions for onsite or offsite access.

- In the UK, a project funded by the Arts and Humanities Research Council is assessing the impact of e-legal deposit on contemporary research. The project, "Digital Library Futures: The impact of E-Legal Deposit in the Academic Sector" (https://www.uea.ac.uk/e-legal-deposit/), will study how legal deposit collections are accessed and used and how those collections might support contemporary research in academic libraries.

Carner concluded that we are probably several years away from scholars having meaningful remote access to digital news through legal deposit. She added that while libraries search for the best model for born digital news preservation, important content is disappearing.

Session 2: Strategies and Models for Library Investment in News Access

Conversation 1: Site Licensing of Major Online News Sources

Ann Okerson, CRL Advisor on Electronic Resources Strategy. *Use Cases and Models for CRL's Site Licensing of Current e-Newspapers*

Alternatives to print-based news collecting and preservation are now emerging. Direct engagement with news producers offers a potential way for academic libraries to support local research and teaching while shaping the news marketplace. Ann Okerson recounted CRL's recent efforts to implement academic site licenses with major news producers.

For newspaper publishers print business models are no longer sustainable: the decline of print advertising revenue and rise of social media are pressuring publishers to give away much of their content, resulting in a loss of depth and quality of reporting. CRL is looking for ways to preserve quality news through site licenses with a few major news providers. A national academic site license for nytimes.com, negotiated in 2014, was the initial effort. Despite certain limitations, the license provides web access to Times born-digital content as well as a historical archive of all *New York Times* articles back to issue number one in 1851.

About sixteen U.S. consortia now participate in the CRL offer, providing an estimated $2 million in revenue to The Times. CRL negotiations with *The Wall Street Journal* have now resulted in a similar arrangement, details of which are on eDesiderata at https://edesiderata.crl.edu/resources/wall-street-journal. Both the Journal and Times publishers fear the academic licenses siphoning off revenue from existing individual subscribers and from aggregators like ProQuest and LexisNexis. To strengthen its bargaining position, CRL is exploring with Canadian Research Knowledge Network, Deutsche Forschungsgemeinschaft, and Jisc the potential for international action on digital news access.

Conversation 2: Direct Library Investment in News Access and Preservation

James Simon, CRL Vice President of Collections and Services, reported on CRL's 2015 assessment of news digitization by libraries and trusted publishers, which highlighted the limitations and drawbacks of those efforts. The vast majority of the materials digitized as of 2015 were older, public domain newspapers, and most had already been preserved on microfilm.

CRL's *World Newspaper Archive* initiative with Readex is a community controlled project that has digitized 3.5 million pages of content to date. The content will eventually to be hosted in open access mode by CRL. Reveal Digital's *Independent Voices* represents another model for partnership with a commercial service provider to produce open access digitized news resources.

Bryan Benilous and Robert Lee, East View Information Service. *East View Information Service's Proposed Global Press Archive*

Benilous and Lee spoke about East View's work with Stanford University Libraries to create a Global Press Archive. While most commercial publishers' newspaper digitization efforts focus on English language, public domain content, the Global Press Archive is intended to make accessible more than 2,500 titles and 40 million pages of newspapers from 125 countries and in 30 languages in a highly curated collection.

CRL is considering possible terms under which members might invest in the project. Challenges include the substantial cost of reformatting such a massive body of paper materials and the formidable task of securing rights to digitize the titles, most of which date from the second half of the twentieth century. A possible by-product of the project would be significant enlargement of CRL's ICON database, which would go a long way toward producing a much-needed union catalog of non-U.S. newspapers.

Clifford B. Anderson, Associate University Librarian for Research and Learning, Vanderbilt University. *Sustaining Television News for the Next Generation.*

Anderson reported on a project funded by The Andrew Mellon Foundation that will examine and address the challenges of library-based preservation of news broadcasts, which are essential sources for information about the world and an important public record. The Vanderbilt Television News Archive is the longest running initiative of its type in the U.S, for the past 50 years. Today online news has changed the broadcast news landscape, presenting significant legal, technical and funding challenges to sustaining the preservation efforts. The challenges include researcher resistance to offline access and analog formats, an accelerated, 24-hour news cycle that requires automated, asynchronous capture of broadcasts, and growing digital file storage requirements. The recently announced Mellon Foundation grant will support an effort by CRL and Vanderbilt to identify economies through synergies with other library-based efforts, like the UCLA Library Broadcast NewsScape project and the Library of Congress-WGBH American Archive of Public Broadcasting.

Next Steps: Library Investment in News Access and Preservation

The annual eDesiderata Forum is a venue for sharing expertise and insights on the changing needs of researchers and on the digital resources that support research. The Forum's ultimate purpose is to inform CRL's collecting and licensing strategy. It was clear from the 2017 Forum conversations that digital news is an entirely different animal, and that today's research presents an entirely new set of preservation challenges for research libraries, and for CRL. How CRL will adapt to those challenges is the topic of a new blog post, *Investing in the Preservation of News: What We Learned from eDesiderata 2017.*