Submitted on: 8/4/2014

Missing links: The digital news preservation discontinuity

Dorothy Carner
University of Missouri-Columbia, Columbia MO 65211, USA.
carnerd@missouri.edu

Edward McCain
Donald W. Reynolds Journalism Institute, University of Missouri-Columbia, Columbia MO 65211, USA.
mccaine@rjionline.org

Frederick Zarndt
Global Connexions, Coronado CA 92118 USA.
frederick@frederickzarndt.com

Abstract

That the spread of printed news has changed dramatically since the Internet and the Web is no news to anyone. The Christian Science Monitor, in print since 1908, ceased daily publication in 2009 to focus on web-based publishing (CSM still publishes a weekly print edition). One month before this, The Seattle Post Intelligencer stopped its print edition. More recently, Lloyd's List, which claims to be the world's oldest newspaper, announced that it would stop its print edition. These are but a few examples of news publishers that no longer print the news on paper.

While these newspapers stopped printing news, they did not stop publishing news. Instead they now concentrate on digital news.

Similarly and until only recently, the IFLA Newspapers Section has focused on cataloguing, collecting, and preservation of printed news. With few exceptions, Section members do not catalog, collect, and preserve digital news with the same diligence as they have in past given to newspapers.

In this paper we will review digital news publishing, both for traditional news publishers like the Christian Science Monitor and the Seattle Post-Intelligencer and for digital only publishers like The Huffington Post, The Texas Tribune, NewsWhip, and others. We will especially look at the publishers' digital preservation policies and practices.

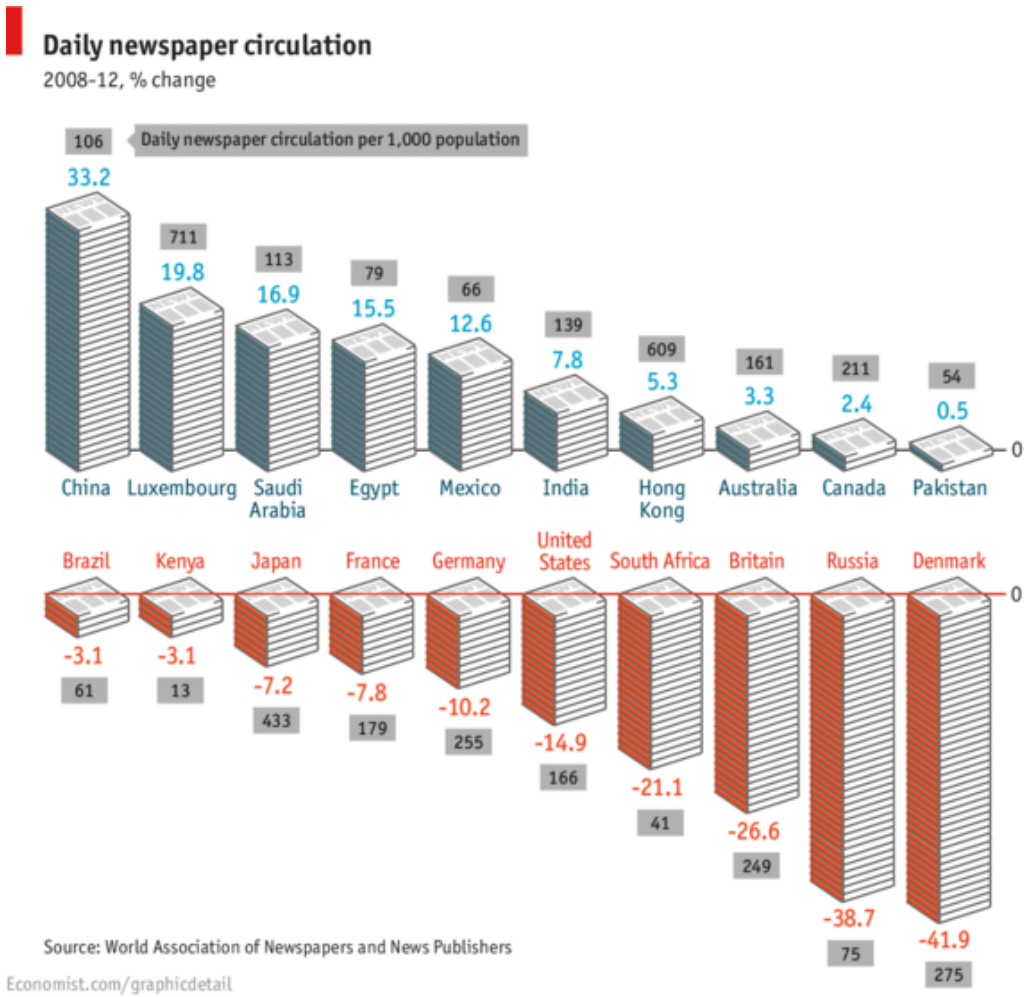
The result? For that you will have to read the paper or listen to the presentation. But you won't be surprised if we hint that there is a humungous collection, catalogue, and preservation gap between the printed and digital.

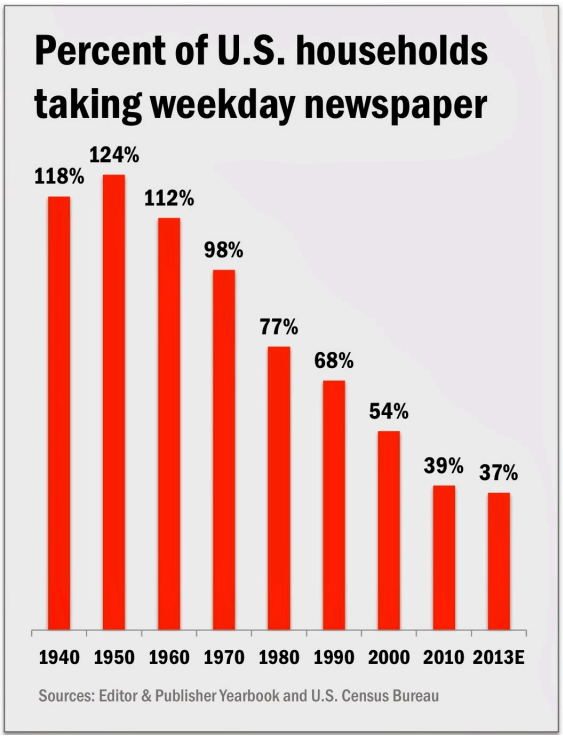
Keywords: *digital preservation, digital news, survey, news publishers, web harvest, born digital news, e-legal deposit.*

Introduction

The Internet, the World Wide Web, and the technologies built on this digital infrastructure have made possible news delivery in ways that pre-21st century news publishers did not imagine. No longer is news first created on media like paper or analogue video / audio tapes: Instead it is created digitally – "born digital"-- and this born digital news is used to create analogue media like newspapers, magazines, non-digital radio or television. Or born digital news is *never* transferred to analogue media and only exists digitally.

Long-running and authoritative newspapers like *The Christian Science Monitor* have stopped or drastically reduced the amount of news published on analogue media (newspapers). Curtailing printed news production and distributing ever more news digitally via websites, social media, YouTube, Twitter, etc. has been a trend which began early this century and continues at an accelerating pace, at least in most of the developed world.





In the United States policies and practices for the preservation of newspapers have been well developed. The National Digital Information Infrastructure and Preservation Program (NDIIPP) ensured newspapers' preservation by copying them to microfilm, which, under proper storage conditions, may last as long as 500 years. And although NDNP historical digital newspapers should not be viewed as a preservation medium, the National Digital Newspapers Program (NDNP) uses microfilm from the NDIIPP program to provide access to newspapers now in the public domain.

Unfortunately there is no NDIIPP or NDNP program for born digital news, at least not in the United States.

Outside of North America, born digital legal deposit policies and practices vary considerably from country to country: Some have well developed born digital legal deposit policies and practices while others find themselves in situation similar to the United States, that is, without well articulated policies and practices for born digital legal deposit. Lack of policy will likely result in a "preservation gap", a period of time when born digital news is simply lost for any number of reasons – disk crashes, bit rot, loss in the migration of news from an old software system to a newer one, etc. News producers may of course preserve their own born digital news, but as you will see below, producers' preservation efforts are often not well.

This paper tells the history of news preservation at producers and at cultural heritage organizations. It discusses the results of a survey of news producers' digital preservation practices conducted by the Reynolds Journalism Institute. The news producers surveyed for this study are all North American, but as with all digital media, the reach of the news they publish is global. We expect that a survey of news producers outside of North America would give similar results, but we leave this to someone else to show.

In a follow-on paper, we shall discuss the results of a survey of born digital news legal deposit policies and practices at cultural heritage organizations around the world.

Literature Review

While the marketplace rewards breaking news, managing previously published news content has historically been someone else's problem, most often a librarian's.

To provide context to the born-digital news preservation survey results cited in this paper, the authors follow the development and management of news archives from "morgue" to born-digital repositories, looking for evidence of intention to preserve along the way.

The first American scholar to investigate the penetration, preservation and role of newspaper archives, often called morgues by newspaper staff, was Robert W. Desmond, a professor of journalism at the University of Minnesota. He wrote:

> Originally, so far as concerned the preservation of information, the only material placed away for future use were prepared obituary sketches of prominent men. It was because of this fact that the newspaper libraries and reference departments were originally given the grim name of "morgue," a name that the best of them have long outgrown (Desmond, 1930, p. 16).

His research was an attempt to change the prevailing attitude that the morgue held "dead" information and was only valuable for writing obituaries. Nathan (1910), twenty years earlier, said that the word morgue "stands for the department of the paper wherein are kept the keys to the news that has passed, the "dead" news, in other words (Nathan, 1910, p. 597).

Desmond was one of the strongest advocates for what he called the newspaper "reference library," the information they contained, so valuable for fact-checking and historical documentation, and the librarians managing them.

Desmond (1930) claimed that some U. S. newspapers had small morgues in place around 1850, but reference collections began to take shape after 1900. He believed that it had been the wars (Spanish American and World War I) that provided the impetus for creating morgues/libraries, because newspapers needed to gather information on a wider range of topics than simply news about prominent men. Newspapers were no longer hyper-local. Because American soldiers were fighting in global conflicts, news from distant parts of the world became important to cover (Desmond, 1930, p. 16).

According to Desmond, *"The New York Herald*, which began publication in 1835, started to build up a library of books about 1845. Clippings were saved beginning in 1852, but not systematically until about 1860 (Desmond, 1930, p. 17).

James Gordon Bennett, Sr., founder of the *Herald,* has been credited with establishing the first counterpart of the modern newspaper library in the United States. In Desmond's book, *Newspaper Reference Methods* (1933), he speculated that the Civil War inspired Bennett to begin keeping an index in 1860. Bennett considered it a valuable enough enterprise to retroactively index the *Herald* back to the first issues of the paper in 1835 (Desmond, 1933, p. 3). However, this was not the first U. S. paper to be indexed. *The Lathrop Report on Newspaper Indexes* (LRNI) indicates that the *Maryland Gazette* (Annapolis MD 1727-1746) was the first to create an index (NLE Index, 2014).

Lydenberg (1915) was the first academic to write about preservation practice for newspapers. In 1914, the New York Public Library (NYPL) was concerned about the severe wear that their bound newspaper volumes were experiencing based on excessive use and the poor paper stock on which news was printed. NYPL librarians experimented with various preservation practices for more than a year. Chemists in Italy and Germany had created "transparent compounds that when applied to paper overcame its brittleness, kept it air-proof, and for ordinary use were said to make it practically imperishable." Also known to protect edges of bound volumes, was a chemical preparation created in Brooklyn that would keep air from volumes only if they were not used. This did not seem the best process since the volumes were in constant use. Only if two sets were acquired (one for use and the other for preservation) would it be an adequate solution (Lydenberg, 1915, p.240).

These early news preservationists at NYPL concluded that binding material could be applied to the sheets that would exclude air. They experimented with Japanese tissue and silk, treating some volumes (applying the tissue and silk with "pure" rice paste) and leaving other volumes without treatment as the control group. They exposed the sheets to the sun for periods between 100 to 150 hours. The unprotected paper turned yellow rapidly; those protected with silk, less rapidly and those treated with Japanese tissue showed only a slight effect. Procedures included using Mullen and Ashcroft tests, standards used for testing paper strength, and the United States Bureau of Standards test for bending and pliability. They also applied muslin to individual newspaper sheets, and because it increased the volume threefold, was easily ruled out as a potential preservation method. Even though the Japanese tissue tests proved to be satisfactory, the cost was prohibitive. They estimated that it would cost $35 per volume or $420 per year for one daily to be treated. That translates to $10,732.04 per year in 2014, based on CPI inflation calculation (Calculator.net, 2014). The trustees felt that the cost was too much and hoped to get publishers to print special editions for the libraries using better paper stock.

Current archivists and digital curators can relate to the research performed by these early preservationists and the predictable outcome that prohibitive costs would ultimately derail attempts to preserve newspaper content for future generations. The optimism they demonstrated in their suggestion that local papers print special editions for the libraries using better paper stock is reminiscent of similar efforts to elicit cooperation from today's publishers of news content.

"Because of the evident need," wrote Desmond in 1933, "for a system of filing suitable to growing newspaper libraries and the desirability of discussing common problems, a group of newspaper librarians was organized in 1923 as the Newspaper Group of the Special Libraries Association" (1933, p. 5). Joseph F. Kwapil, librarian at the *Philadelphia Public Ledger* and the group's first president, sent out the call to meet.

Will C. Conrad, *Milwaukee Journal* librarian, believed that the newspaper library "has become the repository of the materials from which the news story, and the editorials if you please, are built. It has become the most important factor in the news plant, the very center of the plant" (S*pecial Libraries*, Oct. 1928, p. 268).

Probably the best example of a good newspaper library in 1930 was the one Kwapil managed at the *Philadelphia Public Ledger.* The library not only served *The Public Ledger*, but also the *Evening Public Ledger*, *The Saturday Evening Post, The Ladies' Home Journal* and the *Country Gentleman*. It was open 24 hours a day. The staff filed over 1000 clippings each day. The dailies used about 30,000 photographs and cuts taken from the library each year. About 100 newspapers from all over the country were read and clipped daily. In 1930 there were several million clippings on file, with about a million subjects covered with over 20,000 guides and finding aids and more than 80,000 classifications in the subject file. The library had more than 2,000,000 photographs, 85,000 cuts and over 100,000 negatives and employed one librarian and 14 assistants (Desmond, 1930, p. 37). Kwapil (1921) also established a daily index at the *Ledger* (Kwapil, 1921, p. 445), and was instrumental in creating a standard classification system for news libraries since neither the Dewey Decimal nor Library of Congress Classification systems worked well in a newspaper library (Desmond, 1930, p. 44).

Early twentieth-century newspapers began assigning a monetary value to their archives. "In 1924 the *St. Louis Post-Dispatch* collection was estimated to be worth "several hundred

thousand dollars." The *Detroit News*' collection, appraised by a committee of underwriters, was valued at $1,000,000 (Desmond, 1930, p.3). According the Bureau of Labor Statistics CPI inflation calculator, that same collection would be worth nearly $14,000,000 today.

While we continue to grapple with appropriate taxonomies, workflow and access tools, struggling with costly digital storage space, turn-of-the-century news librarians faced similar issues in the analogue world. "We have two chief problems in the reference department," one news librarian complained, "one is the problem of space – getting all of our material into the available space. The other is the problem of how to file information quickly and efficiently and accurately and how to find it quickly when needed" (Desmond, 1930, p. 47).

During the last two decades of the nineteenth century, smaller newspapers began to rely on clipping services rather than hire staff to create and manage a morgue. Robert Luce founded one such service, the Luce Clipping Bureau, in 1888. By the turn of the century, several of these services began to appear. Some offered vanity services to prominent people wanting a record of when and where their activities were published. News librarians were often critical of these commercial clipping services, charging lack of discrimination and poor quality of service, but the cost of staffing a newspaper library with enough people who could read, clip and file hundreds of stories daily, made these services a popular method of "outsourcing" morgue work (Luce, 1913, p. 152-157).

According to McCargar (2011), The Associated Press (AP) "was created to solve a nineteenth-century technology infrastructure problem–the limited and expensive availability of the telegraph during the Mexican-American War" (McCargar, CRL, 2011, p. 5). What began in 1846 as a collaborative service gathering and distributing news content to five New York newspapers, has expanded to include over 1400 U. S. daily newspapers and thousands of television and radio broadcast members (AP, 2014). Not only does the AP provide content to its members, it has been instrumental in creating and providing them access to metadata, enhancing efficiency of storage and access of their news content.

Castells et al. (2004) believed that the introduction of information technologies during the latter part of the twentieth century transformed the news industry and thus their libraries. Adoption of integrated platforms supporting the whole workflow cycle has become the trend over the last two decades. As a consequence of this transformation, a new market of online services for archive news redistribution, syndication, and aggregation has emerged, providing value to a wide range of information consumers (Castells et al, 2004, p. 446).

Hegg (1991) purports that the introduction of the *New York Times Information Bank* (NYTIB) in the 1960's marked the beginning of a massive technological disruption to newspaper libraries and librarians. The NYTIB offered a computerized index and abstract to articles in its publication. By the mid seventies a few newspapers had created hybrid systems involving computerized indexes with news articles available on microfilm. The early eighties saw others create their own full-text retrieval system. Systems could provide both power and speed, satisfying information retrieval needs of journalists and editorial staffs. Librarians could create specialized files and indexes that before computers were not previously available (Hegg, 1991, p.5).

Vittolino (1985) wrote that large newspapers were the first to adopt automation technologies that helped create the electronic news library. Local investigative reporters created in-house databases allowing reporters to analyse "hot issues." New York Times, Inc. (NYTIB), the

Knight Ridder Group (Philadelphia Newspapers Inc.), the creators of VU/TEXT as a result of converting PNI's newspaper clippings to digital ones, and the Oklahoma Publishing Company (DATATIMES) were dominant in the early years of newspaper automation.

Some current news vendors: NewsBank's Readex unit (McCargar, 2006, p. 1) and University Microfilms International (UMI), now ProQuest LLC (McCargar, 2008, p. 3), began as microfilm services for newspapers.

Mead Data Central's Nexis product, for years considered the "gold standard," subsumed the NYTIB in the early eighties and was not only an industry vendor, but began marketing to libraries, local subscribers and even rival newspapers. Librarians at these newspapers no longer clipped stories, but sat at video display terminals (VDT) and enhanced the computerized version of the articles providing metadata for easier access. Smaller newspapers with smaller staffs and fewer resources were slower to automate, but the floodgates had been opened and automation would prove to be the tipping point for news libraries, librarians and news preservation (Hegg, 1991, p6).

In 1982, Investigative Reporters and Editors (IRE) conducted a survey of metropolitan dailies to determine their level of automation. Twenty large newspapers were already using commercial databases for information retrieval (Ward & Hansen, 1982).

In the summer of 1983, then IRE director, John Ullman, surveyed newspapers with a daily circulation of 100,000 or more. The survey was designed to find out which papers were using databases, which of the 1133 commercial databases available from 189 vendors (listed by *Cuadra Directory of Online Databases*) were used and their level of satisfaction with the services (Ullman, J. 1983).

Information Today (1984) reported that by "September 1984, more than 2,400 databases are available through 345 online services worldwide. Cuadra Associates' statistics indicate that the number of online databases available is growing at about 35% annually, while the number of online services is growing about 40% annually."

The Commercial Databank Survey by the SLA Newspaper Division in 1984 identified 39 newspaper libraries using databanks, with 80 newspapers performing online searches, a growth of 100% every two years. (Ward & Hansen, 1986).

In 1986, a survey was sent to 456 members of the SLA News Division with the primary goal to collect information about news librarian salaries and to see if there was a correlation with size of newspapers. Their secondary goal was to see how many newspaper libraries were automated. "Of the 155 responses from newspapers, 51 libraries (32.9%) indicated they were online with at least one system. Of those libraries, "39 were located in newspapers with circulations greater than 100,000." An additional 23 newspaper libraries indicated that they planned to install an automated system within the next 12 months.

Mead Data Central's Nexis was the databank most often used for information retrieval in 1984, followed by Dow Jones News Retrieval, DIALOG, VU/TEXT, Datatimes and four others. By 1986, there were as many newspapers subscribing to DIALOG as there were to Nexis (Hunter, J. 1986, p. 5).

The first issue of the *Cuadra Directory of Online Databases*, published in 1979, described a total of 400 online databases offered by 59 online services. In 1991 the number of online databases available had grown to over 5000 (Cuadra, 2014).

Soffin (1987) wrote about the emergence of national services, which compiled "electronic morgues" from a variety of newspapers. He said that newspapers were faced with two choices: "whether to subscribe to any and whether to contribute their newspaper content to an online service." For many smaller newspapers, they saw the one major benefit of contributing stories to an online database as serving a function of an electronic morgue, eliminating the need to set up and operate their own system. However, the primary disadvantage, at the time, was the cost of accessing stories, even their own (Soffin, 1987, p.4).

With increased automation came self-sufficiency. The perception of the librarian as gatekeeper began to change. Reporters could find their own information. News libraries and news librarians would begin to disappear in all but the largest newspapers, while technology increased. Few news organizations considered the high risk of digital content loss that occurred as the technology evolved and systems changed. Tom Warhover (2011), executive editor of the *Columbia Missourian* wrote:

> Technologies change. For instance: When the *Missourian* migrated from one content management system to another, three years (2001-2004) of digital articles were corrupted. Stories weren't lost, but all the paragraph marks flew away. *Missourian* librarians had to put them back in, by hand. Another local example: Massive failure. *Missourian* librarian Nina Johnson says the files from 1986 to 2002 were lost because of a server crash.

In the early twenty-first century, independent aggregators harvesting news content from the Web, often without the consent of the content producer and in violation of copyright, push content to our desktops and mobile devices. News publishers have pushed back with pay-walls. The worry that someone else will profit from their copyrighted content has deepened the resolve of many publishers to tighten control of their digital content. So when memory organizations approach them about handing off their content to be preserved, there has historically been little trust.

News organizations inclined to handing off content to memory organizations for safekeeping are often not in favor of providing public access to content still under copyright. Zarndt, Boddie, and Lanz note, in their 2010 IFLA presentation, demonstrated that public access to news content still under copyright, can be avoided with the proper software. They share how repository software can be configured to allow access to copyright free news while sequestering news content still under copyright protection. (Zarndt, Boddie, & Lanz, 2010, p. 93) In short, copyright is often an excuse used by news content producers not to preserve or provide access.

LeFurgy (2005) writes about the establishment of the National Digital Information Infrastructure and Preservation Program (NDIIPP), created to address the growing concern that the United States was at risk of losing much of its digital heritage. In 2000 Congress enacted the National Digital Information Infrastructure and Preservation Program (NDIIPP) with the Library of Congress tasked with creating a strategy for implementing it. "Public Law 106-554, providing up to $100 million of funding, was authorized to support NDIIPP,

with $75 million contingent on a dollar for dollar match from non-federal sources" (LeFurgy, 2005, p. 164). The National Digital Newspaper Project (NDNP) and the National Digital Stewardship Alliance (NDSA) were both established based on work emerging from the NDIIPP.

From 1982-2011, the National Endowment for the Humanities (NEH) with the cooperation of the LOC conducted a cooperative effort among states to locate, catalog and microfilm newspapers published in the United States. This is yet one more effort; initiated by memory organizations and librarians to ensure that newspaper content is made available to the citizenry (NEH, 2014). Since 2005, the NDNP, a partnership between the LOC and NEH, has been awarding grants to state libraries, historical societies, and universities for the purpose of digitizing historically significant newspapers in the public domain. The digitized pages are mounted on the Library of Congress' *Chronicling America* site (LOC, 2014).

As a result of the understanding and intervention by NEH, LOC, state organizations (universities, historical societies, etc.), projects and institutes interested in preserving the nation's cultural documents (including newspapers), have initiated programs, created tools and workflows to assist those new to digital preservation. The Educopia Institute, with the assistance of NEH funding, introduced the *Chronicles in Preservation* project from which a set of guidelines for creating readiness activities for institutions wishing to engage in the preservation of digital newspapers was conceived (Krabbenhoeft, Skinner, Schultz & Zarndt, 2013).

The Blue Ribbon Task Force (BRTF) on Sustainable Digital Preservation and Access (2010), comprised of a variety of interested stakeholders, focused its inquiry on materials that were of long-term public interest with diverse preservation profiles. Their interest was in the economic sustainability of preserving products of scholarly inquiry, research data, collectively and collaboratively produced Web content, and culturally significant digital content owned by private entities and protected by copyright. Included in their analysis, was how best to address the issue of preserving digital news content and the challenges that come with that task (BRTF, 2010, p. 1). The BRTF concluded that incentives to preserve in the public interest must be provided and stakeholder roles and responsibilities over the digital lifecycle must be defined (BRTF, 2010, p. 71).

Bernie Reilly is President of the Center for Research Libraries (CRL), an international consortium of university, college, and independent research libraries. Founded in 1949, CRL supports advanced research and teaching in the humanities, sciences, and social sciences by preserving and making available to scholars the primary source material critical to those disciplines (CRL, 2014). With preservation and access to historical content as driving forces, CRL continues to build invaluable relationships with owners and distributors of digital news content. Collaborations with NewsBank and ProQuest, aggregators of news content and universities with unique collections (Latin American and African newspapers) have resulted in new news products with value pricing structures for CRL members.

Access to news content is moot if it is lost. In Alverson et al (2011), CRL tasked authors concluded that to devise "… effective strategies for preserving news in the electronic environment there must be an understanding of the "lifecycle" of news content. "…CRL proposed to examine, analyse, and document the flow of news information, content, and data for four major newspapers from production and sourcing, through editing and processing, to distribution to end users" (Alverson et al, 2011, p. 4). It was the authors' hope that their

report might pinpoint the "high-impact point of entry" in the workflow where libraries and other memory organizations could "capture critical news content and metadata and ensure the long-term survival and accessibility of the American journalistic record (Alverson et al., 2011, p.5). The results, though not surprisingly, showed little consistency among workflows and that a high degree of normalization would be necessary for successful ingestion into and access from preservation level repositories.

"In the age of print research, libraries have played an important role in preserving newspapers for scholarly research", Reilly (2011) described the challenges of preserving electronic news. "…with the current ascent of digital media as the locus of news reporting and distribution, however, traditional preservation models will no longer ensure future access to a comprehensive journalistic record" (Reilly, 2011, p. 54).

The 2011 CRL research project also revealed challenges with current e-facsimile page-image files (pdf, etc.). Exponential movement in the adoption of mobile devices does not bode well for image files having only minimal metadata attached. Reilly suggested that publishers might "export a uniform XML package at the issue or article level, perhaps captured on output from the pagination or editorial system." He pointed out that this "is the moment in the lifecycle of the news item when the annotation is richest and the data most highly structured" (Reilly, 2011, p. 54).

Reilly insisted that a viable approach to preserving electronic news would have to involve cooperation between libraries and newspaper producers, publishers and aggregators. 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" (Reilly, 2011, p. 55).

Government regulations can provide incentives to preserve. Legal deposit of newspapers is common throughout the rest of the world. This is usually one of the missions of a country's national library where each copyrighted periodical must be deposited. For some, like the United States' Library of Congress, this may mean all of the issues within a single month, not every issue published. Most national libraries or the national memory organization tasked with collecting this content have mechanisms for ingesting analogue content and microfilm, but very few have been tasked to acquire digital content. Those harvesting e-content face challenges often arising with evolving technologies.

The International Federation of Library Associations (IFLA) has advocated for legal deposit as an incentive to preserve for years. In 2001, they published a *Statement on Legal Deposit*: "Legal deposit is critical for the preservation of and access to a nation's documentary heritage. Publishers and libraries work together to ensure the worldwide success of legal deposit of content, irrespective of format or technology" (IFLA, 2011).

The International Internet Preservation Consortium (IIPC) Netpreserve.org, another advocate for digital preservation, lists countries with and without legal deposit laws (IIPC, 2014).

More than a century of literature indicates that librarians, archivists and preservationists have been those most concerned with safeguarding the newspaper as cultural record. News organizations have consistently cited too many other concerns (competition, economics, copyright infringement, technological changes) to worry about preserving news content for the public good. Although some may understand the risks of inaction, the lack of expertise in

and understanding of preservation practices; the fear of losing ownership of their content; overwhelming costs; the lack of incentives, and the absence of established trust, combine to keep them from preserving their news content.

As McCargar (2011) suggested in her white paper, *A mandate to preserve*, a *summary of the 2011 Newspaper Archive Summit* held at the Reynolds Journalism Institute, 'the discussion must be elevated above "it's a library problem." It's also a publishing problem, a journalism problem' (McCargar, 2011, p.5).

It is the hope of the authors that this survey and the "Dodging the Memory Hole" conference, scheduled for November 2014, also at the Reynolds Journalism Institute on the campus of the University of Missouri, can start some conversations, build trust, create incentives, and form public/private partnerships that finally lead to positive action.

Survey

Over the past four decades the transition from analogue to digital systems in news media has transformed the way journalistic content is produced and accessed. Like other creators of "born digital" content, news media have employed a series of evolving file formats and technical infrastructures, many of which are now – or may soon be – obsolete. All born-digital content has proven to be fragile: susceptible to bit rot, media failures and human-caused and natural disasters.

There are a number of challenges in this area: the enormity of the amount of content, or data, that is being generated, the disparate nature of news content formats, numerous barriers to acquisition, the need for better discovery systems and a general lack of understanding about the urgent need to save this content. These and other challenges place a large percentage of the world's news content at significant risk.

In April 2011, the Newspaper Archive Summit, held at the Donald W. Reynolds Journalism Institute (RJI) in Columbia, Missouri, convened a diverse group of stakeholders including librarians and news publishers with the purpose of having a conversation about how to ensure the preservation of newspaper content for future generations. From November 10-11, 2014 we will convene a follow-up event, "Dodging the Memory Hole: Saving Born-Digital News Content" narrowing the focus to born digital news. The goal is to produce an actionable plan to manage and preserve this irreplaceable content.

In an effort to inform this group's work, we plan to learn more about how producers of born digital news manage content in today's new media landscape. To accomplish this, we conducted a research study, funded by RJI, of both legacy (print and online platforms) and online only news organizations, asking them:

- About the content they create: types and formats
- Whether their digital content is stored, backed up and accessible
- Who, in their organization, makes preservation decisions
- If they transfer digital content to memory institutions
- What risk factors threaten digital content loss
- What barriers to digital preservation they face and
- If they already have or plan to create digital preservation policies

The results of the survey will assist in informing our conversations at the "Dodging the Memory Hole" forum and provide opportunities for collaboration and pathways to action.

Technical Summary of Born Digital News Content Survey (BDNC)

A telephone survey of 476 interviews was completed by the Health and Behavioral Risk Research Center at the University of Missouri on behalf of the RJI between April 24 and May 21, 2014.

Of the 476 completed interviews, 406 were conducted with journalists of daily newspapers with websites (denoted as hybrid), and 70 with those at online only news organizations in the United States. The American Society of News Editors (ASNE) provided the sample of newspaper journalists, and the sample of online only news producers was derived from a combination of databases from the Columbia Journalism Review, ASNE and other news media associations. Attention was given to the various circulation sizes of daily newspapers ranging from less than 5,000 to 200,000 or more as well as to the number of staff members of the online only news websites.

Data collection was conducted between the hours of 9 a.m. and 5 p.m. Monday through Friday. Interviewers were monitored via unobtrusive call monitoring and the outcome of each monitoring session was reviewed with each interviewer. In addition, database monitoring was conducted to check response coding, and call outcome disposition quality. The response rate of the survey was 34.57% overall, with 35.86% rate for hybrid organizations and 30.70% rate for online only organizations.

Results and Analysis

It is important to recognize that the results of this survey are based on the perceptions of the individuals selected to answer the questions in it. News organizations to call were selected randomly from the survey sample and then asked to connect the survey representative with the "person in charge of digital operations" at the organization. Depending on the size and type of news enterprise involved, there are a wide range of possible responses to this initial and informal request, ranging from a one person operator who does it all to larger organizations with positions so specialized that their scope of knowledge may be limited to a particular silo.

Processing of the survey data is still in its early phases, so these results should be considered as preliminary findings. The RJI/JDNA survey team is anxious to delve deeper into the data, especially the open-ended responses and opportunities for cross-tabulation. Overall, it may be observed that, although most responses fall into an expected range of possibilities, many answers indicate a lack of clarity about common digital preservation terminology. Another factor that remains to be assessed for significance is the role played by organizational longevity, especially as in regard to online only news producers.

The first level of analysis is to examine the difference in responses between traditional newspaper enterprises or "Hybrids," which typically have both print and online distribution and the "Online Only" organizations. With a total of 70, the survey sample of Online Only news producers was smaller than the sample of Hybrids, at 406. Among other factors that may be at play here is the likelihood that Online Only organizations are newer than Hybrids

due to the fact that their distribution model began more recently with the invention of the World Wide Web, HTML and web browsers.

Q21SIZE: In total, what is the size of your organization's news staff, that is, how many people who work in news are now on your organization's payroll?

Survey respondents indicated that (Figure 1), compared to hybrid organizations (22%), about three times as many (60%) online-only news operations fall in the range of 1–4 people working in the news department. Hybrid organizations led in the 5–7 news employee range with 17% versus 10% for the online only group. The percentages were even for the range of 8-10 news employees, but at 14% of the 11-25 employees range, Hybrid enterprises exceeded Online Only news staff payroll numbers by about 60%.

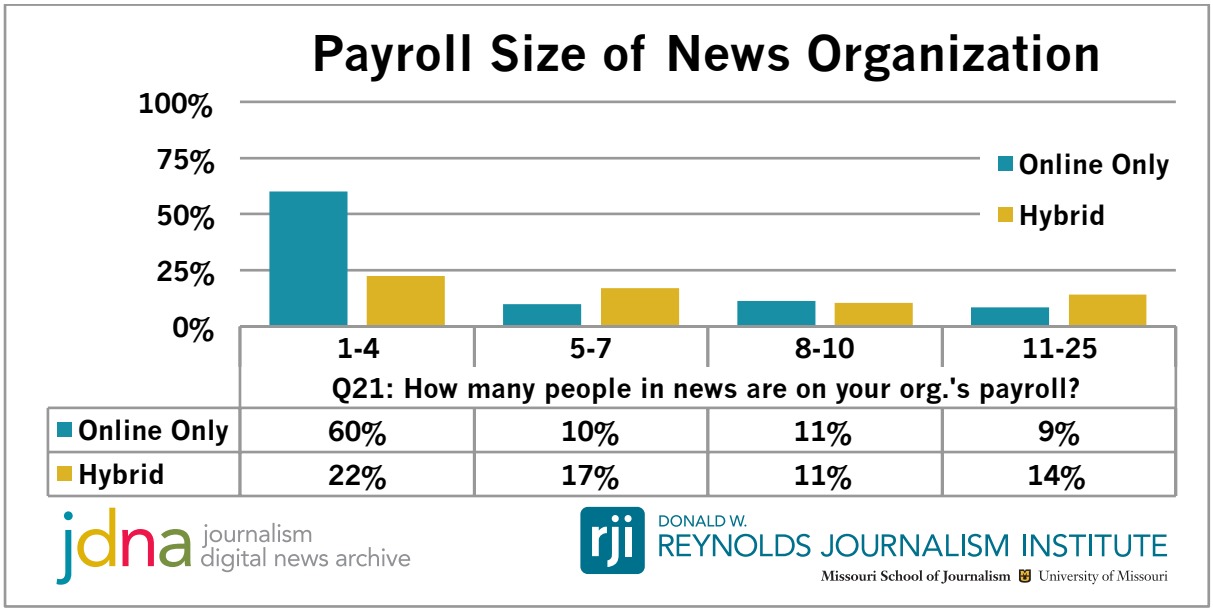


Figure 1. How many people who work in news are now on your organization's payroll?

Q1a: Does your news organization produce born-digital text content?

At the beginning of this survey, immediately before we asked this question, we attempted to provide interviewees with a good working definition of born-digital content:

> Q1. For the purpose of the survey, born-digital content refers to materials that originate in a digital form, not scanned from other media. Examples include digital photographs, digital documents, harvested web content, digital manuscripts, electronic records, and etc. Now please tell me if your news organization produces any of the following born-digital content.

It should come as no surprise that nearly all news organizations produce some form of digital text content (Figure 2). However, it is a bit of a mystery what those 5.9% of Hybrid and 1.4% of Online Only enterprises are doing that doesn't require any textual materials. This brings up the question of how well the respondents understand the question, and if those surveyed don't correctly understand what is probably one of the simplest and most widely-used digital formats, what does that mean for responses to questions that employ more sophisticated terminology and concepts?

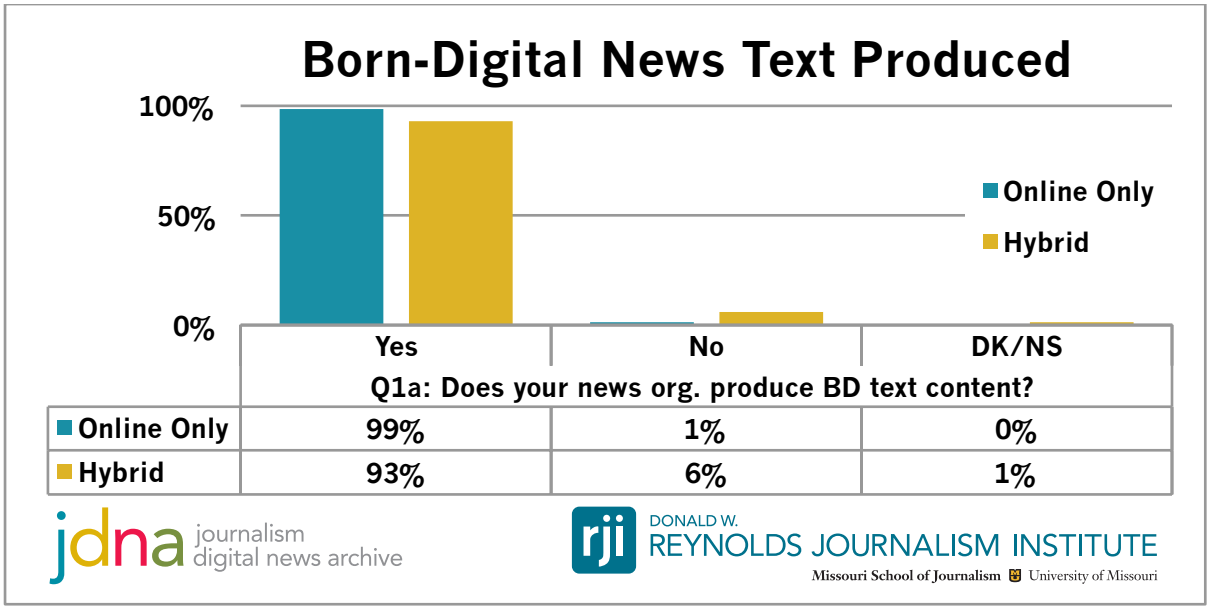


Figure 2. Does your news organization produce born-digital text content?

Q1c: Does your news organization produce born-digital video content?

Video is an increasingly popular format for digital news content. About 70% of respondents indicated (Figure 3) that their organizations are already producing this kind of content, with comparable proportions of Hybrid and Online Only enterprises engaging in video content production. It will be interesting to revisit these response rates in the years ahead, since the demand for video has been rising rapidly in recent times.

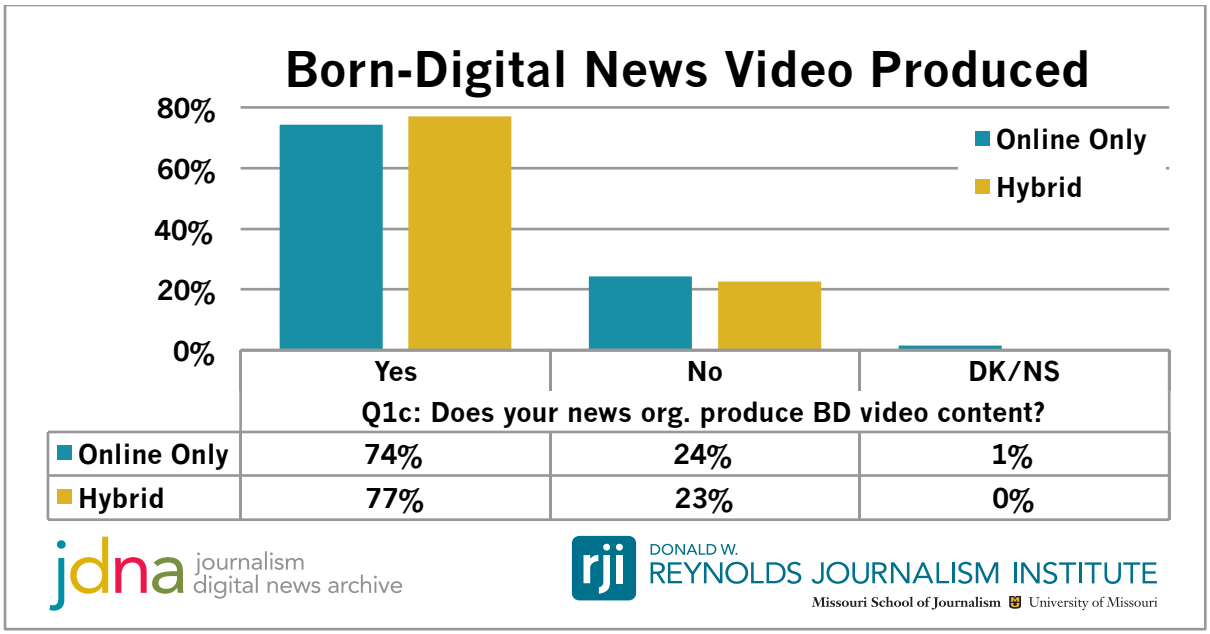


Figure 3. Does your news organization produce born-digital video content?

Q1cc: In what format is the born-digital video content produced?

Here (Figure 4), there seems to be confusion about terminology again, this time between delivery platforms (e.g. YouTube and Brightcove) and video formats (e.g. .mp4 and .avi). It is also unclear whether there is a difference between the content identified as .mp4 and mpeg

formats, since mp4 is a version of the mpeg format. Considering the popularity of .mp4 for web and mobile video delivery, it is likely that many of the responses that refer to streaming platforms may properly apply to .mp4, flash or .mov instead. The variety of video formats in use by news organizations suggests a more complex path toward preservation of these assets, with the need for significant efforts in areas of format verification, normalization, metadata and preservation policy.

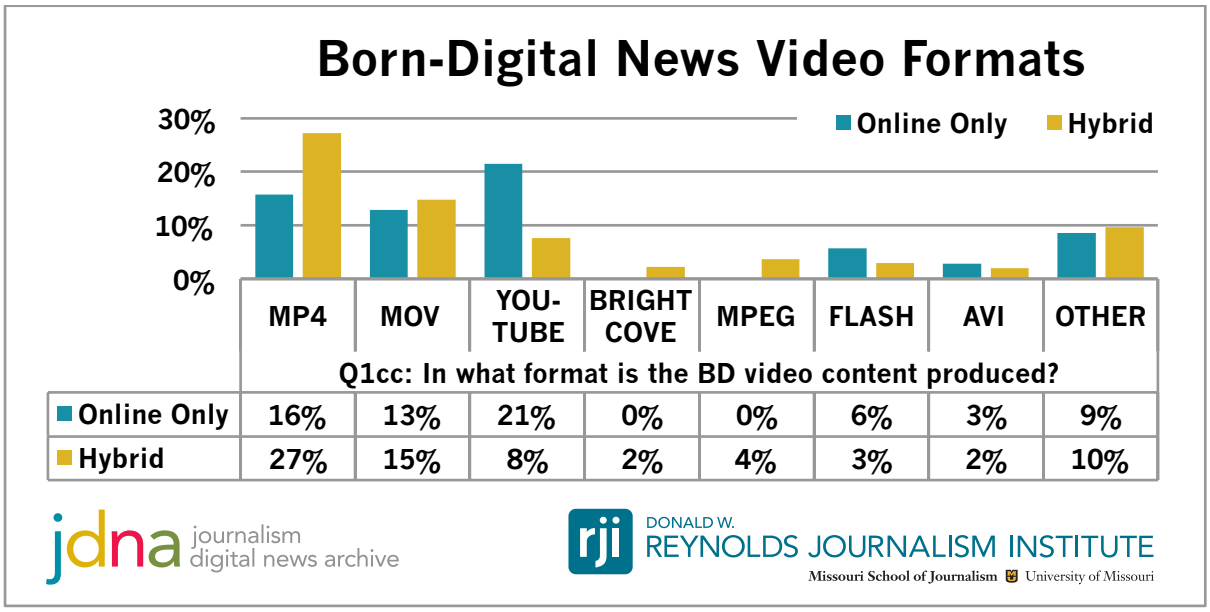


Figure 4. In what format is the born-digital video content produced?

Q5b: About what percentage of your archive of born-digital content from the past 25 years is backed up?

Almost 60% of Online Only organizations indicated (Figure 5) that they back up 100% of their born-digital news content versus only about 12% of Hybrid news producers. More study will be needed to determine the reason for this difference, but one possibility is that the Online Only enterprises are generally smaller and newer than their Hybrid counterparts and were therefore more efficient at backup operations. Likewise, the responses for less than complete backup were inversely related to those indicating 100% rates. Of those reporting that none of their content was being backed up, the Online Only group was substantially higher at 20% compared to the Hybrid category at about 6%. In the Don't Know/Not Sure category, Hybrids slightly more than doubled the rate of the Online Only group at 21% to 10%. Again, it would seem to make sense that older and bigger organizations would have a more difficult time monitoring their backup activities over a quarter of a century.

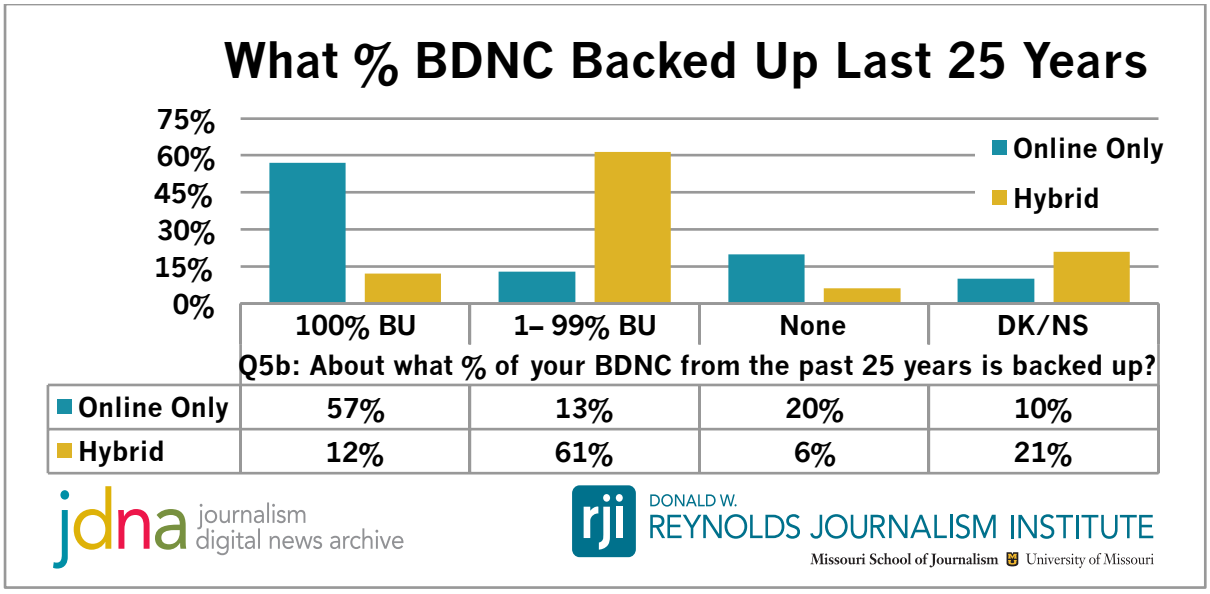


Figure 5. About what percentage of your archive of born-digital content from the past 25 years is backed up?

Q6: Does your content go to a memory institution such as a library, archive or museum?

A full 60% of Hybrid survey respondents indicate (Figure 6) that their content goes to library, archive, museum or other memory institution. In contrast, only about 11% of Online Only respondents said that their content is shared with these cultural heritage organizations. This results in some 86% of Online Only organizations not sharing their content in this way, while about 34% of Hybrid news producers say their born-digital news content does not go to memory institutions.

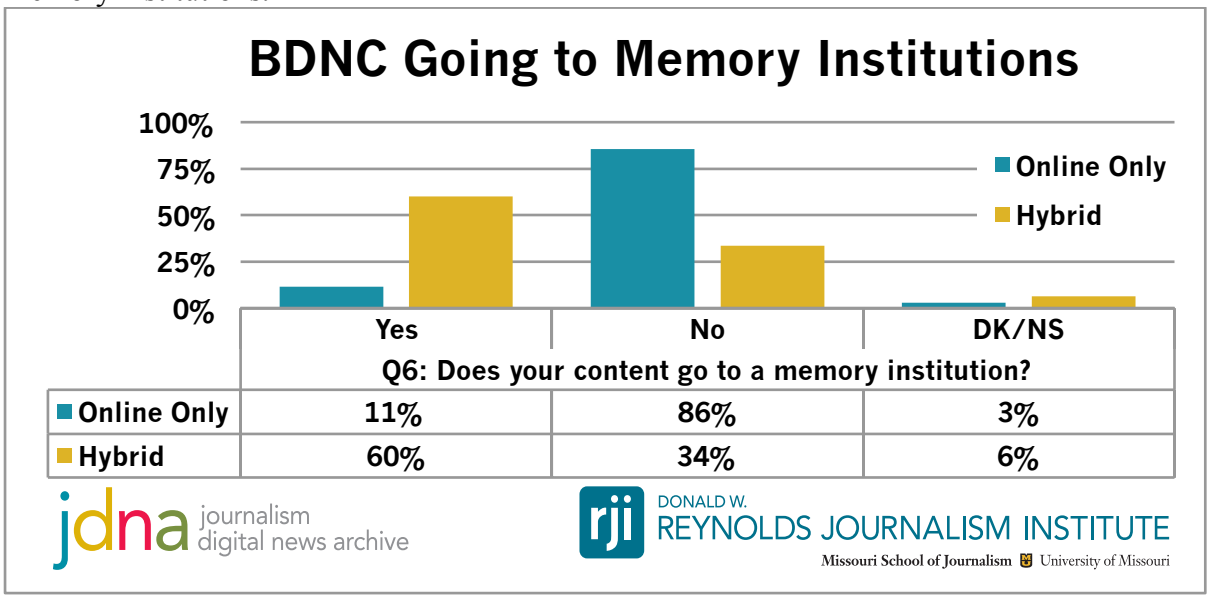


Figure 6. About what percentage of your archive of born-digital content from the past 25 years is backed up?

Q6b: What is the primary reason that you do not use a memory institution?

Of those organizations that are not sharing their content with memory institutions a significant number – 30% of Online Only and 17% of Hybrid respondents – report (Figure 7) that there is no particular reason for this or that they simply don't know of a library, archive or other memory institution with which to share their content. A relatively small number of respondents affirm the statement "We own the content and do not want to give it away" with 7% of Online Only and 2% of Hybrid interviewees in agreement. 14% of Online Only enterprises answered "Don't Know/Not Sure" with 9% of Hybrid respondents selecting this response.

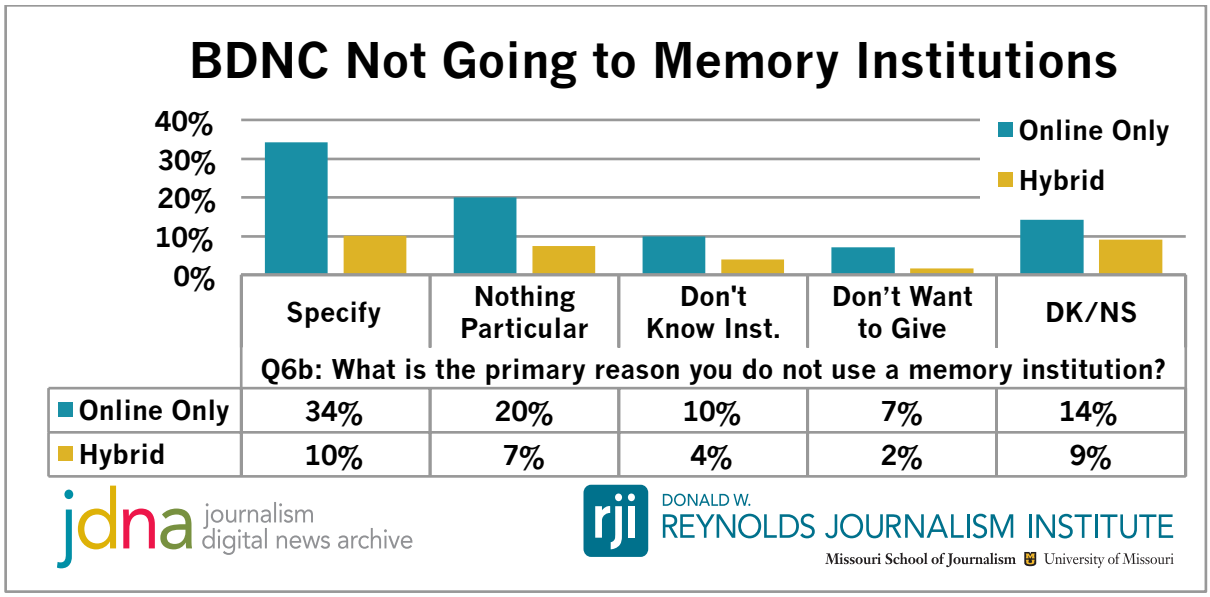


Figure 7. What is the primary reason that you do not use a memory institution?

Q8: Does your news organization currently have any written policies for managing born-digital materials?

When queried about written policies for their born-digital news content (Figure 8), the majority of both Online Only (70%) and Hybrid (64%) organizations indicated that they did not have written policies for managing those resources. Hybrid enterprises showed the highest number of responses for having written policies for BDNC at 26%, compared with Online Only news producers at 19%.

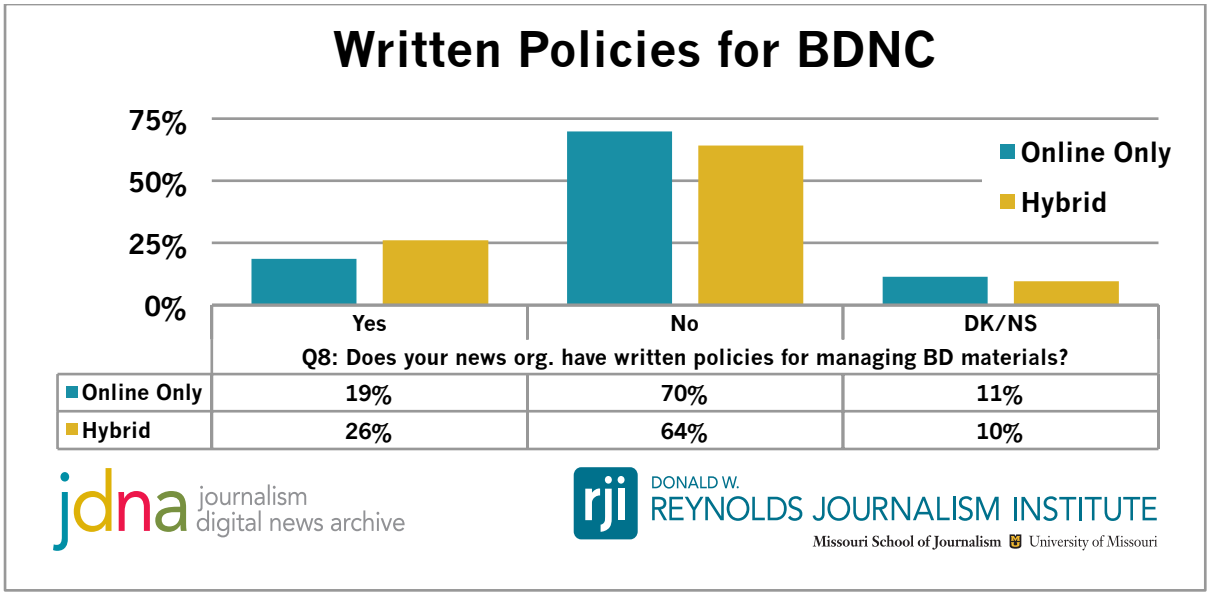


Figure 8. Does your news organization currently have any written policies for managing born-digital materials?

Q10. Do you anticipate developing new policies for born-digital content within the next 3 years?

About half of all organizations surveyed stated (Figure 9) that their organizations will develop preservation policies for their born-digital news content in the next three years, with Hybrid enterprises edging out Online Only organizations by 51% to 49%. Slightly more Online Only respondents said that they do not anticipate new preservation policies in the same period by 37% to 32%, about a 14% difference. A large number of both types of organizations responded in the "Don't Know/Not Sure" category, with Hybrids at 18% and Online Only organizations at 14%.

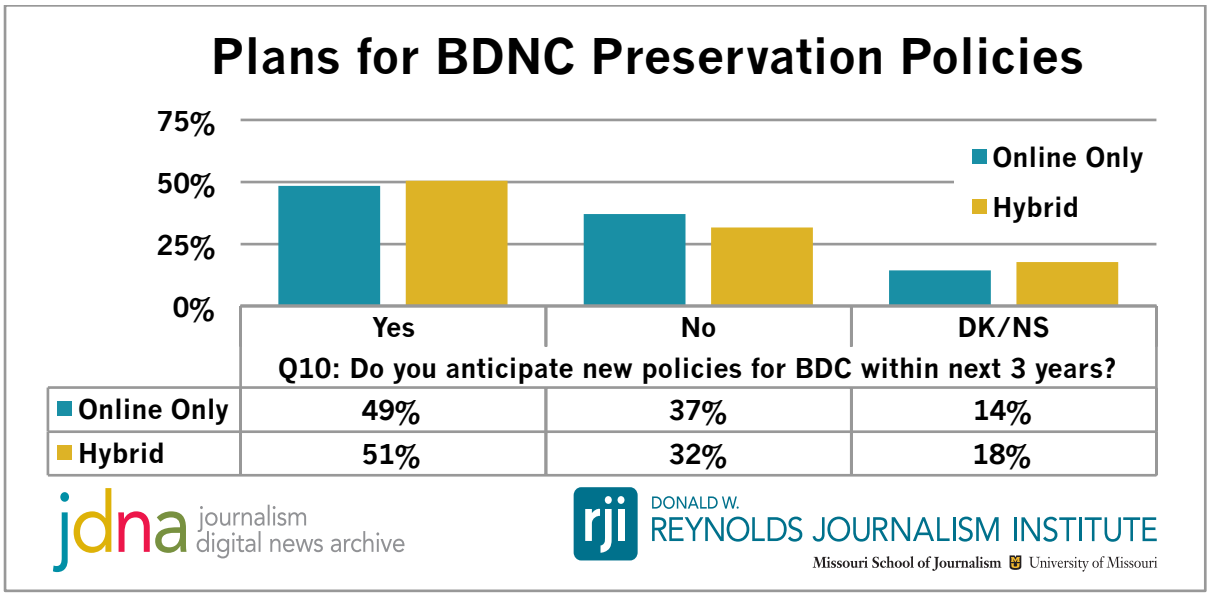


Figure 9. Does your news organization currently have any written policies for managing born-digital materials?

Conclusion

This paper has focused on born digital news backup and preservation practices at news producers in North America. The literature overview and survey show that news producers give attention to preservation only insofar as it aids new news production: News producers principal mission is to create news, not to preserve it. Preservation of news, born digital or analogue, is the primary task of cultural heritage organizations such as the Library of Congress and its sister libraries around the world; figure 7 shows that news producers very often do not share their born digital content with member organizations.

From a cultural heritage perspective and in the absence of born digital legal deposit policies and practices, news producers' answers to questions Q5b, Q6, Q6b, Q8, and Q10 are particularly worrisome. While it is by no means the only country without coherent born digital legal deposit policies and practices, the United States lags far behind policies and practices in Denmark, Sweden, France, and others. This will result in a "preservation gap", a period of time when born digital news disappears because it has not been properly preserved by an organization whose primary purpose is to preserve cultural heritage.

It is the hope of the authors that this paper and the upcoming "Dodging the Memory Hole" conference (http://www.rjionline.org/events/memoryhole) will encourage lawmakers, news producers, and cultural heritage organizations to address the United States' lack of born digital legal deposit policies and practice.

Appendix 1

The following is a condensed version of the telephone survey questions used for the survey whose results are presented in this paper.

Introduction

Hello, my name is [fill in interviewer's name], calling from the Reynolds Journalism Institute at the University of Missouri-Columbia. May I speak to the person in charge of digital operations at your news organization?

I am not trying to sell you anything. I am calling because we're conducting a confidential research study to learn about use of materials that originate in a digital form at your news organization. The results of the survey will help us better understand how news organizations like yours use and manage born-digital content in today's new media landscape.

Your name and phone number have been randomly selected from a database of newspapers and websites that specialize in journalism. Your participation is voluntary. All the information that you will provide will be kept completely anonymous, so that you cannot be identified. The survey only takes about eight minutes to complete.

If you have questions about the survey, please contact Edward McCain at (573) 882-8049 or mccaine@rjionline.org or the University of Missouri's Campus Institutional Review Board at (573) 882-9585. The Campus IRB oversees all research activities carried out at the University.

Is now a good time to talk?

Description of Field	Comments and Values
Q1. For the purpose of the survey, born-digital content refers to materials that originate in a digital form, not scanned from other media. Examples include digital photographs, digital documents, harvested web content, digital manuscripts, electronic records, and etc. Now please tell me if your news organization produces any of the following born-digital content.	[INTERVIEWER: PRESS '1' TO CONTINUE]
Q1a. TEXT	1. YES 2. NO (Go to Q1b) 8. DK/Not Sure (Go to Q1b) 9. Refused (Go to Q1b)
Q1aa. In what format is it produced? By format, I mean XML,ASCII, UTF-8, and etc...	1.Please specify:_____
Q1b. IMAGES	1. YES 2. NO (Go to Q1c) 8. DK/Not Sure (Go to Q1c) 9. Refused (Go to Q1c)
Q1bb. In what format are they produced?	1.Please specify:_____
Q1c. VIDEO	1. YES 2. NO (Go to Q1d) 8. DK/Not Sure (Go to Q1d) 9. Refused (Go to Q1d)
Q1cc. In what format is it produced?	1.Please specify:_____
Q1d. Interactives. They refer to products and services that respond to users' actions by presenting content such as text, graphics, animation, video, audio, games, etc...	1. YES 2. NO (Go to Q1e) 8. DK/Not Sure (Go to Q1e) 9. Refused (Go to Q1e)
Q1dd. In what format is it produced?	1.Please specify:_____
Q1e. Does your news organization produce mobile-only content, that is, content created only for tablets or smartphones?	1. YES 2. NO (Go to Q1f) 8. DK/Not Sure (Go to Q1f) 9. Refused (Go to Q1f)
Q1ee. In what format is it produced?	1.Please specify:_____
Q1f. Are there any types of born-digital content you produce that I have not mentioned?	1.Specify type and format 2.Nothing in particular 8.DK/Not Sure 9.Refused

Description of Field	Comments and Values
Q1ff. In what format is it produced?	1.Please specify:_____
Q2. In terms of percentage, how much of your born-digital news content is backed up?	_____ Percent of born-digital content backed up (Go to Q2b) 555.None 888.DK/Not Sure (Go to Q2b) 999.Refused (Go to Q2b)
Q2a. What is the primary reason that you do not back up born-digital content?	1.Specify 2.Nothing in particular 8.DK/Not Sure (Go to Q3) 9.Refuse (Go to Q3)
Q2aOth. Other Specify	[Open-Ended Response]
Q2b. What format(s) of news content do you back up?	01.ASCII 02.TEXT FILES WITH MARK-UP (HTML, SGML, XML, etc.) 03.EPS 04.TIFF 05.GIF (GRAPHICS INTERCHANGE FORMAT) 06.JPEG 07.MPEG 08.PDF 09.PICT 10.WMF (WINDOWS METAFILE) 11.IMAGE PAC 12.OTHER-SPECIFY
Q2bOth. Other Specify	[Open-Ended Response]
Q3. In terms of percentage, how much of previously produced news content is accessible via search by reporters and editors at your news organization?	_____ Percent 555.None 888.DK/Not Sure 999.Refused
Q4. Do you store and retrieve born-digital content using a web server?	1.Yes 2.No (Go to Q4b) 8.DK/Not Sure (Go to Q4b) 9.Refused (Go to Q4b)
Q4a. What kind of server is that?	1.Apache 2.MS IIS 3.Lighttpd 4.Sun Java 5.Jigsaw 7.Other - Specify 8.DK/Not Sure 9.Refused
Q4aOth. Other Specify	[Open-Ended Response]

Description of Field	Comments and Values
Q4b. Do you store and retrieve born-digital content using a CMS?	1.Yes, on its own 2.Yes, but through a vendor 3.No (Go to Q4c) 7.Other - Specify 8.DK/Not Sure (Go to Q4c) 9.Refused (Go to Q4c)
Q4bOth. Other Specify	[Open-Ended Response]
Q4bb. What system do you use?	1.Specify 2.Nothing in particular 8.DK/Not Sure 9.Refuse
Q4bbOth. Other Specify	[Open-Ended Response]
Q4c. Are there other ways you back up and retrieve content that I have not mentioned?	1.Specify 2.Magnetic Tape 3.Hard Disk 4.Optical Storage 5.Solid State 6.Remote Backup Service 7.No Other Ways 8.DK/Not Sure 9.Refused
Q4cOth. Other Specify	[Open-Ended Response]
Q5. How far back, in years, is your archive of born-digital content 100 percent complete?	1.Less than 1 year 2.Number of years 8.DK/Not Sure 9.Refused
Q5OTH. Record Response	[Open-Ended Response]
Q5a. Of the portion of your archive of born-digital content that is less than 100 percent complete, how far back, in years, does that archive go?	1.Less than 1 year 2.Number of years 8.DK/Not Sure 9.Refused
Q5aOTH. Record Response	[Open-Ended Response]
Q5b. About what percentage of your archive of born-digital content from the past 25 years is backed up?	_____ Percent 555.None 888.DK/Not Sure 999.Refused
Q5c. Have you ever experienced a significant loss of news content?	1.Yes 2.No 8.DK/Not Sure 9.Refused

Description of Field	Comments and Values
Q5d. What was the primary reason for the loss of news content?	1.Specify 2.Nothing in particular 8.DK/Not Sure 9.Refused
Q5dOth. Other Specify	[Open-Ended Response]
Q5e. Is the data backed up on- or off-site?	1.On Site (Specify) 2.Off Site (Specify) 3.Both 4.Neither 7.Other (Specify) 8.DK/Not Sure 9.Refused
Q5eOth. Other Specify	[Open-Ended Response]
Q5f. What is the primary reason you did not back up the data?	1.Specify 2.Nothing in particular 8.DK/Not Sure 9.Refused
Q5fOth. Other Specify	[Open-Ended Response]
Q6. Does your content go to a memory institution such as a library, archive or museum?	1.Specify 2.No (Go to Q6b) 8.DK/Not Sure (Go to Q7) 9.Refused (Go to Q7)
Q6Oth. Other Specify	[Open-Ended Response]
Q6a. In what format?	01.ASCII 02.TEXT FILES WITH MARK-UP (HTML, SGML, XML, etc.) 03.EPS 04.TIFF 05.GIF (GRAPHICS INTERCHANGE FORMAT) 06.JPEG 07.MPEG 08.PDF 09.PICT 10.WMF (WINDOWS METAFILE) 11.IMAGE PAC 12.OTHER-SPECIFY
Q6aOth. Other Specify	[Open-Ended Response]

Description of Field	Comments and Values
Q6b. What is the primary reason that you do not use a memory institution?	1.Specify 2.Nothing in particular 3.Do not know any memory institute 4.We own the content, and do not want to give it away 8.DK/Not Sure 9.Refused
Q6bOth. Other Specify	[Open-Ended Response]
Q6c. Does your organization or the memory institution you use have a digital preservation program? Preservation includes backing up data, but requires a more complex series of activities.	1.Yes 2.No 8.DK/Not Sure 9.Refused
Q7. Next I'd like to ask you a few questions about your management of born-digital content. Who makes decisions about safeguarding born-digital content at your news organization?	1.Specify 8.DK/Not Sure 9.Refuse
Q7OTH. Record Response	[Open-Ended Response]
Q8. Does your news organization currently have any written policies for managing born-digital materials?	1.Yes 2.No 8.DK/Not Sure 9.Refused
Q9. Does your news organization currently utilize consultants or contracts for preservation of born-digital materials?	1.Yes 2.No 8.DK/Not Sure 9.Refused
Q10. Do you anticipate developing new policies for born-digital content within the next 3 years?	1.Yes 2.No 8.DK/Not Sure 9.Refused
Q11. Can you briefly tell me the specific measures your organization takes to preserve born-digital content?	1.Specify 2.Nothing in particular 8.DK/Not Sure 9.Refused
Q11OTH. Record Response	[Open-Ended Response]

Description of Field	Comments and Values
Q12a. On a scale of 1 to 5 where 1 is not at all valuable and 5 is very valuable; please tell me how valuable always having access to your newspaper's coverage of past events through archives is for producing content with historic perspective?	5. Very valuable 4. 3. 2. 1. Not at all valuable 8. DON'T KNOW/NOT SURE 9. REFUSED
Q12b. Using the same scale, how valuable is always having access to your newspaper's coverage of past events for engaging audience or community?	5. Very valuable 4. 3. 2. 1. Not at all valuable 8. DON'T KNOW/NOT SURE 9. REFUSED
Q12c. How valuable is always having access to your newspaper's coverage of past events for producing quality journalism?	5. Very valuable 4. 3. 2. 1. Not at all valuable 8. DON'T KNOW/NOT SURE 9. REFUSED
Q12d. How valuable is always having access to your newspaper’s coverage of past events for producing good returns on investment?	5. Very valuable 4. 3. 2. 1. Not at all valuable 8. DON'T KNOW/NOT SURE 9. REFUSED
Q12e. Are there other values that you think always having access to your newspaper's coverage of past events has for your newsroom?	1. Specify 2. Nothing in particular 8. DON'T KNOW/NOT SURE 9. REFUSED
Q12eOTH. Record Response	[Open-Ended Response]
Q13a. On a scale of 1 to 5 where 1 is not at all a threat and 5 is a great threat, how would you rank the following factors as threats to the loss of born-digital content at your news organization within the next 3 years? Physical condition such as CPU failure, media failure or deterioration	5. A great threat 4. 3. 2. 1. Not at all a threat 8. DON'T KNOW/NOT SURE 9. REFUSED

Description of Field	Comments and Values
Q13b. Technological obsolescence	5. A great threat 4. 3. 2. 1. Not at all a threat 8. DON'T KNOW/NOT SURE 9. REFUSED
Q13c. Lack of or insufficient policy or plan for preservation	5. A great threat 4. 3. 2. 1. Not at all a threat 8. DON'T KNOW/NOT SURE 9. REFUSED
Q13d. Lack of or insufficient resources for preservation	5. A great threat 4. 3. 2. 1. Not at all a threat 8. DON'T KNOW/NOT SURE 9. REFUSED
Q13e. Are there any other factors as threats to the loss of born-digital content at your news organization within the next 3 years that I have not mentioned?	1. Yes-Specify 2. Nothing in particular 8. DON'T KNOW/NOT SURE 9. REFUSED
Q13eOTH. Record Response	[Open-Ended Response]
Q14a. Are you generating revenue through vendors such as ProQuest or Newsbank?	1. Yes 2. No 8. DON'T KNOW/NOT SURE 9. REFUSED
Q14b. What other ways are you generating revenue from archival assets?	1. Specify 2. Nothing in particular 8. DON'T KNOW/NOT SURE 9. REFUSED
Q14bOTH. Record Response	[Open-Ended Response]
Q15a. Do you place obituaries, birth notices and marriage announcements behind a paywall?	1. Yes 2. No 7. Other - specify 8. DON'T KNOW/NOT SURE 9. REFUSED

Description of Field	Comments and Values
Q15aOTH. Record Response	[Open-Ended Response]
Q15b. Do you place legal notices behind a paywall?	1. Yes 2. No 7. Other - specify 8. DON'T KNOW/NOT SURE 9. REFUSED
Q15bOTH. Record Response	[Open-Ended Response]
Q15c. Do you have a news librarian or equivalent position in your newsroom?	1. Yes 2. No 8. DON'T KNOW/NOT SURE 9. REFUSED
Q16AGE. All right, we're almost finished. I just have a few questions that will help us analyze the survey results. How old were you on your last birthday?	_____Age in years [1-120] 8. DON'T KNOW/NOT SURE 9. REFUSED
Q17JOB. What job position do you hold in your news organization?	[Open-Ended Response]
Q18YR1. How long have you worked at your current position?	_____ Number of years [1-120] 777. LESS THAN A YEAR 888. DON'T KNOW/NOT SURE 999. REFUSED
Q19YR2. How long have you worked as a paid journalist?	_____Number of years [1-120] 777. LESS THAN A YEAR 888. DON'T KNOW/NOT SURE 999. REFUSED
Q20OWN. Is your newspaper	1. Independent 2. A Group that is privately owned, or 3. A Group that is publicly traded 7. Other - specify 8. DON'T KNOW/NOT SURE 9. REFUSED
Q20OWNo. Record Response	[Open-Ended Response]
Q21SIZE. In total, what is the size of your organization's news staff, that is, how many people who work in news are now on your organization's payroll?	_____Number of news staff [1-777] 778. MORE THAN 777 NEWS STAFF 888. DON'T KNOW/NOT SURE 999. REFUSED
Q22CSIZ. What is the weekday circulation of your newspaper?	[Open-Ended Response] [ENTER AN '8' IN BOX FOR DON'T KNOW/NOT SURE] [ENTER A '9' IN BOX FOR REFUSED]

Description of Field	Comments and Values
Q23VIS. About how many unique visitors come to your site each month? [FOR EXPLANATION PURPOSE ONLY:] UNIQUE VISITORS refer to individuals who have visited a Web site or network at least once in a fixed period of time. It is a term frequently used by online authorities as a measurement of online traffic.	01. Less than 10,000 02. 10,000 but less than 50,000 03. 50,000 but less than 100,000 04. 100,000 but less than 500,000 05. 500,000 but less than 1 million 06. 1 million but less than 2 million 07. 2 million but less than 5 million 08. 5 million but less than 10 million 09. 10 million but less than 20 million 10. 20 million but less than 30 million 11. 30 million or more 88. DON'T KNOW/NOT SURE 99. REFUSED
Q24GEND. [Record respondent's gender. DO NOT ASK]	1. Male 2. Female 8. Cannot tell
Closing. That was my last question. Thank you so much for your cooperation and time. Have a good day/evening.	[INTERVIEWER: PRESS '1' TO CONTINUE]
(Phone7) Phone Number	Removed from data file
(AAPOR)	Final Disposition Code (110=Completed Interview)
(NATTMPTS)	Number of attempts made to record
(INTVID)	Interviewer ID
(IDATE)	Date of Interview
(COMPANY)	Company Name from sample file
(SEQNO) Sequence Number	A unique number assigned to each record. Sequence numbers beginning with "ONL" are Online Media, "ABC" are from groups A, B, & C, "DEF" are from groups D, E, and F, and "GH" are from groups G and H

Appendix 2

As promised above the authors will publish a subsequent paper with survey results of born digital news preservation policies and practices at cultural heritage organizations around the world. In the meantime here are the survey questions about the country's born digital news preservation policies and practices; the questions were emailed to several national libraries. Suggestions for additional questions are welcomed.

Born digital news includes for our purposes in this survey include

- Stories published on a news organization's website (these stories may or may not be published in print). These stories are primarily text based but may include photos, illustrations, or videos
- PDF files of printed newspapers

Policies

1. Do the laws of your country require publishers to legally deposit born digital news? In this case we mean that publishers MUST send born digital news to one or more legal deposit authorities.
2. Do the laws of your country require cultural heritage institutions (libraries) to harvest news organization websites that are publicly available (not behind a subscription paywall)?
3. Do the laws of your country require cultural heritage institutions (libraries) and publishers to cooperate in order to preserve born digital news when this news is behind a subscription paywall?

Practices

1. Does your library receive born digital news from publishers by FTP or similar means? For this question by "receive" we mean that publishers initiate the transmission of born digital news to the legal deposit authority (library). In tech speak, the publisher "pushes" the news to the authority (library).
2. If publishers "push" news to your library, how does your library decide which publishers? What criteria are used to decide if born digital news from a particular publisher should be preserved?
3. Does your library harvest news websites? If your library does harvest news websites, how frequently does it harvest? Once a day? Once a week? Multiple times per day or week or month?
4. Depending on the publisher, news stories published on the web may be updated several times in an hour, day, or week. Do your library's harvest practices take any action if a news story is updated (new version)?
5. Depending on the frequency of your library's web harvest, the harvest of a news website may miss new versions of a story or may miss entire stories if the publishers updates its website with a higher frequency than it is harvested. If this is the case for your library's harvest schedule, please estimate the number of stories or versions of stories that your library's new harvest misses. ("I don't know" is an acceptable answer.)
6. If your library harvests news websites, how does your library decide which websites? In other words, what criteria are used to decide if born digital news from a particular publisher should be preserved? What criteria are used to determine harvest frequency?

Most of these questions are not simple "yes" or "no". The answers may take considerable time. For this we are very grateful. We would not ask if we did not think that these are important issues. We expect that practices will show that there is a large born digital news preservation gap, with many born digital news stories failing to be preserved.

References

Alverson, J, Leetaru, K., McCargar, V., Ondracek, K., Simon, J., Reilly, B. (2011). *Preserving news in the digital environment: Mapping the newspaper industry in transition. A Report from the Center for Research Libraries April 27, 2011*. Retrieved from http://www.digitalpreservation.gov/documents/CRL_digiNews_report_110502.pdf

AP FAQs. Retrieved from http://www.ap.org/company/faqs

Blue Ribbon Task Force on Sustainable Digital Preservation and Access. (2010). *Sustainable economics for a digital planet: Ensuring long-term access to digital information*. Retrieved from http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf

Castells, P., Perdrix, F., Pulido, E., Rico, M., Benjamins, R., Contreras, J., and Lores, J. (2004). Neptuno: Semantic web technologies for a digital newspaper archive. *The Semantic Web : Research and Applications. Lecture Notes in Computer Science 3053*, pp. 445-458.

Center for Research Libraries Staff. (2013). Retrieved from http://www.crl.edu/profile/umi-dissertation-publishing-proquest-llc

Chronicling America http://www.loc.gov/ndnp/about.html

CPI Inflation Calculator. Bureau of Labor Statistics. Retrieved http://www.bls.gov/data/inflation_calculator.htm

CRL. Retrieved from http://www.crl.edu/about

Cuadra company history (2014). Retrieved from http://www.cuadra.com/about/history.html

Desmond, R. W. (1930). *Newspaper reference libraries: Their history and service*. Minneapolis : University of Minnesota.

Desmond, R. W. (1933) *Newspaper reference methods.* Minneapolis : University of Minnesota Press.

Hegg, J. (1991). Small newspaper libraries: The libraries that time (and automation) passed by. *Special Libraries* 82(4). Retrieved from http://bi.galegroup.com/essentials/article/GALE|A11514813/a29f305209589f5f10f9e6d863eff7b2?u=morenetuomcolum

IFLA Statement on Legal Deposit (2011) Retrieved from http://www.ifla.org/publications/ifla-statement-on-legal-deposit

Information Today. (1984). Cuadra directory reports online database shakeout. *Information Today*. 1(8), p. 1. Retrieved from http://connection.ebscohost.com/c/articles/18353050/cuadra-directory-reports-online-database-shakeout

International Internet Preservation Consortium (IIPC) Netpreserve.org lists countries with and without legal deposit laws Retrieved from http://netpreserve.org/legal-deposit#moreinfo

Krabbenhoeft, N., Skinner, K., Schultz, M., Zarndt, F. (2013). Chronicles in preservation: Preserving digital news and newspapers. *Preservation, Digital Technology & Culture. 42(4)*, 199-204. DOI: 10.1515/pdtc-2013-0029

Kwapil, J.F. (1921). The "Morgue" as a factor in journalism. *The Library Journal. 46,* 443-446.

Lathrop Report on Newspaper Indexes. Retrieved from

http://www.nleindex.com/index.php?pID=LRNI&sID=BrowseIndex&tID=1000E/1129A

LeFurgy, W. G. (2005). Building preservation partnerships: The Library of Congress National Digital Information Infrastructure and Preservation Program. *Library Trends*. *54 (1),* 163-172.

Luce, R. (1913). The Clipping bureau and the library. *Special Libraries*. *4(7-8),* 152-157.

Lydenberg, H. M. (1915). Preservation of modern newspaper files. *Library Journal*. *4 (4),* 240-242.

McCargar, V. (8/22/2006). *Archive Profile: NewsBank Inc.* Retrieved from http://www.crl.edu/sites/default/files/attachments/pages/newsbankprofile.pdf

McCargar, V. (2011) *Repository Profile –The Associated Press*. Retrieved from http://www.crl.edu/sites/default/files/attachments/pages/AP%20Profile%20final%20doc_3_2011.pdf

McCargar, V. (2011). A mandate to preserve: Assessing the inaugural newspaper archive summit. Retrieved www.rjionline.org/news/**mandate-preserve**

Miller, Tim. (1983,September). "Information, Please, and *Fast*: Reporting's Revolution—Databases," *Washington Journalism Review*, *5*(7), pp. 51-53.

Nathan, G. J. (1910). Journalistic "Morgues." *The Bookman. 31 (6),* 597-599.

NEH http://www.neh.gov/us-newspaper-program

Reilly, B. (2013). CRL's Long-lived digital collections project: Working to provide member libraries peace-of-mind. *Against the Grain*. *21(2*), 34-36.

Reilly, B. (2011). CRL preservation analysis of electronic news: Mapping the newspaper industry in transition. *The Charleston Advisor*. *13(1)*, 54-55. http://dx.doi.org/10.5260/chara.13.1.54

Soffin, S. et al. (1987). Online databases and newspapers: An assessment of utilization and attitudes. Paper presented at the Annual Meeting of the Association for Education in Journalism and Mass Communication (70th, San Antonio, TX, August 1-4,1987). Retrieved http://files.eric.ed.gov/fulltext/ED286178.pdf

Ullmann, J. (1983). Large newspaper use of commercial data bases. *IDEAS: Research You Can Use from the Missouri School of Journalism.* 1(3), pp. 11-20.

Vittolino, S. (1985). Bu/text: Newspapers pointing way to profit. Sun Sentinel. Retrieved from http://articles.sun-sentinel.com/1985-03-18/business/8501100563_1_knight-ridder-videotext-philadelphia-newspapers

Ward, J., Hansen, K. (1986). Commentary: Information age methods in a new reporting model. *Newspaper Research Journal*. 7(3).

Warhover, T. (April, 15, 2011). Dear reader: Digital archives don't last: A tale of corruption and crashes. Columbiamissourian.com. Retrieved from http://bit.ly/dNZ7pg

World Association of Newspapers and News Publishers (WAN-IFRA). Retrieved from http://www.wan-ifra.org.

Zarndt, F., Boddie, S., Lanz, D. (2011). Copyright, access policy, and copyright enforcement for digital newspaper collections. In Walravens, H. (Ed.), *Newspapers: Legal deposit and research in the digital era*. *IFLA Publications (Book 150)*. Berlin, Germany : De Gruyter. (pp. 91-101).