

A report on a research project led by faculty and experts at Donald W. Reynolds Journalism Institute and The University of Missouri Libraries.
This project was supported by a grant from The Andrew W. Mellon Foundation.

ENDANGERED BUT NOT TOO LATE THE STATE OF DIGITAL NEWS PRESERVATION



University of Missouri

University Libraries

Donald W. Reynolds Journalism Institute

Right now, a clock is ticking on the longevity of your news content. ... For born-digital content, it's a clock that could strike midnight at any moment when a disk drive or database fails, a power supply dies or a server is corrupted or compromised, wiping out content in the blink of an eye.

Published April 19, 2021

© 2021 by The Curators of the University of Missouri.

Licensed to the general public under the Creative Commons Attribution 4.0 International License (CC BY 4.0); licensing guidelines may be found at <https://creativecommons.org/licenses/by/4.0/>.

Suggested citation: McCain, Edward, Neil Mara, Kara Van Malssen, Dorothy Carner, Bernard Reilly, Kerri Willette, Sandy Schiefer, Joe Askins and Sarah Buchanan. *Endangered But Not Too Late: The State of Digital News Preservation*. Columbia, MO: University of Missouri, 2021. <https://hdl.handle.net/10355/80931>.

This report may also be downloaded from the Donald W. Reynolds Journalism Institute website: <https://www.rjionline.org/preservenews>.

The researchers gratefully acknowledge the support of The Andrew W. Mellon Foundation, which provided a grant for this study.

Table of Contents

Introduction	1
Executive Summary	5
1 Contents: A user's guide to this report	9
2 Background: Digital news is different	11
3 Methodology	25
4 Findings: The state of digital news preservation today	33
Content findings	34
Technology findings	50
Practices findings	78
Mission findings	87
5 Recommendations	93
Immediate actions for any newsroom	94
Medium-term actions for any newsroom	101
Long-term actions for the industry	106
6 Digital News as Historic Record	113
Glossary	117
Bibliography	123
Appendix A: Initial interview questions	127
Appendix B: Revised interview questions	135
Appendix C: Interpreting the SPOT Model for news organizations	139
Appendix D: Keywords list	143

List of Figures

1	How completely does your organization preserve news content?	34
2	Extent of saved content	34
3	What types of news content are preserved?	37
4	What are your primary and secondary channels for news distribution?	38
5	SPOT model analysis of saved content for 24 news organizations	40
6	How news organizations rated on preservation quality using a 5-point scale	41
7	How news organizations compared in resources for preservation	41
8	How news organizations rated on specific SPOT properties	43
9	Preserved content ratings by geographic scope	44
10	Preserved content ratings by primary channel	44
11	Preserved content ratings by origin	45
12	For what purposes is news content preserved?	45
13	Who are the primary users of saved news content?	45
14	Outside of publishing channels, how do external parties access your content?	47
15	Are there limits on who has access to preserved content inside your organization?	47
16	What are the key challenges the organization faces in archiving and preserving news content?	51
17	What are key drivers of technology or system changes?	53
18	Are there any planned technology changes that will impact preservation?	54
19	Does the organization have a disaster or business continuity plan?	54
20	How many of the news organizations interviewed are using each brand or type of Content Management System or other publishing platform?	58
21	Number of major functions served by newsroom systems	63
22	What are the major newsroom systems in use and what functions/roles, in addition to archiving, do they serve?	65
23	Is content uniquely identified? How is this done?	69
24	Does the metadata used in your organization utilize archival specification standards?	70
25	Do your publishing systems use/preserve metadata that connect all parts of a story?	70
26	Do you tailor metadata to search engines, platforms or browsing environments?	70
27	Partial list of standardized sections, tags and pages in use at Lee Enterprises	74
28	Number of staff doing preservation work (sum of all sites)	84
29	Number of total staff defined, assigned in preservation work	85
30	Are your preservation selection guidelines implicit or explicit?	89
31	Does your organization preserve news content for the public record?	91

Research Team

- **Edward McCain**, Principal Investigator, Digital Curator of Journalism, University of Missouri Libraries, Donald W. Reynolds Journalism Institute
- **Neil Mara**, Reynolds Journalism Institute Fellow, Neil Mara News-Tech Consulting; former McClatchy News Systems Director and Journalist
- **Kara Van Malssen**, Partner and Senior Consultant, AVP Consulting
- **Dorothy Carner**, Head, Journalism Libraries/Adjunct Journalism Professor, University of Missouri Libraries/Missouri School of Journalism
- **Bernard Reilly**, President Emeritus and Senior Advisor, Center for Research Libraries
- **Kerri Willette**, Senior Consultant, AVP Consulting
- **Sandy Schiefer**, Journalism Research and Digital Asset Librarian, University of Missouri Libraries
- **Joe Askins**, Head, Instructional Services, Library Research and Information Services, University of Missouri Libraries
- **Sarah Buchanan**, Assistant Professor, School of Information Science and Learning Technologies

Photo credits:

- **Edward McCain**, pages 12, 13, 15, 17, 23, 31, 32, 44, 51 and 53.
- **Neil Mara**, for photos on pages 19, 26, 29 and 72.

Report design:

- **Godat Design** godatdesign.com

For more information, please contact:

Donald W. Reynolds Journalism Institute

401 S 9th Street, Administrative Offices, Suite 300, Columbia, MO 65211

University of Missouri Libraries Administrative Offices

104 Ellis Library, University of Missouri-Columbia, Columbia, MO 65201-5149

Email:

preservenews@rjionline.org

Acknowledgments

The researchers would like to express their gratitude to the following individuals for their invaluable assistance in making this report possible:

Chris Alexander, Thomas Ammermann, Clifford Anderson, Matthew Ballinger, Liz Bauerle, Kerry Bean, Carsten Boe Jensen, Glenn Burkins, Karen Cariani, Brent Carter, Rebecca Chandler, Sherry Chisenhall, Steve Daly, Shelly DeLuca, Ken DuFort, Jim Duran, Deborah Dwyer, Sam Ediger, Emily Egan, Brian Ernst, Annette Feldman, Ross Fitzpatrick, Angela Ford, Rebecca Fraimow, Lee Funnell, Dan Gaines, Ben Gerst, Kurt Gessler, Jeremy Gilbert, Catharine Giordano, Ken Godat, Mark Graham, Abigail Grotke, Gary Hairlson, Brian Heffernan, Rich Hein, Bob Hesskamp, Mark Johnson, Robin Johnston, Torben Juul, Damon Kiesow, Charlotte Kingo Marvig, Brian Kratzer, Walter Kreiling, Heather Lamb, Nathan Lawrence, Roger Macdonald, Chad Mahoney, Jerimiah (Jerry) Manion, Christine Masters, Sarah McLaughlin, Shane Miner, Katherine Monberg, Shula Neuman, Beth O'Malley, Jon Okerstrom, Tim Page, Chuck Palsho, Randy Picht, Ryan Pollyea, Peter Rippon, Julie Rogers, Bob Rose, Alexis Rossi, Kori Rumore, Fred Schecker, Jennifer Selph, Ernest Shaw, James Simon, Erin Sood, Megan Sowder-Staley, Phil Spencer, Amanda St. Amand, Kellie Stanfield, Elizabeth Stephens, Ann Marie Stephenson, David Tenenbaum, Deborah Thomas, Sarp Uzkan, Pamela Vizner, Julia Vytopil, Ron Wallace, Brad Ward, Deborah Ward, Ben Welsh, Dave Whittaker, Jared Wiener, Amanda Wilkins, Brin Winterbottom, Savannah Wood, Veronika Zielinska, and Matthew Zilske, and many others whose names are not listed.

In addition, we are grateful to many previous researchers on these issues, whose work formed an invaluable starting point for this project.

It's no headline that newsrooms across the country today are struggling to survive, battered by multiple economic forces at work for years, the manic march of digital competition and technology, the storm of political attacks on their mission and in 2020 the sudden repercussions of an invisible pandemic predator.

While these are well known across the news industry, one little-recognized, unlisted casualty of the struggle is the impact on an irreplaceable resource that citizens across America rely on: the public record of their communities as recorded by their local newspaper, radio or TV station, online newsroom or other news outlet.

What if that record is going away? What if significant parts of that information stream, digital content especially, are getting lost, erased, chewed up by the machinery of technology, untended in the financial struggle and increasingly allowed to digitally decay, to get disconnected from its various components, to disintegrate? What if, because of the mind-boggling complexity of modern digital publishing systems, our first draft of history is dissolving?

That's the unfortunate fact of what's happening right now in newsrooms across the country. Quietly, in the background of the news industry's public struggles is a nearly invisible but dramatic decline in efforts to preserve our daily news.

In the rush to get the news out, with shrinking resources in the face of expanding competition, today's newsrooms are finding it difficult to devote money or staff time to what seems like an insurmountably daunting effort to save its growing array of digital news content.

... parts of today's news content are simply disappearing, lost in the race to adapt to ever-shifting demands of digital publishing, or caught in the grinding gears of hurried migrations from one Content Management System (CMS) to another ...

The result is that parts of today's news content are simply disappearing, lost in the race to adapt to ever-shifting demands of digital publishing, or caught in the grinding gears of hurried migrations from one Content Management System (CMS) to another, often without even realizing what's lost until long afterward when it's too late to recover.

Where once there were experts who knew exactly how to tap each newsroom's previous content, for example, now most newsrooms have long since let go of their news librarians and archivists. Where once newsrooms had systems and workflows designed around preserving the news, many have dropped archive systems from their tech closets, put preservation on cruise-control and now rely largely on CMS platforms that were designed for publishing to the web as rapidly and efficiently as possible.

And as local newspapers fold in the face of this tidal wave, well-documented in a series of UNC-Chapel Hill studies that track the decline, the outcome too often is a yawning gap in the public record of many communities.¹ These are some of the realities encountered during a year-and-a-half-long research effort into news preservation in the digital era.

¹ Penelope Muse Abernathy, "News Deserts and Ghost Newspapers: Will Local News Survive?" (Center for Innovation and Sustainability in Local Media, Hussman School of Journalism and Media, University of North Carolina at Chapel Hill, 2020), https://www.usnewsdeserts.com/wp-content/uploads/2020/06/2020_News_Deserts_and_Ghost_Newspapers.pdf.

What prompted this research was a growing concern that news organizations across the U.S., mostly due to great financial stress, are not doing what's needed to ensure that news content is preserved for the long term. Lack of attention to preservation appears to be a problem especially with digital content, following the industry-wide shift to the web and digital news channels as the predominant means for news publishing.

The news of the day is increasingly not being saved anywhere except on websites, in content management systems that change so rapidly they can make a previously published news story obsolete in a matter of weeks, even days.

What's clear from our research is that the typical expectation of readers and the public, that news preservation is automatic in the digital age, simply isn't correct. Chances are, in fact, that unless you do something specific and intentional to preserve it, some or all of your born-digital content will be gone in a few years. It will no longer be accessible, readable, searchable or recoverable unless you take deliberate steps to ensure it is.

Right now, a clock is ticking on the longevity of your news content. It's not a 50- or 100-year clock as it would be if your content was on newsprint, subject primarily to slow chemical deterioration. Or a multi-century clock that's allowed some books and manuscripts printed on higher quality materials to survive from the middle ages and earlier times. For born-digital content, it's a clock that could strike midnight at any moment when a disk drive or database fails, a power supply dies or a server is corrupted or compromised, wiping out content in the blink of an eye.

When you look even three or four years down the road, it's virtually certain that, regardless of the computer platform you use, it will be at least significantly, if not radically, different from what it is now.

Consider the simple fact that no current computer has existed intact and without major change for more than 2-3 years at the very longest. Two to three weeks or months is the more common scenario in the era of constant upgrades and rapid technical obsolescence. When you look even three or four years down the road, it's virtually certain that, regardless of the computer platform you use, it will be at least significantly, if not radically, different from what it is now.

Will the content you create today survive those transitions? Will it translate to the modified technology? Will it be readable? Searchable? Will images and videos remain connected with text? Will animated graphics work?

Unfortunately, the answer is usually no. And this issue is turning into a troubling, society-wide problem for those who rely on this information. Not just news reporters who need background for the next breaking news story, but researchers, authors, students in high schools and universities, governments and public libraries that increasingly scramble to serve citizens, and many more.

After meeting and talking with news media companies, tech vendors and memory institutions in North America and Europe over the past 18 months, our group of researchers at the University of Missouri Libraries and the Donald W. Reynolds Journalism Institute (RJI) has found some promising paths forward toward solving this problem.

In a research effort supported by a generous grant from The Andrew W. Mellon Foundation, we visited onsite and in detailed video sessions with nearly 40 newsrooms and related organizations large and small, ranging from small radio stations to some of the largest news organizations anywhere, such as CNN, the BBC and GBH (formerly WGBH); from digital startups to a newsroom that's been covering the Black community in Baltimore for 128 years. We also talked to key memory institutions including the Library of Congress, the Center for Research Libraries and the Netherlands Institute for Sound and Vision.

Will the content you create today survive those transitions?

Will it translate to the modified technology? Will it be readable?

Searchable? Will images and videos remain connected with text?

Will animated graphics work?

In more than 100 separate interviews and face-to-face meetings, we learned in significant detail about the issues and challenges behind the decline of preservation efforts. We learned about the unexpected ways in which some of the common web publishing technologies and workflows seem to be contributing, unintentionally, to the problem. We also learned about the impact of centralization and standardization efforts, the worrisome problems that often occur during conversions from one system to another, and the effect that the loss of archival expertise has had on news preservation.

More hopefully, we also learned about the many terrific efforts, systems and practices now in place at some of the news organizations that are most successful at digital news content preservation. These organizations offer some of the best ideas around on what to do, what tools and systems, workflows and policies work. Some of them are pushing the envelope with artificial intelligence and machine learning tools that expand the value of existing content.

Finally, in talking with a number of key system vendors who provide the technologies used in today's newsrooms, we found promising technology developments and plans that could help light the way for news publishers seeking a better way forward.

This report presents the findings of our research, the issues, the problems and underlying forces that are contributing to losses in the flood of news content racing past us every day. Here we attempt to provide some guidance into the tools, systems, workflows and policies that we hope will help solve these issues and ensure preservation of these irreplaceable community treasures.

Executive Summary

A

Are you concerned about the longevity of your news organization's content?

Have you lost any content or critical metadata through the constant churn of shifting digital technologies? Can you pull up the original, full-resolution videos and photographs your newsrooms produced for that major breaking news story last year? Can you prove definitively that you own the copyright to the story that went with it? And are you wondering whether you can locate and access the evergreen content you need for that proposed new digital product you're considering on food or travel or sports?

If any of these questions worry you, or you wonder about the future of the public record of our communities in the age of massive expansion of digital news channels and sources, there are problems that need to be understood and solved, and steps newsrooms can take to ensure availability, access and control of digital news content and assets.

Can you pull up the original, full-resolution videos and photographs your newsrooms produced for that major breaking news story last year?

That's the purpose of this report, to provide the results of research into what's happening in today's news media when it comes to preserving irreplaceable digital news content. And to share the best ideas and practices news organizations can adopt to address the common problems that can so easily threaten the digital news content we are creating every day.

In an effort to address these questions, a research group from the University of Missouri Libraries and the Donald W. Reynolds Journalism Institute launched an 18-month-long project to assess the status of preservation of born-digital news content across the news industry. Supported by a generous grant from The Andrew W. Mellon Foundation, this report provides the results of that research, conducted through onsite and video conference interviews, including a wealth of information and analysis on a little-known, largely hidden problem that's been developing in the shadow of the news industry's financial crisis and the shift to digital production and publishing.

This report includes a User's Guide to finding and understanding what's in each section, followed by a concise Background on how the switch to digital publishing, and the collapse of old business models helped fuel the upheavals that developed into today's preservation problems. A summary of the Methodology used in this research comes next, followed by the report's Findings, Recommendations, Conclusion and Appendices.

Findings summary

What we found in this research is that news organizations are saving digital news content to at least a limited extent, one that often depends on the kind of technologies where news content resides, their purpose, and other key factors. We found that the degree to which your existing content is accessible and useful depends not only on the technologies used, but also on your policies, if any, about what is saved. Other factors that affect access to content include the workflows used to assemble and store content, the metadata that's saved with your content—or missing depending on how it is managed—whether or not you have staff dedicated to preservation work, and how well content translates when you undergo a transition from one technology platform to another, an inevitable fact of life in today's publishing industry.

News organizations that use either an archive or digital asset management (DAM) system of some kind have the most control of the content used to post, publish, broadcast or stream the news. They are also in the best position to find and access past content, understand its origins and licensing rights, reuse it for new products, tap it for newsroom research, publish it in related content links, and take full advantage of the long-tail phenomenon by reselling to the public or research community.

... the need to act is urgent. Digital news content is fragile and endangered, and unexpected losses happen every day.

Those organizations which do not have a separate archive or asset system, but rely instead on their web CMS or primary publishing/broadcasting technology do not have the same degree of access or control, and are usually limited to versions of content assets that are adjusted to meet the requirements of a web CMS or equivalent system. This includes available metadata, which is likely to be tailored to the needs of the channels served by such systems, rather than full metadata detailing origin, ownership, geographic and descriptive information, as well as linkages to related content objects and past usages and changes.

Recommendations summary

Based on our Findings, we share a number of best practices and recommendations for news organizations to address issues in content preservation. These Recommendations include **Immediate actions** that can be done now at little or no cost; **Medium-term actions** that involve deeper changes or investments; and **Industry-wide actions** that call for longer-term collaborations for systemic change. We include specific resources to help those who want to act.

Our recommendations for **Immediate actions** include creation of a written policy for news preservation in your organization, and how to handle requests for unpublishing. We also encourage every news organization to tap someone to take responsibility for preservation work, even part-time if needed. We recommend reviewing the metadata now available in the systems your news organization relies on for saving content. It's especially important to check metadata for content origin and ownership for future reuse and licensing rights. In addition, we encourage news organizations that need help in preservation to reach out to groups such as the Internet Archive, libraries and other such memory institutions.

For **Medium-term actions**, we recommend that any news organizations which do not already have an archive or asset management platform that can preserve digital news independent of publishing systems look seriously into acquiring one. In the long run this is likely to be the best way to ensure control of your news content assets regardless of what forms are needed for individual digital publishing, broadcasting, posting or streaming channels. In the meantime, be sure to check with your tech vendors to find out what new functionality they have developed that might help support long-term access to your valuable content.

For **Industry-wide actions**, we encourage the industry to tackle this problem collectively, developing common guidelines on what needs to be preserved in the digital news era, how this should be done as best practices, share information on the benefits of utilizing past content, create partnerships to enable collaboration on finding better preservation methods, and work with academia to expand training for the kinds of expertise needed to manage news content in the digital era.

Whichever of these make sense for your news organization, it's important to understand that the need to act is urgent. Digital news content is fragile and endangered, and unexpected losses happen every day. But by starting down the path today to address the issues specific to your situation, you may at least have confidence that your organization is moving in the right direction.

Our goal for this report is to encourage decision-makers at all levels of the news industry to champion the cause of preserving news content in digital formats. News is a vital part of a democratic society, both recording and shaping it. We believe it is too valuable to lose from neglect or lack of knowledge. We hope you will share this report with others, in newsrooms and boardrooms, to increase the awareness of the urgent need to address born-digital news preservation. Complacency is the enemy. The time to act is now.



Contents

A user's guide to this report

In this section, we hope to provide readers with information that orients them and supplies some basic context for understanding and navigating this report. Here's a brief rundown of what is contained in this report and where you can find it:

The **Introduction** and **Executive Summary** sections above have already addressed the question of why it is important to preserve born-digital news content and provided high-level summations of what we consider the most important discoveries from the 18 months our team spent asking questions and analyzing what news organizations told us about their operations.

The sections following this one include:

The **Background** section, which provides a brief overview of the history of news preservation, including that of digital news content, especially as it relates to the rationale for this research project. This includes the evolution of digital preservation research and theory that advanced a wider recognition that preserving digital news content is a vital pursuit. This section also highlights some pioneering practical efforts to capture and preserve news in digital formats.

The **Methodology** section explains the details of how our research team performed our investigation. This portion of the report includes the following:

- The kinds of news and other organizations involved and how they were approached and selected for participation in this study.
- How the research team's approach evolved over the course of the project, beginning with an exploration of workflow and ending with a more structured investigation.
- Areas where the research project ran into challenges.

The **Findings** section is a detailed dive into what the report team considers to be the most important results of our research. If you want to get the scoop on what we learned, this is the place to begin reading.

The **Recommendations** section provides the research team's best ideas for solving the problems and leveraging opportunities that we learned about during our newsroom visits and interviews.

The **Appendices** contain project documents as well as supplemental information. They include material generated as part of the study, such as a list of organizations involved and sets of interview questions. Additional materials such as suggested readings in our **Bibliography** and a **Glossary** provide context and help explain terminology specific to the news industry or to digital preservation.



Background

Digital news is different

What's different about digital news preservation?

There's a fundamental difference between preserving ink on paper and preserving digital content. The essential nature of analog and digital objects is different. To understand the research behind this report it's important to understand the differences between them. Simply put, compared with ink on paper, it is harder to maintain access to digital news content over time—much harder. This difficulty is due to the fact that the physical and chemical processes at work to deteriorate newsprint take place at a relatively slow rate. In such situations, print archives can often be characterized as preservation by benign neglect. Indeed, the text and images stored in a book can sit on a cool, dry and dark shelf for decades, sometimes centuries, in a ready state for someone to open its pages and consume its contents. Barring flood or fire, paper decays relatively slowly, fading and crumbling away in a predictable and gradual farewell. Digital content is different. It can be gone in an instant without warning.

The factors behind digital fragility are many. In contrast to analog content, for a person to read, listen, see or hear digital information, a host of intermediary technologies are involved, requiring very particular hardware and software. Because of these complex technical requirements, access to digital content is subject to a wide variety of ongoing threats, both physical and technological. The ongoing evolution of computing processors, formats, operating systems, applications and other avenues required to allow a person to experience digital information in a meaningful way means that preserving access to digital information requires carefully considered and nearly constant attention in order to guarantee access over time. The renowned Bodleian Libraries provide the following definition of digital preservation: "The formal activity of ensuring access to digital information for as long as necessary. It requires policies, planning, resource allocation (funds, time, people) and appropriate technologies and actions to ensure accessibility, accurate rendering and authenticity of digital objects."²

Digital preservation is active—like life support

Experience and research show that digital content that is not organized and managed does not persist. Benign neglect doesn't work for digital content. True digital preservation involves much more than simply backing up files, folders and systems. To preserve content over time, specific actions must be taken. Content must be migrated to new disks to avoid hard drive failures. File formats that are unstable or in danger of obsolescence must be carefully converted to preservation-friendly formats. Safeguards must be in place to prevent accidental or deliberate change. Content loss or alteration

² Edith Halvarsson, "Introduction to Digital Preservation: What Is Digital Preservation?," Oxford LibGuides, February 12, 2021, <https://ox.libguides.com/digitalpreservation/whatisdp>.

must be detected and restored from multiple backups. Metadata about important characteristics of the content such as authorship or date are needed to enable its authentication. If news organizations don't address the need for informed and coordinated efforts to ensure present and future access, news content in digital formats will inevitably disappear forever.



The archive for KOMU consists of multiple hard drives with digital video footage organized by year. The station has plans to implement a DAM or MAM in the future.

By identifying and mitigating circumstances that put digital content at risk, it may be possible to avoid the loss of that information. Threats to content, including digital news, generally fall into two categories: threats to the news enterprise itself and threats to news content. This report focuses on the latter, but the financial viability of news organizations poses an equal, if not greater menace to the long-term survival of news content.

For whom are we preserving?

In this report, we recognize there are several user communities interested in the fate of digital news content.^{3,4} The first and foremost of such stakeholders are news organizations themselves. They own and hold the content they produce, often building new or updated stories supported by previously published reporting. If news organizations somehow lose their archive, the opportunity for other communities to access saved news content now or in the future is also lost. That includes scholarly researchers and the general public.

At this moment, it is difficult to chart a clear course for how news content will make its way into some form of long-term public access for scholars and others, but it is time to start toward that destination. Digital preservation has advanced to the point where the questions are not so much about technical issues as they are about determining priorities. This much we know: for news content to remain accessible for the long term, the current goal must be—at minimum—to keep content from disappearing now. The logic behind this is based on the premise that digital preservation is not necessarily a once-and-for-all or all-or-nothing proposition.⁵ Saving born-digital news today will preserve the opportunity for long-term access later, not only for news organizations, but for broader use by society as a whole.

3 Digital preservationists call such groups a “Designated Community,” defined as “An identified group of potential Consumers who should be able to understand a particular set of information. The Designated Community may be composed of multiple user communities.”

4 “Reference Model for an Open Archival Information System (OAIS),” Recommendation for Space Data System Practices (The Consultative Committee for Space Data Systems, June 2012), <https://public.ccsds.org/Pubs/650x0m2.pdf>.

5 Blue Ribbon Task Force on Sustainable Digital Preservation and Access, “Sustainable Economics for a Digital Planet: Ensuring Long-Term Access to Digital Information,” February 2010, https://www.cs.rpi.edu/~bermaf/BRTF_Final_Report.pdf.

How is digital news content stored and accessed?

It's important to consider that decisions and actions about preservation will be made over time under changing circumstances. In an effort toward longevity, born-digital news content might be held in one or more of the following software systems over its life:

- **Content Management System (CMS):** used to create, modify and distribute digital content, usually via the web. This web content may include text, graphics, photos, audio, video, maps and code for interactive features.
- **Digital Asset Management System (DAM):** used to store, organize, manage, access and distribute file-based digital objects. A DAM may integrate with a CMS or other production tools, but keeps an original, unchanged version of the file to guard against change or loss.
- **Media Asset Management System (MAM):** similar to a DAM, but with features that integrate with video production workflows.
- **Preservation Repository:** a complex set of components designed to ensure the access, viability, security, usability and discoverability of its content for the long term. This kind of repository can be certified as a "trustworthy digital repository" as specified in ISO 16363:2012.⁶



In addition to storing content in its CMS and DAM, The St. Louis Post-Dispatch employs other backup methods including optical media (DVDs) and hard drives, especially for photo and video content.

Although it is common for news organizations to utilize a CMS to hold their news content, such a practice puts their digital assets at greater risk due to the CMSs' greater focus on production. Both DAMs and MAMs offer more protection against loss of digital objects than CMSs. Although the phrase preservation repository is sometimes used to denote a storage system, it is unlikely that a news organization would utilize a full "preservation repository" for its content. It is important to note that although DAMs and MAMs can be used as part of a long-term access strategy, their use does not mean that an organization is preserving its content. Very often, DAMs and MAMs are employed at news organizations for their ability to provide short-term access to facilitate production processes. The word preservation indicates a much longer timeline.

⁶ "ISO 16363:2012: Space Data and Information Transfer Systems – Audit and Certification of Trustworthy Digital Repositories," February 2012, <https://www.iso.org/standard/56510.html>.

What is an archive?

The word archive gets used a lot and its meaning is frequently not agreed upon. The Society of American Archivists (SAA) acknowledges that “The most central term to the field of archives is also the most fraught.”⁷ As our research team spoke with reporters and editors, we heard frequent references to archiving news content in the CMS, DAM or other storage system or of retrieving material from an archive. For purposes of this report, we consider archives to be news content or records preserved because of their continuing value. This means a basic requirement for calling something an archive at a news organization is that it provides access to content for the long term; an archive’s purpose goes beyond the day-to-day needs of the newsroom where much of the raw material needed for production processes may be deleted shortly after it’s used. News archives keep content for the long-haul.

What is born-digital news?

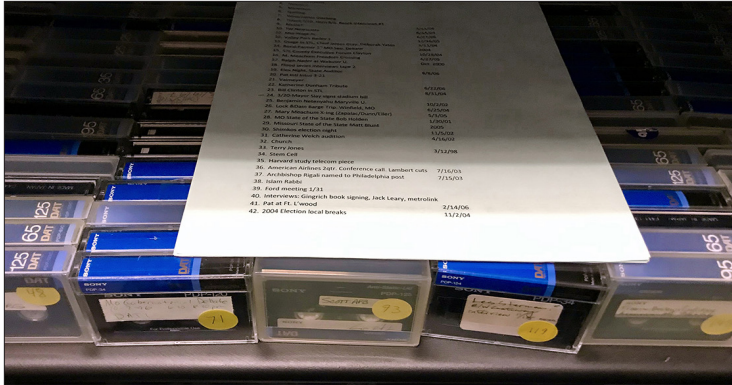
What do we mean when we talk about preserving born-digital news? First of all, when we use the phrase “born-digital” in this report we are not talking about digitized content such as scans or photographs of printed pages or conversions of analog recordings from magnetic audio or video tape into binary formats. In contrast to digitized materials, born-digital content has its origins in computer chips and sensors, along with the software programs that provide the instructions for how the hardware is supposed to behave, and is thus “born” digital. Virtually all news content is now created in digital formats, including but not limited to text, photos, graphics, maps, audio, video, databases and interactive elements. In some cases, born-digital content may be used in the production processes of analog content, such as printed newspapers. For purposes of this study, we consider all these kinds of news content as born-digital.

The shift from analog to digital

The shift from analog to digital media happened in phases over the span of decades. Much, probably most, textual news content has been born-digital since the 1980s due to the clear advantages of computerized systems that empowered reporters and editors to create and modify their stories from newsroom terminals. The text formats used by computers changed during this time, which undoubtedly caused some difficulties during system migrations and resulted in some loss of information, but generally speaking the archives composed from text survive largely intact and hold some of the oldest and most extensive records available from any sort of news archive.

By the late 1970s some of the first digital audio workstations arrived on the scene, but it took another 20 years to overcome limitations of high storage cost and slow computer processing speeds to make equipment suitable for use in radio newsrooms. The late 1990s saw heavy use of digital audio recorders and digital audio workstations for radio news production.

7 “Archives,” Dictionary of Archives Terminology, accessed February 18, 2021, <https://dictionary.archivists.org/entry/archives.html>.



Digital Audio Tape (DAT) cassettes stored in a cabinet with a list of their subject matter and date recorded, KWMU, St. Louis Public Radio, St. Louis.

Although much more complicated and requiring more memory and processing power than their text or audio counterparts, digital editing systems for broadcast television came into play in the late 1980s. At that time, video capture was still on analog tape. It was then converted to a digital format that could be cut and rearranged much more efficiently than using the analog processes of the day. Within the decade, professional digital video cameras allowed the seamless capture and transfer of footage without the need for conversions to digital formats. End-to-end born-digital video production became the industry standard in the early 2010s, likely accelerated by the Fukushima disaster in 2011, which destroyed the factories that provided broadcast-quality videotape.⁸

The Federal Communications Commission (FCC) approved digital broadcasting for U.S. radio stations in October of 2002.⁹ Broadcast TV newsrooms shifted from analog signals to digital signals starting in the early 1990s with the transition to digital ending in 2009 when the Federal Communications Commission (FCC) stopped allowing analog transmission for most stations. There was no similar mandate for radio broadcasters, but the FCC currently allows the simulcast of analog and digital radio signals.¹⁰

The role of PDFs in born-digital preservation

Although PDFs are often generated as part of the digitization process for newspaper pages, they also play a key role in the preservation of modern born-digital news content at legacy news organizations. Searches for historic news content at sites such as the Library of Congress' (LOC) Chronicling America retrieve PDFs displaying scans of newspaper pages.¹¹ This may result in the assumption that all PDFs represent analog news content. This situation is worth clarifying.

After seeing the advantages of working with digital text, print publishers gradually began incorporating other forms of digital technology into their production processes. As part of the process of transferring digital versions of fully-designed print pages from computerized design programs to the analog printing process, Adobe's Portable Document Format (PDF) was often utilized. PDFs are ideal for this purpose because they can hold a variety of digital objects, provide an accurate rendering of what the printed page would look like and can be used as a bridge from the digital realm to the physical

8 Carolyn Giardina, "Industry Scrambling After Japan Earthquake, Tsunami Lead to Tape Shortage," *The Hollywood Reporter*, March 20, 2011, <https://www.hollywoodreporter.com/news/industry-scrambling-japan-earthquake-tsunami-169456>.

9 Gary Krakow, "Radio Is Going Digital," NBC News, March 11, 2003, <https://www.nbcnews.com/id/wbna3078252>.

10 Kathleen A. Hansen and Nora Paul, *Future-Proofing the News: Preserving the First Draft of History* (Lanham: Rowman & Littlefield, 2017), 171.

11 "Chronicling America," Library of Congress, accessed February 18, 2021, <https://chroniclingamerica.loc.gov/>.

printing process. Thus, PDFs can be considered to contain born-digital news, even if they were generated as part of a print production process.

PDFs are also well-suited to preservation purposes, since—unlike web pages that can change over time—they are discrete digital objects and represent a stable snapshot of the page’s content at a given time. These properties allow them to be ingested and preserved in digital preservation systems—often the same systems that preserve digitized images of printed pages. For these and other reasons, over the past few years the Library of Congress shifted from collecting microfilm captures of print pages to requiring PDF submissions for mandatory deposits or copyright registration. The LOC collects between 300–350 daily papers from the United States in the PDF format.

However, saving PDFs alone is an insufficient answer to preserving born-digital news, since they contain only the print version of stories and leave out broadcast news sources entirely. Anecdotal evidence from this study’s interviews estimates that PDFs of newspaper print editions might contain somewhere between half and three-quarters of the news content represented at a given news organization’s website. While PDFs have the advantages of a standardized format that is well-suited to existing digital preservation models, they lack significant amounts of text and photo content published online, not to mention photo galleries, audio, video, interactive presentations, databases and more that can be delivered via web pages.

Archiving online news

A popular approach to preserving born-digital news is web archiving. With 525 billion web pages available through the Wayback Machine, the Internet Archive is the largest collector of web pages in the world.¹² The IA provides the infrastructure for much of the web archiving going on today, either directly or through contracts with the Library of Congress and other memory institutions. Web archiving technology makes a copy of one or more pages (URLs) and then encapsulates that information in an emulated environment designed to keep not only the data, but also the look and feel of the original pages. While web archiving represents a welcome addition to the arsenal of digital preservation methods for online news, it has its own limitations. For one, there are simply an unknown and enormous number of web pages out there and they can change quickly and without warning. Together with research partners such as the GDELT Project¹³, The Internet Archive has built a list of approximately 170,000 URLs from news organization websites in over 200 countries. However, even the Internet Archive doesn’t currently have the resources to find and preserve more than a sampling of those web pages. Even if the IA could preserve every page of news content, copyright laws put restrictions on who can access them and how they can be used or even archived. News organizations are also moving toward placing greater restrictions on public access by using paywalls, meaning anyone who wishes to see their content will need to purchase a subscription.

Television and radio news preservation

Significant efforts have been made to preserve TV and radio news broadcasts, beginning with the Vanderbilt TV News Archive (VTVNA) which started collecting analog content on August 15, 1968. It has collected daily news broadcasts of major networks since that time. VTVNA shifted to acquiring digital

¹² “Internet Archive,” Internet Archive, accessed February 23, 2021, <https://archive.org/>.

¹³ “The GDELT Project,” The GDELT Project, accessed February 23, 2021, <https://www.gdeltproject.org/>.

content since the U.S. broadcast industry transitioned in the late 2000s. Today VTVNA is a unique source of approximately 80,000 video files of news programming, including some 1.2 million stories and clips which occupy about 250 Terabytes of storage space.

Other broadcast preservation initiatives include the Corporation of Public Broadcasting-launched American Archive of Public Broadcasting (AAPB). AAPB is now a collaboration between the Library of Congress and GBH Boston (formerly WGBH) to create and maintain a digital archive for public broadcasting in the U.S. The project will transcode about 5,000 hours of born-digital files, with an additional 40,000 hours of public TV and radio to be digitized.¹⁴ A separate AAPB project to archive every episode of PBS NewsHour plans to ingest born-digital files in the future, after digitizing the videotapes recorded prior to 2015.



Karen Cariani, Executive Director of the GBH (formerly WGBH) Media Archives, retrieves an item from the GBH archive vault which mostly holds film and other analog media.

In addition to capturing web pages for the Wayback machine, the Internet Archive also has two other media preservation efforts, the TV News Archive and the Radio Archive. Each of those programs captures video or audio media formats directly from the web, replicates them in multiple locations and provides a means of searching for and accessing content.

Preserving the opportunity for long-term access

The essence of born-digital news, at once mercurial and never-ending, presents formidable challenges to transforming it into a stable, permanent record. Contemporary news content has its genesis in digital systems that are designed for creation, editing, production and distribution, not for providing long-term access. Previous practices for archiving print versions of news content are not adequate for preventing the loss of born-digital news. Many news organizations have been producing and publishing digital content for decades. Over this time, for a variety of reasons, CMSs have often been called on to play the role of digital archive, one for which they are generally not designed. As a result, in today's newsrooms a great deal of digital news content is being stored in production systems by default and without a plan for long-term access or preservation. Given this situation, when it comes to born-digital news content, a significant part of our strategy might be to do our best to preserve the opportunity for future access—simply not to lose what's left. On the other hand, there are a good number of news organizations that are succeeding in their quest to make sure their digital "first rough draft of history" persists for readers, listeners and viewers in 2121 and beyond.

¹⁴ "American Archive of Public Broadcasting Permanent Entity Grant," American Archive of Public Broadcasting, accessed February 23, 2021, <https://americanarchive.org/about-the-american-archive/projects/permanent-entity>.

How was news preserved before content came in bits and bytes?

To fully grasp the challenge of preserving digital news it helps to understand how preservation happened in the print and analog broadcast era. A surprising amount of news from that era has come down to us. The photo morgues of several major newspapers such as The New York Times, Denver Post, and The Chicago Daily News, and the video and film archives of The Associated Press (AP), National Public Radio (NPR) and the BBC have survived in the care of those news organizations. And, incredibly, at least a few issues of most American newspapers known to have been published in the last three centuries can still be read today in research libraries, historical societies, and other memory institutions. And those collections are growing. The American Antiquarian Society alone, through gift and purchase, adds an average of 15,000 issues a year to its holdings. These sources have been mined over the years for countless books, documentary films, and museum exhibitions.

The survival of so substantial a public record is due to the combined efforts of the news industry, academia and government. The three sectors put in place interlocking mechanisms that together created a lively secondary market or afterlife for broadcast and print content.¹⁵

Naturally, the news industry itself played a key role. Organizations such as The New York Times and the BBC maintain extensive archives of published and unpublished content as a matter of record, and for the use of newsroom staff in search of historical background for current reporting. Other organizations that have become defunct, including the Chicago Daily News and Rocky Mountain News, eventually entrusted their archives to historical societies and public libraries for long-term safekeeping. Still others were lost, casualties of the growing pressure to streamline newsroom operations and reduce costs.

National and state libraries, big city public libraries, and major academic and independent research libraries all had a role. They systematically collected newspapers and documented their publication histories in catalogs and in tools like Winifred Gregory's American Newspapers, 1821–1936 union list. Two mechanisms largely drove those efforts: legal deposit laws related to copyright and library patrons' demand for current news.

Thousands of local and national publishers routinely deposited copies of their dailies and weeklies at the Library of Congress and other national libraries. In return they received copyright protection. University libraries and large urban public libraries subscribed to local and national newspapers to provide access for students and the public, and for scholars. Papers so acquired eventually piled up in some libraries and form the core of massive collections and historical archives mined by academic researchers, local historians and genealogists.

During the postwar period microfilming emerged as a practical means of alleviating the inevitable storage crunch, and as a medium for broadening distribution of newspapers to libraries. Miniaturized page images on film soon replaced mountains of cumbersome and embrittled paper previously maintained largely at publisher and library expense. And with the explosive growth of American universities in the 1950s and 1960s, demand for news from all world regions surged and gave rise to a robust microfilm distribution industry. Companies such as Readex, IDC and ProQuest answered

¹⁵ Bernard Reilly, "The Library and the Newsstand: Thoughts on the Economics of News Preservation," *Journal of Library Administration* 46, no. 2 (2007): 79–85, https://doi.org/10.1300/J111v46n02_06.

the demand for these miniature formats and generated a modest but significant secondary stream of revenue for newspapers. Some, like Readex's parent company NewsBank and ProQuest, went on to provide online storage and rudimentary content management services for publishers, serving as supplementary archives and taking on some of the functions of the in-house newsroom libraries. They also managed to assemble complete, or nearly complete, runs of thousands of newspaper titles.



Research team members learn about the archive workflow including digital asset management from Archivist/Researcher Jennifer Selph at the St. Louis Post-Dispatch.

In the early 1980s the National Endowment for the Humanities (NEH) created the United States Newspaper Program, a multi-decade endeavor that eventually gathered and preserved on microfilm local newspapers from all fifty states.¹⁶ The NEH's National Digital Newspaper Project followed on that work beginning in 2005, and seeks to create a comprehensive, searchable database of U.S. newspapers and to begin digitization of the early content.¹⁷

Libraries and historical societies also inherited and preserved the photo morgues of local news organizations that downsized or closed. When The New York World Telegram and Sun shut down in 1967 the Library of Congress salvaged and archived its morgue of about a million photographs.

Libraries became the logical resting places for broadcast news as well. In 1968 the Vanderbilt Television News Archive, now a program of Vanderbilt University Libraries, began recording evening news broadcasts from the three major U.S. television networks: ABC, CBS and NBC, later adding CNN and Fox News. The resulting archive of broadcasts (on tape and later digital files) were stored and cataloged by Vanderbilt and serve as a source of historic footage for the networks themselves.¹⁸ Both Purdue University's C-Span archive, started in 1987, and the AAPB, instituted in 2013, were created with public and private funding to preserve the video and audio outputs of public television and radio. Much of the AAPB work to date has focused on conversion to digital media of video broadcasts produced during the analog era, including more than 14,000 PBS NewsHour programs, some dating back as early as 1975.

More recently, news organizations such as the Charlotte Observer, under McClatchy ownership, needed a home for physical news archive materials such as clips files and microfilm when it vacated downtown offices. Fortunately, an agreement was quickly reached to donate these materials to the Public Library of Charlotte-Mecklenburg.

¹⁶ "U.S. Newspaper Program," National Endowment for the Humanities, accessed February 23, 2021, <https://www.neh.gov/us-newspaper-program>.

¹⁷ "National Digital Newspaper Program," National Endowment for the Humanities, accessed February 23, 2021, <https://www.neh.gov/grants/preservation/national-digital-newspaper-program>.

¹⁸ Marshall Breeding, "Building a Digital Library of Television News," *Computers in Libraries*, June 2003, <https://librarytechnology.org/document/10346>

Curatorial efforts, and sustained investment, over half a century by public, private and academic sector actors operating in the orbit of the news industry, aided the preservation of news for over half a century during the analog era and created an enduring record of three hundred years of American and foreign journalism.

First recognition of the problem

In the face of the unprecedented volume, velocity, and variety of digital media output, the print and analog broadcast systems began to falter. There was an obvious mismatch between the traditional approaches, designed as they were around fixed objects—the article, edition, broadcast — and the technical realities of networked, dynamic digital media.

Legal deposit, for one, began to fail as a wholesale means of supplying national libraries copies of domestic newspapers. Copyright is built around the notion of discrete “works”, i.e., finite packages of information, like the daily newspaper and the nightly news broadcast. Deposit libraries have had difficulties retooling to cope with the constant stream of dynamic content delivered by the web and the unbroken news cycle. Trying to keep pace, the U.S. Copyright Office recently changed to allow a PDF file of the static pages of the print edition, which the Library of Congress now receives but lacks the right to make available beyond its own premises. The Digital Millennium Copyright Act of 1998 imposed new restrictions on the ability of libraries to distribute digital news content by extending the duration of the copyright by publishers.¹⁹

Digital delivery of news content has reduced both the demand for major research libraries to provide access to current news from local collections and their ability to do so. Contemporary researchers prefer the convenience of online aggregator databases and round-the-clock access to online sources and real-time reporting. At the same time, libraries generally lack the resources and technology to capture digital news content and the legality of such collection is murky.

Companies such as ProQuest, Readex, and Gale reinvented themselves in the 1990s as purveyors of databases aggregating newspaper content in digital form. While their products do afford 24/7 online access and impressive searchability, much of their content is derived from print newspaper page-image files rather than born-digital source materials and lacks the rich underlying functionality of the original online news.

Framing the problem: focusing on born-digital news content

Threats to the long-term survival and integrity of digital information emerged in the 1990s, starting at NASA. The space agency experienced irrecoverable losses of electronic data gathered during missions in the 1970s due to obsolescence of storage media and software. By the mid-1980s the torrent of images and other digital data from NASA’s programs overwhelmed the agency’s ability to keep it.

NASA’s dilemma set the stage for some of the earliest efforts at digital preservation. In 1995, the international Consultative Committee for Space Data Systems (CCSDS) began to coordinate “the development of standard terminology and concepts for the long-term archival storage of various types of data.” Under CCSDS, experts from academia, government, and private research worked

19 “Digital Millennium Copyright Act,” American Library Association, January 24, 2019, <http://www.ala.org/advocacy/copyright/dmca>.

together to develop what is now called the Open Archival Information Systems (OAIS) Reference Model. First published in 2002, the OAIS Reference Model provided a plan for architectures, standards, and protocols for system design, metadata requirements, assessment, and other issues central to digital preservation.²⁰

That same year the Columbia Missourian, owned and operated by the University of Missouri since the establishment of the Missouri School of Journalism in 1908, experienced a server crash that wiped out fifteen years of textual content and seven years of photographs held in the news organization's CMS. The massive loss of mid-Missouri history prompted the School of Journalism and the University of Missouri Libraries to task Victoria McCargar, a journalist, librarian and early digital news preservation expert, with probing the underlying causes of the loss. McCargar found that certain conditions at the Missourian—a faulty backup system, key staff attrition, and others—were prevalent throughout the news industry. Among her suggestions to the emerging School of Journalism's Donald W. Reynolds Journalism Institute were:²¹

- Own the problem of preserving born-digital news content and lead the response to it
- Foster outreach efforts such as conferences or a symposium on the topic
- Create a Digital Curator of Journalism position to oversee efforts locally and industry-wide

Early digital news preservation efforts

In December of 1994, recognizing the challenges digital media posed to libraries and archives, the Commission on Preservation and Access and the Research Libraries Group created a Task Force on Archiving of Digital Information to articulate the challenges and identify solutions. The task force consisted of individuals drawn from industry, museums, archives, libraries, publishers, scholarly societies, and government. Their landmark 1996 report described a number of key problems, organizational, technological, legal and economic, needing to be resolved to ensure “continuing access to electronic digital records indefinitely into the future”. The task force recommended a number of measures to resolve such problems.²²

The report laid the foundation for subsequent planning at the national level. In 2000 the U.S. Congress appropriated \$100 million to develop and implement a strategic plan and program to enable the “sorting, acquisition, description, and preservation of electronic materials.” Within the broad scope of the National Digital Information Infrastructure Preservation Program (NDIIPP) mandate, the preservation of digital news received some modest attention at first. Considerable NDIIPP funding was invested in collecting content from websites, and the Library targeted sites reporting news about congressional races, major events like 9/11, and other topics of national interest. To date much of the harvesting has been outsourced, relying on the San Francisco-based nonprofit, the Internet Archive.

20 Brian Lavoie, “The Open Archival Information System (OAIS) Reference Model: An Introductory Guide (Second Edition)” (Digital Preservation Coalition, 2014), <https://www.dpconline.org/docs/technology-watch-reports/1359-dpctw14-02/file>. The OAIS reference model was approved in January 2002 as ISO International Standard 14721; a revised and updated version was published in 2012 as ISO Standard 14721:2012, and the model serves as the frame of reference for all subsequent digital preservation metrics, including SPOT analysis.

21 Victoria McCargar Consulting and Victoria McCargar, “Missouri J-School and the ‘Backstory,’” Report (Victoria McCargar Consulting, 2008), <https://mospace.umsystem.edu/xmlui/handle/10355/45033>.

22 Donald Waters and John Garrett, “Preserving Digital Information: Report of the Task Force on Archiving of Digital Information” (Commission on Preservation and Access and The Research Libraries Group, May 1, 1996), <https://clir.wordpress.clir.org/wp-content/uploads/sites/6/pub63watersgarrett.pdf>.

National libraries in other countries, including the British Library and the Bibliotheque nationale de France, have since adopted web archiving to supplement or as a substitute for legal deposit.

As a preservation solution web archiving displayed certain limitations. The harvesting engine normally skims content, from home pages for example, and cannot reach deeper regions where more and more content now resides behind paywalls and authentication systems. Many archived links and multimedia content are non-functional. Rather than a snapshot of a complete website as it exists at a given moment, pages are often gathered from the same site at different times. In effect, good at capturing individual websites, but not the sprawling platforms and highly-enriched content maintained by today's news organizations.

In September 2009 the Library of Congress convened a workshop Preserving Digital News, to explore possible strategies for the systematic preservation of digital news content. Library and industry experts gathered at the event concluded that devising an effective national strategy for preserving news would require a fuller understanding of the lifecycle of digital news than currently existed.

The Library's Office of Strategic Initiatives commissioned the Center for Research Libraries (CRL) to document that lifecycle. CRL examined the workflows for born-digital news content, media and data at four major newspapers from end to end. The study found that the nature of the digital lifecycle and web-first news workflows were of a complexity that would challenge traditional library preservation methods. The study further concluded that the news organizations themselves were better positioned for long-term management of their own content than libraries and archives.²³

In 2011, in response to the earlier loss of digital content at the Columbia Missourian, the University of Missouri Libraries and the Donald W. Reynolds Journalism Institute launched the first in a series of conferences that brought together the news industry and library worlds. The initial event, called the Newspaper Archive Summit, attracted 125 professionals from stakeholder communities in the United States and Europe to focus on issues such as orphaned, lost and born-digital newspaper archives. From 2014 through 2017, five Dodging the Memory Hole forums, research projects and outreach initiatives illuminated the operational challenges the industry faced in preserving its digital content, engaging journalists, scholars, technologists and digital preservation experts in a network of specialists to identify areas for specific work on born-digital curation: awareness, legal framework, policies, resources, standards and practices and technology.

In November 2017, WGBH and the Council on Library and Information Resources (CLIR) convened a small group of library and archival professionals, technologists, and representatives of funding organizations who are long-term supporters of public television and radio for a discussion of preservation strategies for public broadcasting content. Participants reviewed the current state of archiving efforts, discussed diverse use cases for public media archives, and explored potential models of public-private partnerships that could replace the mechanisms that sustained archiving in the analog era. It was the consensus that the goals of preserving and providing access to born-digital content would have to be led by the news industry itself.²⁴

23 Jessica Alverson et al., "Preserving News in the Digital Environment: Mapping the Newspaper Industry in Transition" (Center for Research Libraries, April 27, 2011), https://www.crl.edu/sites/default/files/d6/attachments/pages/LCreport_final.pdf.

24 "Sustaining Public Media Archives: Summary of Sustainability Discussion Hosted by CLIR and WGBH" (WGBH and the Council on Library and Information Resources, November 2017), <https://clir.wordpress.com/wp-content/uploads/sites/6/2016/09/Sustaining-Public-Media-Archives.pdf>.

Alarms continue to sound regularly about threats to the survival of the journalistic record. In 2019 a Tow Center for Digital Journalism survey of 21 news organizations found that “the majority of news outlets had not given any thought to even basic strategies for preserving their digital content, and not one was properly saving a holistic record of what it produces.”²⁵ Our project is an attempt to answer that call to the industry.

²⁵ Sharon Ringel and Angela Woodall, “A Public Record at Risk: The Dire State of News Archiving in the Digital Age” (Tow Center for Digital Journalism, Columbia University, 2019), <https://academiccommons.columbia.edu/doi/10.7916/d8-7cqr-q308>.

3

Methodology

For more than a decade the University of Missouri Libraries and the Donald W. Reynolds Journalism Institute have demonstrated a commitment to examining and addressing issues related to the loss of born-digital news content. With a grant from The Andrew W. Mellon Foundation, a team of librarians, archivists, journalists, and technology and audio-visual specialists engaged in this research project to identify key factors in the technology stacks, workflows and policies that affect news organizations' ability to preserve born-digital news content and to understand best practices for properly archiving and retrieving that content.

Two research questions guided our work:

- 1:** How are print, digital, radio and television news organizations in the United States and Europe preserving born-digital news content?
- 2:** What are the best practices, problem areas and changes needed to avoid unintentional loss of content by news enterprises?



The research team at a planning meeting in Oct. 2019 in the Journalism Library at the University of Missouri, Columbia, Missouri.

We interviewed a variety of news organizations and related enterprises

To answer our two research questions, the team created a list of United States, the United Kingdom, and European news organizations we hoped to interview. Organizations were organized into three categories: News Producer Type, Legacy/Digital Native and a Geographic Coverage designation. Each grouping had its own subdivisions.

- **News Producer Type** indicated the distribution channel or combination of channels used by each organization and was further categorized into one of three subcategories. "Broadcast/web" news producers used both the airwaves (for radio or television or both), cable and the web to deliver

news content. “Web/print” news producers used both the web and print media (newspapers or magazines) to deliver news and then there were news producers who only used the “web” to deliver news content.

- **Legacy/Digital Native** identified those news organizations that had begun distributing content through analog channels and then added digital, as “legacy” news organizations. If the news organization had originated in the digital age and has since distributed its content only through digital channels, we gave it the “digital native” label.
- **Geographic Coverage** defined a news organization’s territorial distribution in one of three ranges: “regional” (encompassing wider coverage and distribution than the metropolitan area which is home to the outlet), “national” (nation-wide coverage and distribution) or “international”, (global news coverage and distribution).

Categories and terms used to describe news organizations in this study:

NEWS PRODUCER TYPE	LEGACY/DIGITAL NATIVE	GEOGRAPHIC COVERAGE
broadcast/web	legacy	regional
web/print	legacy	national
web	digital native	international

The research team also interviewed organizations and institutions that play significant roles in production, workflow, distribution and/or preservation of news content, to provide a better overall picture of what technologies were available; how effectively those technologies were used; and what role third party aggregators and memory institutions play in the preservation process.

Some of these additional interviews were intended to enrich our understanding of the roles played by news technology vendors such as Denmark-based Stibo DX, SCC in Atlanta and MerlinOne in Boston. The businesses we spoke with provide content management and digital media asset management tools for production, storage and access that are vital in managing the workflow of news content.



Standards editor Margaret Holt leads the research team on a visit to the Chicago Tribune newsroom in January 2020.

We also interviewed news aggregators, such as Newspapers.com and NewsBank, engaged by news organizations to provide storage and access to published news content both internally and externally. These enterprises often process and provide access to news content via digitized versions of print content, born-digital renderings used in the printing process, or ASCII text. Both the digitized and born-digital page-image content are often kept in the PDF format. Intentionally or not, news aggregators often seem to serve the role of surrogate archivist for news organizations.

In addition, we interviewed memory institutions, such as the Library of Congress, Vanderbilt Television News Archive and the Internet Archive, involved in archiving or preserving published news content, to have a better awareness of what digital news content is currently being preserved, how that process works, and the challenges involved.

Our ambitious and diverse list eventually was narrowed to include only those with whom we could set up interviews during a pandemic year. We are grateful for the expertise and time shared with our team from individuals at the following organizations:

ABC News	Invisible Institute	PBS NewsHour
The Associated Press	KBIA (NPR)	QCity Metro
Baltimore Afro-American	KOMU	Quincy Media
BBC	KWMU (St. Louis Public Radio)	Software Construction Co. (SCC)
The Boston Globe	Library of Congress	St. Louis Post Dispatch
Center for Research Libraries	Los Angeles Times	Stibo DX (CCI)
Chicago Sun-Times	McClatchy Corp.	Stars and Stripes
Chicago Tribune	MerlinOne	TownNews
CNN	Netherlands Institute for Sound & Vision	Vanderbilt TV News Archive
Columbia Missourian	NewsBank	Vox Magazine
The Dallas Morning News	Newspapers.com	Vox Media
Gannett	Newsy	The Washington Post
GBH (formerly WGBH)	NPR	
Internet Archive	Obsidian Collection	

From a total of 40 organizations where we conducted interviews, we eventually selected 24 news organizations to include for our analysis. Inclusion in this list of 24 was based on those newsrooms that provided the most complete data set.

Our methodology evolved in response to COVID-19

Initially, our scripted questions focused on workflow and how newsrooms used technologies, following news content from creation through production stages to distribution and storage or preservation.²⁶ Interviews were always conducted with the permission of the interviewees and in compliance with the requirements of the University of Missouri Institutional Review Board (IRB), whose policies and procedures strive to advance research that is fair and ethical.²⁷ Almost all interviews were recorded and if that was not possible, notes were taken. The interview phase of our research began in September 2019 with in-person visits to the newsrooms of our local Missouri School of Journalism news outlets and a local start-up. We used these early interviews to educate ourselves about different newsroom workflows and technologies involved in the production of content for web, print, television and radio. This knowledge helped us refine and improve our interview process. Throughout this process, we conducted interviews with multiple participants to reflect different experiences and perspectives of the journalists involved. Quotes used in the report, though anonymous, are actual comments during interviews.

In-person visits to McClatchy Company offices in North Carolina in November 2019 provided more answers to our scripted questions. A third set of interviews in January 2020 were part of a regional

²⁶ See Appendix A.

²⁷ "Institutional Review Board," University of Missouri Office of Research and Economic Development, accessed February 19, 2021, <https://research.missouri.edu/irb/>.

road-trip encompassing Missouri and Illinois, with visits to the newsrooms of the St. Louis Post-Dispatch, TownNews and Quincy Media. A fourth set of interviews in Chicago followed in February 2020, including on-site interviews with the Chicago Sun-Times, the Chicago Tribune, the Invisible Institute and The Center for Research Libraries. A visit to Boston in early March of 2020 marked the fifth and last round of in-person interviews, which included WGBH (now GBH), MerlinOne and The Boston Globe.



The WGBH (now GBH) newsroom as of March 2020 when the research team visited the large PBS operation in Boston.

By mid-March 2020, the COVID-19 pandemic shut down travel, providing an opportunity to regroup and refine our strategy. At that point, after visiting with 18 institutions and conducting 43 face-to-face interviews, we realized that we should take some time to assess the data we had collected and to review what we had learned up to that point. For example, on a practical level, we more fully recognized that recording multiple voices talking at the same time made taking notes and transcribing interviews much more difficult. In addition, over time it became increasingly clear that future interviews would need to be conducted using videoconferencing software, ideally with one person at a time. This required more precise scheduling than on-site visits and greater attention on everyone's part to speaking clearly into their microphones and not talking at the same time.

The team conducted a mid-course review, led by AVP

In order to facilitate the review process, team members gathered our transcripts and notes to date and identified some useful directions for further structuring and analyzing our data. Consultants from AVP who were part of the research team provided invaluable assistance as we reworked and refined our set of questions so that answers could be more definitive and quantifiable. To answer our two research questions, we needed to know more about news organizations' mission and policies for preservation, content production and distribution, tools for access and retrieval and storage technologies. To accomplish that our interview questions would now include more specific information about the news outlets, divided into four main areas:²⁸

- **Organizational Infrastructure**
- **Content and Collections**
- **Systems, Search and Metadata**
- **Storage and Management**

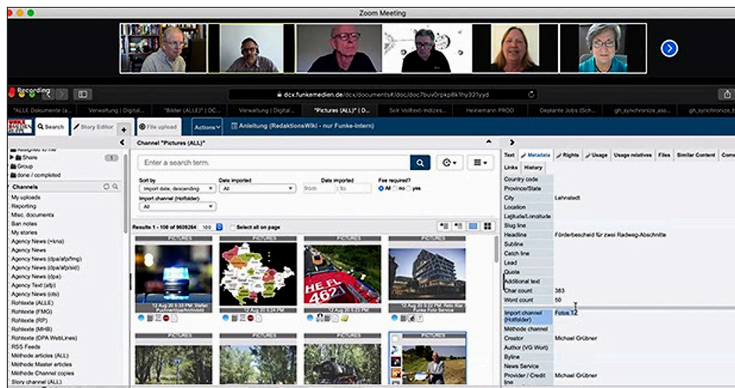
²⁸ See Appendix B.

An additional set of questions was developed specifically for vendors, aggregators and memory institutions. Each of these five topic groupings would include multiple questions that would elicit specific information which could be more easily quantified. A Google spreadsheet, called our “data tracking sheet” tabbed for each category and containing rows of questions specific to the particular category was created. We populated it with columns for each news organization and their responses to each question.

In addition to providing an improved conceptual framework for developing questions and processing interview data, we found that organizing groups of questions in this way allowed us to target certain questions to specific interviewees, depending on their area of expertise. For example, it was generally more productive to ask senior management questions about organizational infrastructure and to query the IT department about storage and management.

At times it was necessary to retroactively fill in all of the answers to the new questions with information from the previous set of more exploratory interviews, since many of the older questions had not been specifically asked. In cases where answers from either older or newer interviews were lacking information or missing, the team reread and analyzed all of the notes and transcripts, populating rows in each organization’s column with answers to the new questions. Some would remain empty, while some were either gleaned from the transcripts or sought through follow-up conversations.

Going forward, when possible we scheduled interviews with one person at a time rather than groups, which was the earlier default. This change made reviewing transcripts and gathering and analyzing the data much easier. However, COVID-19 and the 24/7 news cycle for the upcoming presidential election put a strain on the time journalists had available to be interviewed. Ultimately, these unusually difficult circumstances limited the number of new interviews we were able to schedule.



Screen capture of videoconferencing interview with Thomas Ammerman of Digital Collections, in Germany.

Immediately after each interview, the team met to talk about key takeaways from each interview. Next, the automated speech-to-text transcripts were reviewed, identifying speakers and cleaning up mis-transcriptions. Cleaned up transcripts, notes and any graphics or screen shots from the interview were added to Google Doc interview files and folders. Team members would then extract the relevant data from the interviews and add it to the data tracking sheet. In an attempt to gain a more complete understanding of the news organizations, if information from interviews was sparse or missing, the research team attempted to supplement it with additional research. Researchers would also quantify some of the data and represent it graphically.

As the data tracking sheets were being completed, team members began the process of quantifying the collected data in spreadsheets for numerical analysis. Key to this process was normalizing the interviewee responses, making sure to consistently apply definitions so that the data values were assigned using a common scale. In some cases, this required consulting the interview transcripts to gather additional context or background information. Sometimes the interview questions had multiple parts which resulted in complex responses. In those cases, the team broke the questions and answers into segments (such as Q1A, Q1B and Q1C) that could be more clearly defined and counted. In some cases, Q7 for example, “What is the primary news content the organization produces?” the answers were not consistent or uniform enough to apply a consistent structure to. In those cases, the data was interpreted using a qualitative approach.

After the spreadsheets were complete, pivot tables were utilized so that the data could be further analyzed using filters for News Producer Type, Legacy/Digital Native and Geographic Coverage, if needed. In addition, tables and charts were produced to provide the research team with graphical representation of the underlying data. Most of the tables and charts found later in this report are derived from this question-by-question quantification of the interview data. The researchers have included reference to the interview question or questions that were the main source of data for tables and charts.

Assessing digital news preservation using the SPOT model

In examining the systems employed by news organizations which hold and allow access to their content, the team decided to use the Simple Property-Oriented Threat (SPOT) Model for Risk Assessment²⁹ because it describes the threats to digital preservation and focuses on preservation outcomes rather than specific causes of a threat. We gravitated toward this more flexible model because the systems in place at news organizations were not designed with content preservation in mind, but many of their functions align with preservation functions as identified in the SPOT Model.

For purposes of using the SPOT model for the news organizations in this study, the team identified the internal users of the news organizations in question as the primary designated community. In short, these internal users’ main interest is in access to digital content for their own organization’s use and reuse. (For more about designated communities, see “For whom are we preserving” in Background.)

The SPOT model identifies six essential properties of well-preserved digital content:

- 1: Availability**
- 2: Identity**
- 3: Persistence**
- 4: Renderability**
- 5: Understandability**
- 6: Authenticity**

29 Vermaaten, Sally, Brian Lavoie, and Caplan, Priscilla, “Identifying Threats to Successful Digital Preservation: The SPOT Model for Risk Assessment,” *D-Lib Magazine* 18, no. 9/10 (October 2012), <https://doi.org/10.1045/september2012-vermaaten>.

A detailed explanation of how the research team interpreted the SPOT model and applied it to content from the news organizations in this study is available.³⁰ In addition to viewing the data through the lens of the SPOT model, the team analyzed a variety of data during a workshop led by AVP. This analysis helped us identify patterns from our findings and to distill the collected information into higher-level “meta-findings,” which helped the team answer our research questions and create observations and recommendations.

Over an 18-month period, the team interviewed 115 individuals from 29 news organizations, four news technology companies, two news aggregators and five memory institutions. This represents an estimated total of 188 hours of interviews and hundreds of hours reviewing transcripts and analyzing data.

³⁰ See Appendix C.

4

Findings

The state of digital news preservation today

This section presents the main findings of our report. We include findings in the form of quantitative values and comparisons where they could be discerned from our interviews with news media organizations. This section also includes other, qualitative and anecdotal findings, where we found patterns and trends worth reporting even if they could not be distilled into numbers.

Findings are based on interviews with news media organizations that covered a wide range of questions (see Appendices A and B for complete list of survey questions) about born-digital news content preservation, from policies and purposes to systems and workflows, technology stacks and metadata, along with descriptions and extents of the content types saved up to now.

Here's what we found, starting with findings that illuminate the current state of digital news preservation in today's news media. The first section below called **Content** shares information on current preservation, what is being saved, what types of digital content, and how trends in digital publishing are affecting these processes.

Next we cover **Technology** in a section that shows the types of publishing systems and content management system (CMS) platforms, digital asset management (DAM) systems and other tools now in use. We also offer insights into how each affects the ways that news content is saved, and how variations in the ways that publishing technologies are implemented in each newsroom can be highly influential in preservation outcomes.

Following this, the findings turn to **Practices** that have emerged in response to the enormous financial pressures affecting the industry, the role of digital workflow and the potential issues in how news organizations sometimes use their web CMS, and their impact on preservation.

The final group of findings, a section called **Mission**, covers the difference it makes when an organization spells out preservation as part of its mission. This section also looks at related factors such as strong, clear policies on preservation, a commitment to the community and track record on preservation, and how company culture impacts all of these efforts.

Content Findings



Finding 1

Newsrooms save some but not all digital content

Once we analyzed the interviews and responses, one finding jumped out as especially surprising: the simple fact that every news organization we interviewed is saving digital news content now being produced, at least partially. Among the 24 news organizations we interviewed in depth, original news content is not being deleted, neither intentionally nor on a large scale.




We learned that the content news organizations are saving may not be everything they produce or publish. In many cases it may not be the highest quality in size or resolution for visual or audio content. It may not have the most elaborate metadata. And it may be saved in a web CMS or DAM system rather than a dedicated preservation platform. But the core content you see today on news websites and apps, and on TV and mobile devices, stories and photos especially, and increasingly video as well, is being saved in one form or another, at the 24 newsrooms we interviewed in depth.

Figure 1: How completely does your organization preserve news content?

	NEWSROOMS RESPONDING YES
Yes, fully	7 
Yes, partially	17 

Source: Interview responses to part of Question 1; see Appendix B for full question text.

Figure 2: Extent of saved content

	NUMBER OF NEWSROOMS THAT SAVE EACH
Final published/broadcast	17 
Everything	5 
Partial	2 

Source: Interview responses to part of Question 11; see Appendix B for full text of questions.

Meaning, impact, observations: What does this result mean? One basic but important question we considered was: does digital content persist beyond an immediate production period in today's news platforms? This question is also the basis for the interview questions and analysis that followed. The responses we received were universal: Yes, all the news organizations we talked with are keeping at least some of the content they are producing, the parts they consider to be essential to telling stories.

We consider this significant. So long as the content continues to exist in digital form, the potential remains to take additional steps where needed to ensure long-term preservation. It may require additional or different technology, it may need assistance from an outside organization, grants or other funding to accomplish. But the potential is there if the content exists.

Such findings are encouraging, even if they differ somewhat from existing literature and studies on the topic, many of which raise alarms about the degree to which news organizations fall short of what many digital preservationists believe is needed to ensure long-term survival of news coverage as part of the public record. That was our starting assumption as well, and the thrust of a number of papers and conferences over the past decade done under the auspices of our sponsors, RJI and the University of Missouri Libraries. We began with that perspective: that as the news industry made the shift to digital, in many cases it had ceased to fulfill an essential role that's critical to our communities and the public record at large, the task of saving their news content, all of it.

Content

But as we began to interview and meet with news organizations, we learned this did not reflect the situation in today's news organizations. For one, what we saw is that newsrooms are saving content in some form, albeit limited forms in many cases. But it also became clear that major improvements in these practices, likely requiring investments in preservation or staffing, may be unattainable in the face of current financial challenges.

In every newsroom we visited we heard and saw evidence of deep economic distress: staff layoffs left-and-right; products and channels shut down, even profitable ones because there's nobody left to produce them; the closing and selloff of newsroom and press buildings, many of them local architectural landmarks, and downsizing to leased office space that is usually a small fraction of their former space.

In newsrooms that only a year earlier were full, it was not uncommon to see half the desks empty, sometimes more. And in response to the pandemic, we saw numerous newsrooms let go of offices completely as they shifted to full work-from-home operations. In short, the industry is beset by what seems nearly insurmountable financial difficulties. This is especially true with the news organizations that still produce the largest share of today's news coverage, what we used to call local newspapers and now often call legacy media.



Empty desks at the former offices of the Tucson Citizen, which ceased publication in 2009 after 139 years.

Most of those we interviewed were unaware or only vaguely aware of the fragility of digital content, although they could often cite examples of how recent digital systems changes resulted in losses that affected presentation or rendering quality, and how parts of stories such as images or database components become disconnected and fragmented. More importantly, we saw little awareness that there was anything that could be done in their systems, workflows or policies to address these issues.

The lack of awareness was confirmed through our interviews and meetings. We heard again and again the generalized assumptions that existing systems seemed sufficient, that anything published on the internet is being saved by "somebody out there," even if most were unaware of who that was. We heard statements that the IT department is doing backups, and isn't that enough? Or that the Internet Archive or the Library of Congress or "somebody like that" is surely saving everything. Most we spoke with had never considered preserving social media posts, podcasts, data journalism projects and other relatively new content types.

This perspective in the news industry is not too dissimilar from the general public's assumption when it comes to digital information, that it's all out there somewhere and one does not need to take any deliberate steps.

Content

In contrast to the general pattern of practices we saw, we did interview some news organizations that are doing extensive, remarkably comprehensive work for news content preservation. The Associated Press, for example, has built and maintained an extensive digital asset management operation for its content over the past nearly two decades; the BBC, NPR and GBH (formerly WGBH) all run significant preservation operations, as does CNN for its cable news video content.

While they may offer models for other media organizations to try to emulate, they are still the exception. Our interviews confirmed that most news organizations today are taking only limited steps to preserve born digital news content. Much of the efforts that remain are holdovers from the pre-digital era, when well-established analog preservation processes were successful in saving print content. Few have been able to update their processes to establish the deliberate activities needed to properly preserve digital content for the long term.

Finding 2

Saved content is mostly text, images, video

One focus for our interviews was to understand the content typical news organizations are preserving; what types of content, how broad a range of content types is saved, and what gaps and limits there might be on any set of content across the years.

To get the broadest possible picture, we met with and interviewed news organizations across the spectrum of today's news media landscape. In the media segment of newspaper or legacy print media we interviewed long-established news organizations such as the Chicago Tribune and The Boston Globe, the 30-newspaper group McClatchy, the 75-newspaper group Lee Enterprises and the 128-year-old Baltimore Afro-American. We also interviewed broadcasters in radio and TV, including KOMU, an NBC affiliate in Missouri; GBH, the large public television operation in Boston; KWMU, the primary NPR radio station in St. Louis and Quincy Media in Quincy, Illinois which counts 30 television stations among its media holdings, and is planning to sell them. We also met with a small digital-only startup, QCity Metro, which serves the Black community in Charlotte; and with Vox News, a large national media outlet that's part of Vox Media.



The Master Control room at WGBH (now GBH) in March 2020. The monitors on the wall display the outgoing broadcast signals for 13 channels in and around Boston.

Content

Types of content: Our interviews also show a wide array of content types being saved in the current digital environment. Here are our findings on what content is preserved:











- Text predominates: Text is by far the most commonly preserved content type, cited by 23 of 24 news organizations, followed by photos, cited by 20 of 24, video (17 of 24) and graphics (12 of 24). Source: Interview responses to part of Question 1; see Appendix B for full text of questions.
- Final, not raw content: Most outlets preserve the final published or broadcast content (17 of 24), and only a few saved everything, including raw content materials (5 of 24). Source: Interview responses to part of Question 11.
- Selections mostly implicit: Most of the 24 sites have only implicit selection policies for preservation, totaling 16, while 8 reported clearer or written explicit policies. Source: Interview responses to part of Question 11; see Appendix B for full text of questions.

Meaning, impact, observations: One issue that stood out across the board was the enormous challenge newsrooms face with the exploding number of digital publishing channels, the new or modified content types they require and how best to preserve them, if at all. As newsrooms added channels at a seemingly breakneck speed in recent years, re-allocating or in some cases adding resources to launch podcasts or newsletters or expand focus on social media and audience development functions, we heard almost no cases where resources for preserving these content channels were also taken into account.

Many sites we spoke with acknowledged that they are simply overwhelmed with the multiplicity of content channels and have not even begun to preserve content for these emerging digital channels, or much of anything beyond their immediate primary content types.

Their challenge is illustrated in the following table, which shows that text, photos and video are by far the content types most often saved by the newsrooms we interviewed. Few news organizations are taking any deliberate steps to save content types such as podcasts, e-newsletters, database content or news reports delivered via smart speakers and other IoT (Internet of Things) devices. And only one news organization we talked with is preserving replicas of their web pages in a dedicated preservation system, beyond the content all sites store in their web CMS publishing systems.

Figure 3: What types of news content are preserved?




















	SITES THAT KEEP EACH TYPE	
Text	23	
Photos	20	
Video	17	
Graphics	12	
E-edition (incl. PDFs)	11	
Audio	10	
Scripts/logs	7	
Podcasts	3	
Data content (tables, etc.)	2	
Web	1	

Source: Interview responses to part of Question 1; see Appendix B for full text of questions.

Content

To better understand the source of the content, we reviewed the publishing or broadcasting channels that each organization utilizes. Here's what we found are the primary and secondary publishing or broadcast channels used by the media companies we interviewed.

Figure 4: What are your primary and secondary channels for news distribution?

	IDENTIFIED AS PRIMARY CHANNEL	IDENTIFIED AS SECONDARY CHANNEL
Broadcast radio	6 	1 
Broadcast TV	6 	1 
Digital (web, native apps)	23 	10 
Newsletters	0	6 
No response	1 	3 
Podcasts	5 	11 
Print	11 	1 
Smart devices (Alexa, etc.)	1 	8 
Social media	5 	18 
Wire services/APIs	4 	1 

Source: Interview responses to part of Questions 8 and 9; see Appendix B for full text of questions.

As the table shows, the digital channels of websites and native apps are the dominant ones among these news organizations. Nearly all publish a news website, the single most important channel for many regardless of the size of the news organization, and regardless of whether their readers consume news through a desktop browser, mobile web browser or a native app on any device.

The example of the Chicago Tribune, for example, illustrates these issues. The Tribune publishes photos to a highly popular Vintage Chicago channel on Instagram, the social media service preferred by many newsrooms for photos because its visual orientation. But like all other news organizations we interviewed, none of their social media content is preserved outside of the social platform on which it originated.

"I run a vintage photo Instagram account," said one photo editor. "I do this every day. Have we thought about how that is preserved? What will happen with those images? Not many people are thinking about these things."

"With so many platforms, have we thought about what it means to be published now?" she asks.

"We can really only say what was published in the paper, because it's fixed. We can't really track what has been published online, including when, with what images, maps, etc., things like that. ...

So it's kind of ridiculous that we don't know this and can't track what went online."

While the challenge of managing and preserving content is more complex in the digital publishing era, technical issues have impacted news content increasingly for decades, ever since computer technology first began making inroads into news operations with video display text terminals and Macs for digital photography and graphics.

Content

These challenges are reflected in the wide variety of content and metadata gaps among all of the news archive collections we surveyed. Here are some examples of these variations in the newsrooms we talked with:

- At one large metro newspaper, photo editors stressed the need for staff photographers to follow protocols for preserving the unused digital frames from each news story, the “outtakes” as they are called, along with critical metadata. But like newsrooms everywhere, they acknowledge it’s a daily struggle with this massive set of content, five-to-ten-times as many images as the ones that actually get published, often much more. It’s like working against a flood. The results, estimated the news librarian: 80% of outtake images from past decades are not usable because the metadata is so minimal.
- When another large metro paper moved to the new Chorus publishing platform a few years ago, they brought in more than 7,000 videos, but the data goes back only to May 2019.
- At one NPR station, content published on their website is missing prior to 2006, when they moved to the NPR-based Core Publisher system.
- At one broadcast network, they found significant gaps in their website content prior to 2007, primarily in metadata. They’ve worked to fill in content using digitized versions of content that was originally analog, including printed transcripts that go back to 1990.
- One media company is facing a challenge that stems from acquisition of other media outlets, each with its own technology and workflows. Aligning these content sources into one DAM platform is a huge challenge, said one manager. The primary goal: do it right with new content going forward. “Looking backward, there are a lot of gaps. This is a stake in the ground moving forward.”

Finding 3

Public media have better resources, better archives

As part of this project we conducted a limited analysis of news content now being preserved by the news organizations we interviewed. To do this we used the Simple Property Oriented Threat (SPOT) model, one of a number of standard assessment tools used in the preservation field. See *Methodology* for more on this.

When viewed through this lens of the SPOT Model’s six data risk properties, one pattern becomes clear, if unsurprising: news organizations with more resources dedicated to preservation do a better job than others at preserving their news content. This also correlates with other findings that show a major difference between public and private sector news organizations in staffing resources involved in preservation work.

Our analysis showed that public or nonprofit news organizations have significantly higher SPOT ratings on the quality of saved news content than the private, for-profit sector of news organizations, with public outlets scoring more than 22% higher on the SPOT scale than private news outlets, overall. See table on next page.

Content

Figure 5: SPOT model analysis of saved content for 24 news organizations

	AVERAGE MEDIAN SCORE
Public news media	2.48
Private news media	2.04
Difference	0.45
Percent difference	21.9%

Source: Scores are averages of median ratings by project team members based on information about content preserved at each news organization.

This difference in saved news quality is consistent across each of the six SPOT properties, with a measurable difference between public and private news outlets on each property. Public outlets scored:

- 14% higher on content *Availability*, measuring whether or not content has been saved.
- 32% higher on *Identity*, measuring how clearly saved content can be uniquely identified.
- 16% higher on *Persistence*, measuring reliability of the data's storage systems.
- 27% higher on *Renderability*, measuring the ability to reproduce the data as originally published.
- 16% higher on *Understandability*, measuring the ability to comprehend the content as originally intended.
- 28% higher on *Authenticity*, measuring the ability to ascertain that the news content is what it purports to be, that it's faithful to and unchanged from the original.

Meaning, impact, observations: Digging deeper into these results it's important to note that some of the largest quality differences between public and private groups appear under the more advanced, more difficult-to-achieve properties, such as these (see table also):

- **Renderability**, with a 27% difference in the ability to reproduce the original news content. This is directly related to factors detailed in other parts of this report which scored higher in our SPOT analysis, including whether or not the news organization saves all content elements rather than just one or two, such as the text only, but not images or video or audio. It's also related to whether or not there are clear, usable linkages between these content types, to enable the reassembly of news content into a package similar to its original presentation.
- **Authenticity**, with a 28% difference in the ability to ensure the content is faithful to the original, unchanged. As shown in this report, the difference is related largely to the quality of the metadata saved with the content. For example, is the metadata sufficient to prove the content was created by staff in your news organization rather than a news service or freelancer? Is it sufficient to show that you have rights to republish the content, or license it to others?

These SPOT differences between public and private also correlate at least in part with greater staff resources, as measured by the number of news library or archive staff for each group. Our analysis shows a clear difference on this factor between public news outlets, with an average of 5.3 staff members each, compared to 1.3 in the for-profit sector. The difference is remarkable, with public outlets showing a 4-to-1 advantage in news preservation staffing over for-profit news outlets in our analysis.

Content

Figure 6: How news organizations rated on preservation quality using a 5-point scale (0-4)

SPOT PROPERTIES	PRIVATE NEWS ORGS MEDIAN RATING	PUBLIC, NON-PROFIT NEWS ORGS MEDIAN RATING	DIFFERENCE	PERCENT DIFFERENCE
Availability	2.3	2.6	0.3	14%
Identity	2.0	2.6	0.6	32%
Persistence	2.2	2.5	0.4	16%
Renderability	2.0	2.5	0.5	27%
Understandability	2.0	2.3	0.3	16%
Authenticity	1.8	2.3	0.5	28%
Overall rating	2.0	2.5	0.4	22%

Figure 7: How news organizations compared in resources for preservation

NEWS LIBRARY STAFF FOR EACH SUBGROUP	PRIVATE NEWS ORGS	PUBLIC, NON-PROFIT NEWS ORGS
Average number of news library staff	1.3	5.3

Source: Scores in Figure 6 are average median ratings by project team members based on information about content preserved at each news organization. For SPOT definitions used in this project see Appendix C. For background on SPOT model:

<http://www.dlib.org/dlib/september12/vermaaten/09vermaaten.html>

Staff counts in Figure 7 are from interview responses to Question 4; see Appendix B for full text of questions.

Underlying process, policy and technology factors influence quality of saved news

A deeper look into the analysis shows some of the underlying factors that shaped the pattern of ratings from one SPOT property to another. Here are some observations on these factors:

Most news organizations interviewed did better on the primary SPOT property, **Availability** of content, than on other more granular factors that depend on metadata and other features of a news organization's preservation process and technology. While Availability is essential, it is the bare minimum needed for good preservation, only the starting point. Knowing that content has been saved is critical, but that alone will not ensure that content can be reproduced faithfully for its intended purpose. As noted elsewhere in this report, the high degree to which at least basic news content elements are being saved was unexpected.

On the second property, the **Identity** of content, this factor applies to the precision and consistency with which a content object is distinguished from other objects in a system or repository, determined by features of the technologies in use, such as the types of unique identifiers assigned by a system, file names or identifiers such as alpha-numeric unique ID numbers or text strings that are stored as part of the metadata in a database, along with query or search systems used and their influence on the ability to identify content objects. Most of the news organizations we interviewed rated fairly high on this property. However, we generally gave higher ratings to outlets or systems that used multiple metadata techniques to identify content compared to those that relied on one, such as a filename string in a file system that contains a date and sequence number.

Under the **Persistence** factor for data storage, one trend we noted is that news organizations are increasingly relying on off-site, vendor-provided cloud data storage systems such as Amazon's AWS, by far the most common service reported by news outlets we interviewed. In many cases AWS and other cloud services have largely or fully replaced on-premises data storage systems, the racks of SAN

Content

storage devices that were ubiquitous at news media data centers and are still in use by many outlets. It's important to note, however, that most of those we interviewed were unaware of whether, or to what degree, AWS meets high-level digital preservation storage standards such as ISO 16363.

Only a handful of interviewees reported using one of the main AWS redundancy tools, the Cross-Region Replication feature, which duplicates data in separate S3 data storage "buckets" that are located in a different part of the U.S., or abroad. Several sites also cited the use of AWS Glacier, a lower-cost storage service that is slower for retrieving data, as a secondary storage service, as a form of backup. Only one outlet, a large broadcaster, reported that they do not fully trust any one service, especially off-site commercial cloud services, and make local LTO tape copies of their digital content to be sure.

Under the **Renderability** factor, most news organizations have no way to ensure the ability to re-render web pages in the future as originally published. In fact, most change their web presentation structures and underlying code so often that it's not uncommon for parts of news content to show gaps or missing elements within just a few weeks.

Only one, a large broadcasting organization, has its own process for preserving web pages, a function they outsource to an external company. Several outlets we interviewed said they have used, and relied upon, the Internet Archive and its periodic preservation snapshots of some web page content.

Looking at the **Understandability** property, one of the key factors in this part of the ratings is whether or not systems generate and retain strong, reliable structures of linkages between related content elements of a story. Most reported that linkages are limited, and leave out one or more content types (video, graphics or social media embeds for example). We heard of many such cases, where key content elements become disconnected after publication or broadcasting, especially video, audio or graphics content.

In addition, we learned from sites that do have strong linkage systems that most of these apply only to content created within the current CMS since its installation date; few are able to extend this to previous content packages created under a different technology.

Under the **Authenticity** property, most news organizations we interviewed did not score well in this area. The main reason is that most of their publishing systems, as well as asset management systems and preservation platforms are unable to consistently track changes in content made after initial publication.

Content

Figure 8: How news organizations rated on specific SPOT properties

NEWS PRODUCER TYPE	LEGACY/DIGITAL NATIVE	GEOGRAPHIC COVERAGE	MEDIAN OF PARTICIPANT RANK VALUES FOR EACH SPOT PROPERTY						AVERAGE OF MEDIANS
			AVAILABILITY MEDIAN	IDENTITY MEDIAN	PERSISTENCE MEDIAN	RENDERABILITY MEDIAN	UNDERSTANDABILITY MEDIAN	AUTHENTICITY MEDIAN	
web/print	legacy	international	4	4	1	4	4	4	3.5
web/print	legacy	regional	1	1	1	2	2	1	1.3
broadcast/ web	legacy	international	4	4	4	4	4	4	4.0
web/print	legacy	regional	2	1	2	2	2	2	1.8
web/print	legacy	regional	2	2	2	2	2	2	2.0
web/print	legacy	regional	3	2	3	3	3	3	2.7
broadcast/ web	legacy	international	3	3	1	3	3	3	2.6
web/print	legacy	regional	2	3	3	2	2	2	2.3
web/print	legacy	regional	3	3	3	2	3	2	2.5
broadcast/ web	legacy	regional	2	2	2	1	1	1	1.4
broadcast/ web	legacy	regional	1	1	1	1	.5	1	0.9
broadcast/ web	legacy	regional	2	2	3	2	2	2	2.0
web/print	legacy	regional	3	3	3	2	3	3	2.8
web/print	legacy	regional	3	3	4	3	3	2	2.9
broadcast/ web	digital native	national	2	1	2	1	2	1	1.4
broadcast/ web	legacy	national	4	4	4	4	4	4	4.0
broadcast/ web	legacy	national	3	3	3	3	2	3	2.8
web	digital native	regional	2	2	1	1	2	1	1.4
broadcast/ web	legacy	regional	2	2	2	1	1	1	1.5
web/print	legacy	regional	3	2	3	3	3	2	2.6
web/print	legacy	international	2	1	1	2	2	1	1.5
web/print	legacy	regional	1	1	2	1	1	1	1.1
web	digital native	national	1	1	2	1	1	1	1.2
broadcast/ web	legacy	regional	4	4	4	4	3	4	3.8
Average of medians per property			2.46	2.25	2.35	2.25	2.10	2.04	2.24

Source: Scores are average median ratings by project team members based on information about content preserved at each news organization. For SPOT definitions used in this project see Appendix C. For background on SPOT model reference: <http://www.dlib.org/dlib/september12/vermaaten/09vermaaten.html>

In many cases we encountered there is no process, policy or technology in place to designate a permanent canonical published version of a news story package for preservation. Several sites do have these functions, including workflows and technology steps in dedicated preservation systems to declare a canonical copy and then track changes from that point forward. However, most of these were manual steps done by staff members. Only one, a large broadcasting organization, uses an automated process. In this case the process allows a week for any content errors to be corrected or missing elements to be added before committing the package to the archive as the canonical copy, rigorously tracking after that point.

Content

Quality varies by type and age of news organization, and by geographic market

One additional set of outcomes that emerged from this analysis shows differences between other subgroups such as the geographic size of the market the news organization serves. The larger the market (international, national) the higher the saved content ratings. This fits with other data in this study and with general expectations that outlets that serve wider audiences have greater resources for preservation.

Figure 9: Preserved content ratings by geographic scope

	AVERAGE OF MEDIAN AVERAGES
International	2.9
National	2.3
Regional	2.1

Source: Scores are average median ratings by project team members based on information about content preserved at each news organization. For SPOT definitions used in this project see Appendix C. For background on SPOT model reference: <http://www.dlib.org/dlib/september12/vermaaten/09vermaaten.html>

Two other outcomes, however, were unexpected. One was a difference between news organizations based on the primary channels used to distribute news. Our analysis showed that broadcasters who also operate news websites had the highest average ratings. Newspapers that now also operate active news websites had the second highest scores. And, most surprisingly to our research team at least, those news outlets that publish only on the web scored much lower. (See table below.)

Figure 10: Preserved content ratings by primary channel

	AVERAGE OF MEDIAN AVERAGES
Broadcast/web	2.44
Web/print	2.24
Web	1.29

Source: Scores are average median ratings by project team members based on information about content preserved at each news organization. For SPOT definitions used in this project see Appendix C. For background on SPOT model reference: <http://www.dlib.org/dlib/september12/vermaaten/09vermaaten.html>

When results are viewed from a slightly different angle, it still tends to confirm this outcome. We grouped so-called “legacy” news media organizations together, combining broadcasters with newspapers. We compared these to newer organizations that utilize only more modern channels such as web pages and OTT digital streaming services, the so-called “digital natives.” The results we see are similar, that scores for digital native news organizations are far lower than those for legacy organizations. This was unexpected as we assumed native news organizations have inherently better knowledge of technology and ways to ensure content is retained.

Our analysis showed otherwise. This marked difference appears to have some relationship to the long-established nature of pre-digital preservation processes, workflows and systems. There appears to be a bleed-over effect on digital news preservation, especially for news organizations with significant preservation staffing already in place and continuing through the transition to digital publishing.

We heard from many news organizations across the spectrum that a general lack of standards and clarity on best practices for digital news preservation makes it difficult to know what should be done. Combine these factors and it’s less surprising that digital native news organizations are not doing as well on content preservation as their legacy cousins.

Content

Figure 11: Preserved content ratings by origin

	AVERAGE OF MEDIAN AVERAGES
Legacy	2.37
Digital native	1.33
Difference	1.04
Percent difference	77.98%

Source: Scores are average median ratings by project team members based on information about content preserved at each news organization. For SPOT definitions used in this project see Appendix C. For background on SPOT model reference: <http://www.dlib.org/dlib/september12/vermaaten/09vermaaten.html>

Finding 4

Internal use is primary, public access important but often outsourced

One of the goals of this research project was to understand who the expected users or audience for the content that is being preserved are, whether internal to the news organization or outside. We also sought to understand current practices around access to preserved content: who has access and how news organizations control and manage access.

In most cases newsrooms told us their primary purpose for saving content was for internal use. This was the key factor driving existing efforts.

Of the 24 news organizations we interviewed, 23 told us their primary purpose was for use by news staff members who need previous content for reference on upcoming news stories, to rerun past content in whole or part, and for use by others within the company.

Figure 12: For what purposes is news content preserved?

	NUMBER OF NEWSROOMS CITING EACH PURPOSE
Internal use	23
Public use	14
Licensing	11

Source: Interview responses to part of Question 1; see Appendix B for full text of questions.

The next most often cited purpose are the 12 news organizations who also see the public as important users, through libraries or other third-party archive services; 8 who see third-party publishers as important users; and 7 who cited researchers as key users, such as those at universities or non-profit organizations.

We also found that many are leveraging the potential to license content to other publishers or broadcasters. In some cases, such as with CNN and The Associated Press, this was acknowledged to be a significant revenue source, although none were willing or able to share revenue data.

Figure 13: Who are the primary users of saved news content?

	NUMBER CITING THESE USERS
Internal journalists/newsroom	23
Public	12
Third-party publishers	8
External researchers	7
Other	2

Source: Interview responses to Question 3; see Appendix B for full text of questions.

Content

Meaning, impact, observations: Our findings show that, while internal newsroom research and reuse of past content is the main purpose for content preservation, public access is also offered by nearly all of the news organizations we interviewed.













For the most part they do not use the same tools or systems for both internal staff access and external public access. With few exceptions, internal systems are accessible only to news and other company personnel. Separate systems or third-party services are provided to handle access by the public. Most of the newsrooms interviewed said they have outsourced this function to third-party news reseller services, along with institutions such as libraries or universities, as the primary means by which members of the public can access past news content.

Here is what we learned on access issues:

- Most of the newsrooms interviewed said they provide public access to past content by feeding their news to third-party news resellers or syndication services (20 of 24). This includes access to such services at no additional cost for subscribers, a model some news organizations use as an incentive for paid subscriptions. Source: Interview responses to Question 22; see Appendix B for full text of questions.
- Of the news syndication services or resellers cited, the most commonly used was Newspapers.com followed closely by ProQuest. Other common service providers include Lexis-Nexis, NewsBank and Factiva. Most news organizations have agreements for content resales with two or more of these services. Source: Interview responses to Question 22.
- Public access is also commonly available only through institutions such as libraries or universities, along with high school libraries (16 of 24). Source: Interview responses to Question 22.
- In addition to public access through the methods above, most of the news organizations we interviewed make content available to other businesses, institutions and researchers through content licensing operations handled by licensing agencies such as PARS International.
- Larger news organizations have internal licensing teams, such as AP, CNN and the BBC, which run significant licensing operations in part because their content has broad, international interest.
- News publishing systems in general are accessible internally to all news staff and to other key company personnel, but most have limits or controls on this access (16 of 24), through group or role-based permission arrangements; five news organizations provide unlimited internal company access. Source: Interview responses to Question 21.




Content

Figure 14: Outside of publishing channels, how do external parties access your content?

	NUMBER CITING EACH	
All news syndication services	20	
Library/institution	16	
Licensing	9	
Other syndication/news services	6	
By appointment/request	5	
Newspapers.com	5	
ProQuest	4	
Direct access to system (DAM, other)	3	
No response	3	
Lexis Nexis	2	
NewsBank	2	
Factiva	1	

Source: Interview responses to Question 22; multiple responses counted. See Appendix B for full text of questions.

Figure 15: Are there any internal limits on who has access to preserved content inside your organization?

	NUMBER CITING EACH	
Yes, limited or controlled	16	
Yes, unlimited	5	
Response unclear	3	

Source: Interview responses to Question 21; see Appendix B for full text of questions.

News services are mostly text-only, but more comprehensive than print

It should be noted that many of the external systems through which the public has access to news content provide only text or page-image data. Some do offer access to searchable renderings or PDF-style displays of print newspapers, where photos and maps can be viewed, for example. But none currently provide access to digital news content in forms similar to the way it was originally published online. In most cases, presentations such as original web pages or mobile app presentations are not saved at all. Only one organization of the 24 we interviewed, the BBC, had a mechanism to capture web browser news content presentations.

One other factor that's important to note: In most cases, the news syndication services are receiving content feeds that are generated from the company's web CMS. This means the feeds usually include a more complete set of all their own published content than what is commonly available in print.

Print and online content sets differ for multiple reasons. One key factor is the limited space available in newspapers, but there are others as well. Time-related factors tend to discourage print publication of news stories that are newsworthy only for short periods as they occur, such as coverage of temporary events such as weather issues or traffic jams. Content like this may be published and updated frequently on the web and in other digital channels as breaking local news, but is unlikely to appear in print, especially as the decline in advertising shrinks available newsprint space.

Content

These differences can be a major challenge for preservation, especially for news organizations that have not completed the shift to fully digital-first publishing workflows, for reasons of technology limits or internal newsroom culture. Several of the news organizations we met still have print-oriented systems or workflow steps ahead of digital publishing, which can slow the process and, in some cases, require manual cut-and-paste workarounds to get content online. For several of the newsrooms, this leads to separate sets of print and digital content that are not always kept synchronized.

Institutions, government fill some gaps

In addition to providing access to digital content, many of the news organizations we met also provide access to older analog content going back many decades, often under the care of library and university institutions. For example, the Baltimore Afro-American has an agreement with nearby Morgan State University, where they preserve physical newspapers, photos and other historical materials dating back almost to the founding of the newspaper in 1893. The Boston Globe has a similar arrangement with Northeastern University. Others, including the Chicago Sun-Times have similar arrangements.



A portion of the newsroom at The Boston Globe during the research team's visit in March 2020, just before offices were closed due to the COVID-19 pandemic.

In the U.S., some digital news content as well as significant print collections are available through the Library of Congress (LOC), the only American government institution whose mission includes preservation of materials such as news content. The LOC for years has been collecting copies of print newspapers and analog copies of news program tapes through its mandatory deposit laws and copyright laws. But these functions are limited to designated titles only, rather than all publications. For example, of McClatchy's 30 daily newspapers, the LOC required mandatory deposit for about half of them.

The Library of Congress has also faced significant funding or budget constraints that have also severely limited their ability to preserve digital news content. The LOC is currently operating a limited web page capture and preservation effort for current digital news in cooperation with the Internet Archive. This covers only limited, manually-selected sets of news content chosen by LOC curators according to judgements of the most important news or cultural activities that change periodically. In addition, access to this web content is embargoed for one year and has other limitations, to protect news company copyrights.

Content

Overseas there are similar institutions run by governments that have broader missions and funding. One that we interviewed is the Institute for Sound and Vision in the Netherlands, near Amsterdam. This is a collaborative organization of the Dutch government that works with private and public television broadcasters to capture and preserve much of the news and entertainment content generated in that country.

Another is run by the BBC, which has one of the largest and oldest collections of news content anywhere. For the BBC, however, the method for providing public access is not directly through their systems or operations, but through another government agency. For UK citizens who want access to the huge BBC Archives, they go instead to either The British Library, where onsite and some online access is provided; or to the British Film Institute for television content. In addition, residents of Scotland can access Scottish content through The National Library of Scotland; similarly through the National Library of Wales for Welsh citizens.

Technology Findings

Technology is a major factor in whether, and how well born-digital content is preserved

What we found is that design and capabilities of publishing systems are critical, along with the way systems are installed and how they are used. Key factors include the type and breadth of metadata tools available and how they are configured, the kinds of workflows a CMS can support and how these functions are used to implement policy choices, the way systems are installed and user training is conducted, and how all of these factors come together to support overall company culture and values.

The details around these factors, we found, can have considerable influence on the way content is preserved, the extent of metadata saved with content, the degree of difficulty and potential manual effort involved in news preservation, and more.

In this section we will cover the basic findings on what technologies are in use in the news organizations we interviewed. Following this we'll review several of these critical technology factors and their impact on preservation. Lastly, we'll share what we learned in talking with key technology providers to the industry, where new developments have potential to help significantly improve news content preservation.

Finding 5

Top tech challenge is managing multiple digital channels

Keeping up with the changing and increasingly competitive landscape for content delivery is an enormous challenge for today's news media companies. While most of the news media companies in the U.S. have successfully made the switch to digital publishing systems and workflows over the past decade, our interviews showed that the continuous evolution of that technology is one of the major challenges not just for news content preservation but for daily publishing as well.

There's a seemingly insatiable need for tech resources and services to feed new sites and products, mobile devices and streaming channels, social media platforms and more to generate new revenue, consuming nearly all available resources. In this high-pressure environment, long-term content preservation is often ignored or too low on the list of priorities, its needs unmet, the interviews showed.

While some of these challenges are to be found in almost any modern business or institution, these factors affect the news industry in particularly challenging ways as it struggles to evolve toward a new business model while the existing one dissolves seemingly beneath their feet.

It's no wonder, then, that the cost and pace of changing technology, drawing resources away from preservation work, tops the list of key challenges that news organizations told us they are facing today. These technology issues, along with related factors in workflows and policies, are the substance of daily struggles.

Technology

Here are the results of discussion on these questions. When asked to identify the biggest challenges they face in digital news content preservation, the newsrooms reported the following:

- Dealing with demands for new technologies and support systems to handle the growing and dizzying array of multimedia channels and formats (13 of 24). Source: Interview responses to Question 5; see Appendix B for full text of questions.
- Avoiding technology obsolescence, by effectively completing complex migrations of news and other workflows from one tech platform the next (9 of 24). Source: Interview responses to Question 5.
- Ensuring the quality of content data, the effectiveness of search systems and the ability of news staff to find, or discover, what they need in publishing systems (9 of 24). Source: Interview responses to Question 5.
- Ensuring production systems are in compliance with backup and disaster-recovery policies, changing content ownership and licensing requirements, version and error tracking demands, increasingly complex security issues to ward off hackers, and more (8 of 24). Source: Interview responses to Question 5.
- Other factors as well, such as the problem of information silos, dramatic expansion of data volume, the focus on production, decline in staffing and funding and others shown in the table below. Source: Interview responses to Question 5

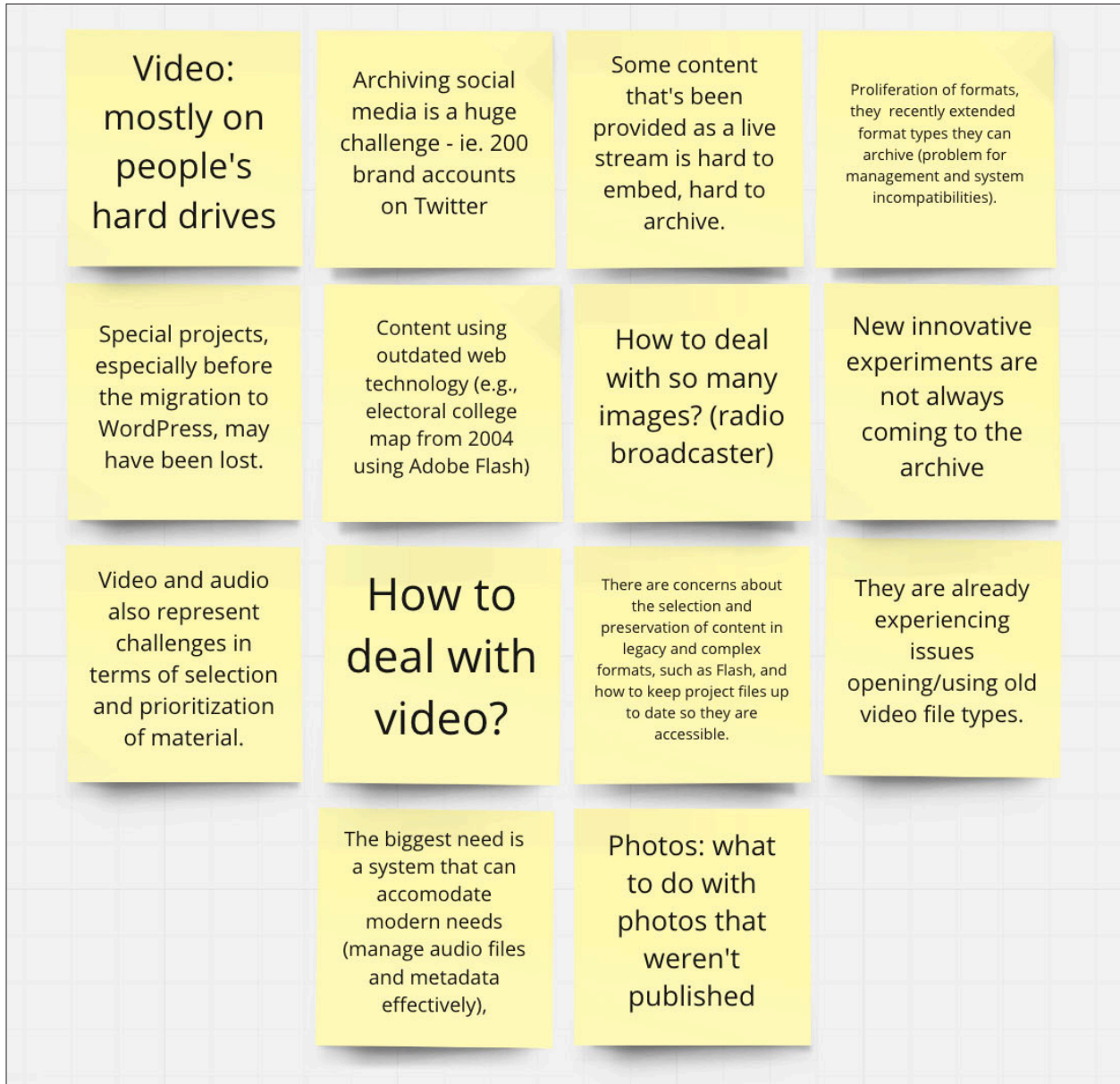
Figure 16: What are the key challenges the organization faces in archiving and preserving news content?

	NUMBER OF ORGANIZATIONS CITING EACH
Handling multimedia formats, video, audio, photos, web (variety and complexity)	13
Technology migration (switching systems)	9
Data quality, search/discovery (difficulty finding content)	9
Production compliance with rigorous policies in production technology	8
Information silos (between systems and staff groups)	7
Data volume (growing file sizes of digital media, cost of high-end storage)	7
Production is the priority, not preservation (competition for resources)	6
No preservation plan (no policies, not part of mission)	6
Staff: numbers/time (insufficient or no preservation staff)	5
Risk of reliance on vendors (worry about proprietary systems if vendor changes focus)	5
Legacy media and structures (problems in adapting and blending analog content into modern digital systems)	5
Managing content relationships/versioning (maintaining linkages for continuously updating/changing news assets)	4
Workflows (existing archive processes that don't account for digital content fully, or at all)	3
Staff: knowledge/skill (need for training in digital processes for new and existing staff)	3
Rights/legal (difficulty determining rights for reuse, licensing on digital assets)	2
Funding (declining budgets, including COVID-related cuts)	2
Lack of technology (need for DAM instead of production-only systems)	1

Source: Interview responses to Question 5; multiple responses counted. See Appendix B for full text of questions.

Technology

Handling multimedia formats: video, audio, photos, web



Examples of the statements about key challenges for preservation made by newsroom and technology managers during our interviews. Some are quotes, some paraphrased for space. Comments were grouped into summary statements using the Miro whiteboard tool.

Source: Interview responses to Question 5.

Technology

Meaning, impact, observations: As detailed above, news organizations are constantly struggling with new technology demands alongside competing issues of cost constraints. Here are the key factors driving technology changes cited by the news organizations we interviewed:

- Financial issues top this list. The constant need for lower technology costs even amid expanding functions and services, was cited by 13 of the 24 news outlets as one of the key drivers or new technologies. Source: Interview responses to part of Question 30; see Appendix B for full text of questions.
- The need for greater efficiency and streamlined workflows was cited by 9 of 24, as part of the constant effort to cut costs as the industry struggles with declining revenue. Source: Interview responses to part of Question 30.
- Changing business or organizational requirements were cited by six sites, while related changes in overall mission were cited by four, along with other factors. Source: Interview responses to part of Question 30.








One of the most difficult challenges of all in technology is ensuring that your systems are able to survive potential disruptions, ranging from weather phenomenon such as hurricanes or tornadoes to more common electrical outages or technology events such as disk storage failure, database corruption or security intrusion attacks. Source: Interview responses to part of Question 30.

To deal with these potential threats, most large news organizations maintain full-scale backup systems, tested in detailed, time-consuming disaster plans and exercises. But in the face of today's financial challenges, it was not surprising that less than half of the news organization we interviewed had full, written disaster plans (9 of 24). Another nine news organizations reported having informal or unwritten plans, or partial or pending plans. Source: Interview responses to part of Question 30.

To complete this picture, we asked what technology changes, if any, are under way or planned that could have an impact on digital news content preservation. Here's what the news outlets reported:

- Almost half reported they are conducting or planning new technology systems for production, editing or publication functions (11 of 24). Source: Interview responses to part of Question 30.
- This includes moving to an entirely new Content Management System (5 of 24). Source: Interview responses to part of Question 30.
- Meanwhile, a handful of others (3 of 24) said they were hoping to improve or update data storage or hosting systems soon). Source: Interview responses to part of Question 30.






Figure 17: What are key drivers of technology or system changes?

	NUMBER CITING EACH	
System costs/financial	13	
Demand for streamlined workflows/efficiencies	9	
Broader organizational needs/acquisitions	6	
Mission	4	
Poor performance/obsolescence of existing tech	3	
New features/improved usability	3	
Not answered	1	

Source: Interview responses to part of Question 30; see Appendix B for full text of questions.







Technology

Figure 18: Are there any planned technology changes that will impact preservation?

	NUMBER CITING EACH	
Yes: new production/editing/publication systems	11	
Yes: moving to new web CMS	5	
Not answered	3	
No	3	
Hoping to update storage/hosting solutions	3	
Total (KWMU has two entries)	25	

Source: Interview responses to part of Question 30; see Appendix B for full text of questions.

Figure 19: Does the organization have a disaster or business continuity plan?

	NUMBER CITING EACH	
Yes: written/explicit	9	
Yes: unwritten/inexplicit	6	
Partial/pending	3	
Not sure	2	
No response	1	
No	3	
Total	24	

Source: Interview responses to Question 31; see Appendix B for full text of questions.

Meaning, impact, observations: The growing multiplicity of content channels is an issue for all media organizations, not only legacy media such as newspapers. It also affects broadcasters, both television and radio stations, and newer digital media organizations as well.

In broadcasting, in particular, we saw the increasing importance of news websites and related digital channels for TV stations and especially for radio stations that once relied exclusively on over-the-air signals. Now, radio stations are posting more and more content to their websites, and doing it more frequently. Reader traffic has responded.

One NPR radio station we interviewed, KWMU in St. Louis, acknowledged that news consumption on its website has grown as former listeners switch to phones and other mobile devices to get their news. This growth has been so rapid in recent years that it is now competitive with its radio broadcast audience, and in one recent month exceeded the number of listeners over the air.

Additionally, KWMU recently modified its workflow sequence to prioritize their website. For all but the most urgent breaking news stories, reporters now file first to the web, and then generate audio for radio broadcast.

This kind of evolution poses a huge challenge for media organizations, which struggle to adapt their technology and workflows to handle such changes. One response: NPR's central office in Washington, DC recently developed a new nationwide web-based content management system, called Grove, that it's now offering to all local NPR stations. The system will facilitate sharing of content among any or all NPR affiliates, and will be backed by a central tech team to better ensure content is preserved.

Technology

Surprisingly, we found that these and other preservation issues also affect digital native news outlets, a small but rapidly growing segment of the news industry. While digital startups may be cutting-edge when they first launch, they face the same technology rat-race that others do as the pace of technology change continues to accelerate over time, requiring continuous adaptations that can be as hard on them as any media outlet.

For this project we interviewed two such digital natives, Vox Media, the large and prominent media organization based in New York; and QCity Metro, a small news startup that serves the Black community in Charlotte, NC. Both shared examples of the same kinds of challenges others reported, including content that gets lost in the transition between one tech platform and another, and other issues with ensuring that all content is preserved and available long term.

This also showed up in our content analysis. While we expected digital native news outlets to be better at preserving digital news content, in our analysis they actually scored lower than legacy news outlets on an analysis of saved news content using the SPOT preservation model. This may be related to staffing as much as anything else.

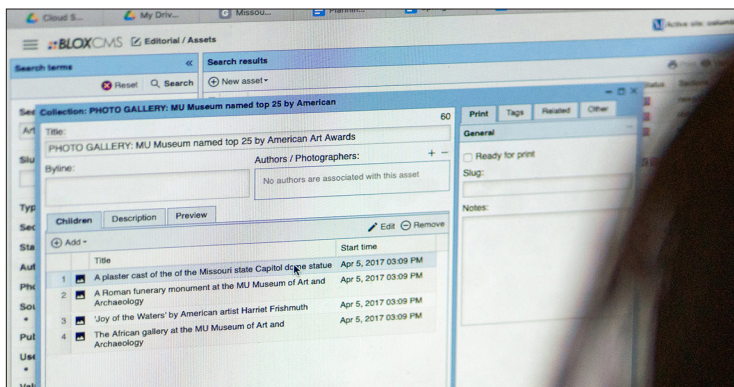
Finding 6

Web CMS is central, often doubles as archive

Our interviews showed that modern web Content Management Systems have become the single most important platforms for news publishing. This is driven by the push toward digital publishing as the dominant way of reaching readers and customers. This is the case for most news media organizations, including those originating in print and many of those originating in broadcast media.

In the drive toward fully digital publishing, the web CMS has become the central multifunction platform for digital news. Most newsroom interviewees told us they use their web CMS to generate much more than web content, which was the CMS's original intent. Today the web CMS is also used to feed a wide array of content channels, including native digital and mobile apps, newsletters, third-party news services such as Apple News, social media posts and links, smart speakers, IoT devices, print publishing systems and more.

This section also covers findings related to technologies beyond the web CMS, including systems used in the publication or broadcasting process, the systems that make content available publicly to consumers, and systems used for asset management and archiving.



A photo editor at the Columbia Missourian sets up a photo gallery using the BLOX CMS from TownNews.

Technology

There are also relevant factors in systems used to produce content in print, for newspapers, magazines, e-publications and others. In many cases, for example, versions of web and print content are shuffled back and forth between parts of CMSs or different publishing systems used for specific channels. Although print distribution is declining steeply, for example, we found several newspapers where print workflows remain the origin for news stories online, including some workflows that unintentionally contribute to delays in digital publishing and can also result in preservation losses if content is not synched between web and print or other systems.

Here's what we learned on the use of content management systems of all types, including systems for asset management and preservation:

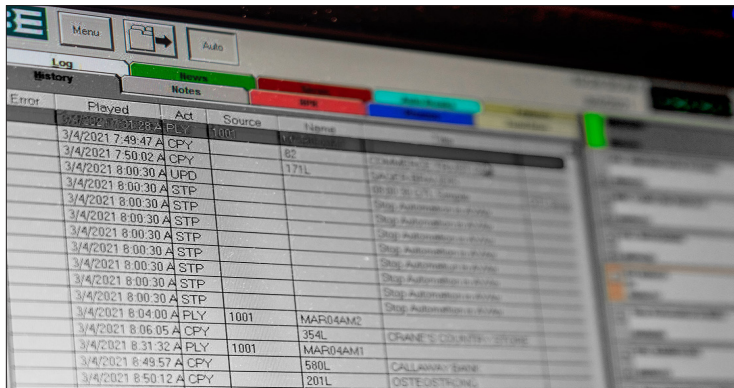
- Of the 24 news organizations we interviewed, all use a web content management system of some kind at or near the heart of their news production and publishing operation. Web CMS platforms are the dominant technology system in today's newsroom, in many cases the only major tech platform. Source: Interview responses to parts of Questions 15 and 16; see Appendix B for full text of questions.
- The CMS platforms most commonly used by the news outlets we interviewed are custom (in-house) developed systems. We found 14 different custom-developed systems, used for many different functions, at 10 of the newsrooms we talked with. Source: Interview responses to parts of Questions 15 and 16.
- Following this, the most commonly used commercially available systems, including DAM platforms are: the BLOX web CMS from TownNews and CUE Print (formerly NewsGate from CCI) from Stibo DX with five sites each; and WordPress, BLOX Total CMS from TownNews, MerlinOne, SCC MediaServer, and Arc XP (formerly Arc Publishing), with three sites each. Source: Interview responses to parts of Questions 15 and 16.
- In addition to their web CMS, most of the news organizations we interviewed use one or more additional content or asset management technologies, defined broadly to include any major systems involved in news production that includes an independent database. Six of 24 identified four or more major content management systems in simultaneous and interconnected use. Two of these newsrooms reported as many as five separate major content management technologies and two newsrooms were in transition to replacement platforms. Source: Interview responses to parts of Questions 15 and 16.
- The sheer multiplicity of content systems and other tech platforms in modern newsrooms, and the frequent issues with integrations and workflows between them, itself contributes significantly to the loss or corruption of digital content over time, we learned in a number of cases. We'll go further into this in the Practices section.

Technology

Meaning, impact, observations: In the technology universe at large, the term Content Management System, or CMS, has become almost synonymous with web publishing software. But in the news industry, and in this study, we found the term applied to many software applications used together or separately to manage one or more of the content types used in publishing. The findings in this section count major internal content management systems, not what might be called “minor” technologies, sub-systems or modules, or the host of external systems and services commonly found in newsrooms today serving specific niche functions for digital publishing.

The types of content management systems have multiplied in recent years, and during our interviews, we encountered a wide range. Examples of these are asset management systems (DAM or MAM) that have become common in recent years are those for managing video content, such as Brightcove, an external service; along with video sub-systems that work as modules for larger CMS platforms such as Field59 from TownNews, MediaDesk and Graphene from Brightspot and VideoCenter from Arc XP.

We also learned about broadcast-related content management systems such as AudioVault, used at the NPR stations we interviewed; BitCentral at TV stations, NewsFlex for managing video content streams for broadcast and digital distribution, and high-end video editing and production software from companies such as Avid and Vizrt.



Error	Played	Act	Source	Name
	3/4/2021 7:49:47 A	CPY	1001	52
	3/4/2021 7:50:02 A	CPY		171L
	3/4/2021 8:00:30 A	UPD		
	3/4/2021 8:00:30 A	STP		
	3/4/2021 8:00:30 A	STP		
	3/4/2021 8:00:30 A	STP		
	3/4/2021 8:00:30 A	STP		
	3/4/2021 8:00:30 A	STP		
	3/4/2021 8:00:30 A	STP		
	3/4/2021 8:00:30 A	STP		
	3/4/2021 8:04:00 A	PLY	1001	MAR04AM2
	3/4/2021 8:06:05 A	CPY		354L
	3/4/2021 8:31:32 A	PLY	1001	MAR04AM1
	3/4/2021 8:49:57 A	CPY		580L
	3/4/2021 8:50:12 A	CPY		201L

AudioVault storage and playback software in use in the studio of KBIA radio, an NPR member station operated by the University of Missouri School of Journalism, Columbia, Missouri.

For web publishing functions specifically, the most commonly used among our interview sites were in-house developed systems. This was followed by BLOX, used at five of the sites we interviewed. Also common were WordPress and Arc XP, each found at three sites; CUE Web from the Danish news tech provider Stibo DX, and Chorus, from Vox Media, each at two sites; and single instances of other web CMS platforms such as Ruby Shore, Polopoly and Graphene, the name used at the Los Angeles Times for a system built on the Brightspot commercial platform.

Technology

Figure 20: How many of the news organizations interviewed are using each brand or type of Content Management System or other publishing platform?

	NUMBER IN USE		NUMBER IN USE
Custom developed	14	Preservica	1
CUE Print (NewsGate)	5	Polopoly	1
BLOX	5	n/a	1
WordPress	4	Nikon Snapbridge	1
BLOX Total CMS	3	NewsFlex	1
SCC MediaServer	3	Naviga	1
MerlinOne	3	MirrorWeb	1
Arc XP	3	Methode	1
Vizrt	2	MediaDesk	1
VideoCenter (Arc XP)	2	Graphene	1
CUE Web	2	Field59	1
Chorus	2	Brightcove	1
AudioVault	2	BitCentral	1
YouTube	1	Avid MediaCentral	1
Spring CM	1	Avid Interplay	1
Scisys	1	Avid Access	1
Ruby Shore	1	Avid	1
Primestream	1	Artemis	1
Prestige	1	Archon (ArchivesSpace)	1

Source: Interview responses to parts of Questions 15 and 16; see Appendix B for full text of questions.

It's important to note one recent trend for web CMS platforms that's significant for news preservation: the emergence of the so-called headless CMS. This development, in which the "head," or rendering components of web publishing (how the content looks, often called the "front end") are separated from the rest of the system. This model allows for the user experience to be customized in-house by news organizations. The flexibility of the headless CMS has allowed it to become the standard for how web CMS platforms are implemented at the large news publishing companies.

This means the actual HTML and CSS (visual presentation style sheets) generated for web pages are handled largely by internally developed software systems built by each news organization. These custom systems take structured news content data from the vendor-provided web CMS, ready for rendering, and prepare it for presentation to readers in web browsers, in mobile apps and any other channels in which the news organization controls the appearance.

This combination of vendor-provided core CMS functionality, topped by customer-developed presentation functions has become popular with news publishers seeking to focus internal technology resources on the most important part of the process, the look and feel, organization and navigation of their website, an often unique and critical company branding function. It also means that underlying web CMS technologies are becoming more standard and somewhat commoditized, and often include open-source software components.

Technology

Off-platform tech tools are off the radar for preservation

While the section above covers major internal systems installed and implemented specifically for each newsroom, this only scratches the surface of the full range of technologies in frequent use across modern newsrooms, especially for digital publishing. In addition to the major systems above there are hundreds, even thousands of different technology tools and subsystems in use in today's newsrooms.

These tools and systems are not all covered or tracked in this report but are worth a brief mention because they present a unique set of preservation challenges. What they represent is the increasing fragmentation and dispersal of information and news data that is now a common part of the process of reporting and presenting news content. Of the 24 newsrooms interviewed in depth for this project, only one has any kind of formal process to capture or preserve such material, in this case limited preservation of social media posts by the BBC. None of the other tools and systems are included in preservation processes, if there are any, at the remaining newsrooms, and few have more than limited plans beyond backup copies of data.

To summarize, these range from the innumerable applications and tools found on news staff laptops and mobile devices to the large assortment of hosted software systems tapped as part of their news creation and presentation processes, such as Document Cloud, AirTable, Google Maps and Google Docs, along with individual services such as open-source Apache Tika content analysis software, Timeline.js for news timelines and hundreds more.

For content planning within newsrooms, a surprisingly high number of newsrooms now use Google Docs for their so-called news budgets, the documents used to track daily lists and descriptions of planned news content. This is due to the high degree of flexibility in structuring information, especially compared to the often-limited planning functions found in large-scale publishing platforms. Many newsrooms find Google Docs easy to adapt to continuously changing needs of news planning, news assignments, content sharing among different newsrooms and content tracking.

Tuesday, Nov. 12										
Status	A1 potential?	Display potential?	Scheduled Pub time	Slug or title	Notes to Print	Reporter/Editor	Video	Other assets	Words	Headline or description
Holding				census	holding	strading/schrader				Census will be hiring hundreds of people in the Triangle
Filed	yes			11092019-speakerretirementbill	Good for all NC, expecting to post for 5 am Wednesday, can hold for weekend	kane/schrader				North Carolina legislature hit with \$141,000 bill over retiring aide's pension costs
Filed	no			1112-voting	Good for all NC, will post this during night shift	Carolina Public Press	no	photo from CPP	1600	How voters with disabilities factor into election equipment debate
Planned	no			1112-gerrymandering	Tentative	doran/schrader				
Final posted	yes			1109-causeyprofile	Good for all NC, can hold for weekend	specht/schrader	shot by Casey		2187	Lesson from a pair of scandals: This North Carolina Republican won't be pushed around
Planned				1112-DurhamScooters	Good for both, anytime	schultz/tba			600	Hey Durham, how about those electric scooters? City seeks feedback
Final posted	no	no		1112-markjohnson	Good for all NC	hui/ogburn	file	file	800	Mark Johnson, state school superintendent, plans to run for lieutenant governor. He joins a crowded field.
Initial posted	poss			1112-CashBailLawsuit	Good for all, timely	schultz-doran/tba			1100	Class action lawsuit challenges Alamance cash bail system
Final posted	poss	no		1112-nativeamericanschool	Good for all NC	hui/ogburn	file	file	600	State officials are denying a charter school request because of its Native American education curriculum. They view it as too controversial.
Final posted	no	poss		1112-N&Ocondos	Good for Rai and Dur	sanchez-guerra/ogburn	no	handout drawing	500	The latest on the planned condos that will be at the site of the former News & Observer offices.
Planned										
Filed	maybe	No			Good for all NC	doran/hendrickson	No	mug	475	Wayne Goodwin says he's running, again, for insurance commissioner.
Initial posted										
Holding										

This shows an example of a news content budget using Google Docs, a popular tool used for news planning by many newsrooms today. This budget lists and describes news stories for that day for one publishing channel.

Source: McClatchy NewsDesk

Technology

Couple this with a growing number of communication systems now commonly found in today's newsrooms. Chief among these is Slack, a popular messaging platform that spread to newsrooms in recent years based on its use among digital development teams. There are many others as well, including video conferencing systems such as GoogleMeet, Microsoft Teams and the Zoom system.

For data analysis, reporters have for years tapped into massive government, business, sports, real-estate, crime and other datasets using tools ranging from desktop and hosted spreadsheets and database tools such as AirTable and Excel to web scraping tools and complex applications such as SAS and ESRI for mapping, analysis and visual presentation of news data points embedded with news stories on the web and in apps and social media. Many of the newest digital features tap into highly popular cloud-based tools and solutions for websites offered by Amazon Web Services, including Lambda-based functions for highly flexible data transformations and automated processes.

For audience development, a relatively new function in modern digital newsrooms, a wide array of tools and systems such as ChartBeat, Parse.ly, Google Analytics and numerous internal and external platforms are used to track customer behavior on news websites and in apps, newsletter, podcasts and other channels. And social media also has its own set of hundreds of tools and apps in frequent use, including CrowdTangle, TweetDeck and the dozens of social media services themselves, from Twitter and Facebook to Instagram, What'sApp, Parler and so many more.

The key issue with proliferation of these services and tools is that they are external to newsroom control and usually have no connection other than a URL embed to the systems where digital news content is saved. In cases where there is some form of integration, it's usually one-way, is activated intermittently or sourced only at the time of initial publication. These fragile linkages fall far short of preserving the original content in long-term news preservation systems or saving more than an image representation.

This is especially unfortunate for news preservation in the case of newsrooms using Google Docs for news budgets. This information is highly valuable for descriptive and indexing purposes but lies outside any internal preservation platforms, and usually ends up being deleted or overwritten. As we'll detail later in this report, news budget data is one of the richest sources of news metadata from any source, and in most newsrooms these days it is merely used once and discarded.

Related to the rise of external technology tools is the increasing use of collaborative news reporting efforts that share news investigation resources for content that is published simultaneously across two or more newsrooms. These include local and regional collaboratives, as they are called, created organically by groups of newsrooms or with the aid of organizations such as the Center for Cooperative Media at Montclair State University, which supports news collaboratives across the country.³¹

In some cases these collaborative efforts involve many dozens or hundreds of news organizations sharing content and publishing, tapping into the work of news enterprises such as ProPublica and Bellingcat, non-profit news organizations that conduct deeply-reported investigative news projects. These projects often have high impact, such as the cooperative effort of more than 100 newsrooms involved in The Panama Papers investigation into a worldwide money laundering operation that involved huge datasets of leaked banking and other information.³²

31 "Center for Cooperative Media," Center for Cooperative Media, accessed February 24, 2021, <https://centerforcooperativemedia.org/>.

32 "International Consortium of Investigative Journalists," International Consortium of Investigative Journalists, accessed February 24, 2021, <http://www.icij.org/>.

Technology

Finding 7

Some use asset systems as archives, others rely on web CMS

When it comes to technologies used for news preservation, we found three kinds of systems used for this purpose among the news organizations we interviewed: archive systems, also called preservation repositories; Digital Asset Management (DAM or MAM) systems doing double-duty as an archive; and sites that rely solely on their web CMS platform.

Here are the key data points on the types of technologies used for digital content preservation at the news organizations we interviewed for this project:

- Most of the news organizations we interviewed use some kind of technology to preserve at least part of their content internally (22 of 24 news organizations). Source: Interview responses to parts of Questions 15 and 16; see Appendix B for full text of questions.
- Most also rely on their web CMS or production system for internal news preservation for at least part of their content (22 of 24). Source: Interview responses to parts of Questions 15 and 16.
- Four of the news organizations we interviewed rely solely on their web CMS for internal content preservation. Source: Interview responses to parts of Questions 15 and 16.

Meaning, impact, observations: It's important to note here that there are essentially no standards in the industry for the kinds of technology systems used to preserve digital news content. What we found in our interviews is a wide range of systems and approaches, from formal archives to essentially no specific tool or system, only production platforms doing their primary tasks and doubling as the place where content is kept. Here's a deeper look at each of the three types of technologies we found in use.

Archive systems, or preservation repositories, are tech platforms designed specifically for preservation, with capabilities developed by expert media preservation staff, and often utilizing techniques from the larger preservation and academic communities, including metadata standards such as the PBCore and Dublin Core metadata systems.^{33 34}

One example of this type is the Artemis system designed and built by NPR as their archive, dating back to the early 1970s when the popular All Things Considered program was first aired. Artemis is designed to serve NPR's internal newsroom and producer needs for access to past content as well as provide long-term preservation. In this role it is NPR's "official record of broadcast," according to its mission statement. And it's a good example of a system that utilizes a metadata standard, in this case the PBCore standard.

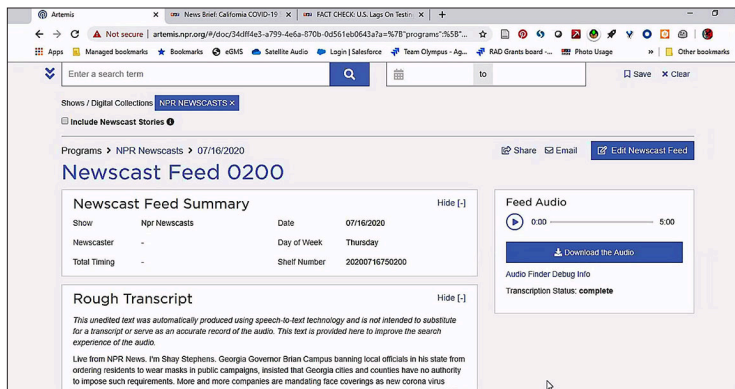
"Our metadata schema was from PBCore," said one member of the Research, Archives and Data Strategy team at NPR, which manages the system. "And since then, it's based off needs, it has definitely become much more unique. ... (but) we try and make sure that we're adhering to the majority of those that we can."

33 "PBCore Metadata Standard," PBCore, accessed February 24, 2021, <https://pbcore.org/>.

34 "DCMI: Home," Dublin Core Metadata Initiative, accessed February 24, 2021, <https://dublincore.org/>.

Technology

Artemis is also an example of the multi-function nature of some of these systems, since it began as an asset management system at first, then developed over time into a more formal archive. Another mark of its preservation nature is the way it stores content data. The Artemis system employs data storage systems that use two formal preservation techniques, maintaining multiple copies of data at two separate locations to ensure redundancy and also using a technique called fixity checks, which regularly test stored data to ensure no changes have been made, and to provide a record of any changes.



Screen capture of web-based interface for NPR's archive system called Artemis, which began as an asset management system in the 1970s.

The second type of approach is the use of Digital Asset Management (DAM or MAM) systems. These are typically designed for daily news production work, often for photo and video content, but also have some of the capabilities of permanent archives and are used for this purpose by many sites. In our interviews we encountered a number of cases where digital assets systems were used as archives for many content types, in addition to production functions. They are used to preserve text, photos, graphics and pages used in print products. Some also save content elements that appeared only on the web, not the rendered HTML pages but the core content, such as full-length versions of a story that was shortened for print.

For example, the commercially available DAM platforms such as SCC MediaServer or MerlinOne are used in many cases for daily production functions such as managing news assignments and other news planning functions, and to feed photo or video content to publishing systems for legacy news organizations. But in their secondary role as archives (The Dallas Morning News for example), they also provide long-term access to original photos, graphics, video, text or other digital content objects used in news production. This often includes large collections of past content as far back as digital data exists.

McClatchy and Tribune, for example, did this when they installed centralized publishing systems across all of their respective newspapers in the period between 2007-2017. As part of their centralized systems, they collected all available past content that was in digital form at the time of installation for each newspaper. This past content included text in digital form, back to the early 1980s, photos in digital form back to the 1990s, and page PDFs and news graphics back to the same period, stored permanently in their SCC MediaServer systems. In McClatchy's case this added up to 24 million news stories, 14 million news photos and 7 million PDFs of pages from 30 newsrooms across the U.S.

In the third type of approach, the web CMS is also acting as the sole preservation platform, especially for news organizations that do not have any other major data systems. This is especially the case for many newer, digital-native news organizations, we found, which often have only a web CMS in use.

Technology

Across the board in our interviews, we found most news organizations now relying at least in part on their web CMS for content preservation, even those that employ one or more of the other two systems outlined above. To clarify, this means that at least some of the news content exists only in the web CMS.

This approach to preservation is an example of the multiple roles that web CMS platforms are called on to perform in today's newsrooms, in addition to its primary role of managing content to produce rendered web pages of news content.

Here's a table showing all the major digital and media management systems at the 24 newsrooms we interviewed in detail. This shows the number of individual major roles served by the content management, asset management or publishing systems at each newsroom we visited; some systems serve multiple roles.

Figure 21: Number of major functions served by newsroom systems

PUBLISHING ROLE OR FUNCTION	NUMBER OF FUNCTIONS SERVED BY MAJOR SYSTEMS
Web publishing	23
Archiving	21
Print publishing	10
Video	8
Broadcasting	8
OTT	1
CMS/distribution	1

Source: Interview responses to parts of Questions 15 and 16; See Appendix B for full text of questions.

Archive or DAM use improves quality

In our analysis, we found a connection between the types of technology in use and how well the news organization does in content preservation. The key was whether or not the organization's overall technology architecture included the use of an archive system or asset management system of some kind other than a production CMS to preserve at least part of their content.

The link between these two was apparent during our interviews. To confirm this, we compared content collections at sites that rely on web CMS platforms for part of their news preservation with those that do not. Here's what we found:

- Those news organizations that use an internal, dedicated preservation system of some kind score much higher on the SPOT scale of archiving assessment than those that rely on their web CMS for part of their content preservation, with median scores of 2.7 vs 1.65, respectively, on the six SPOT model attributes. Source: Interview responses to parts of Questions 15, additional analysis of SPOT ratings in Figure 8. See Appendix B for full text of questions.
- Even more significantly, news organizations with dedicated asset management capabilities scored far higher than those that rely solely on their web CMS with median scores of 2.7 vs 1.3, respectively, on the six SPOT model attributes. Source: Interview responses to parts of Questions 15, additional analysis of SPOT ratings in Figure 8.

Technology

Tech providers constantly developing new capabilities

To better understand the influence of technology on preservation, for this research project we not only gathered data from the 24 newsrooms we interviewed. We also set out to meet with and interview as many of the key technology providers in the news industry as possible.

While the field of competition for publishing systems declined dramatically as the industry began to struggle over the past decade, we interviewed four that represent a good cross section of the news publishing and CMS industry, serving both small and large publishers, in North America and overseas.

We conducted lengthy interviews and in two cases visited with these firms:

- **TownNews:** owned by Lee Enterprises, whose web and print CMS and related modules are used by many small-to medium-sized newspapers across the U.S. This was an in-person meeting and interview.
- **Stibo DX:** formerly CCI, is a Denmark-based firm whose CUE Publishing platform is used by numerous large metro news organizations in the U.S., Western Europe and Asia, via video conference calls.
- **MerlinOne:** based in the Boston area, has been providing DAM systems to the news industry for several decades, and is branching out into the medical DAM field. This was an in-person meeting and interview.
- **SCC:** or Software Construction Company, based in suburban Atlanta, has also been providing DAM systems for more than 30 years, and offers options to use their platform for newsroom planning and editing workflows as well as archiving, via video conference call.

In addition, we participated in a webinar by WordPress on its NewsPack offering, a package of tools and services for digital news organizations. And through one limited interview we gained some information on Arc XP, which is used by a number of the newsrooms we interviewed as part of this project. We were unable to set up a full interview with Arc XP, which emerged a few years ago from technology development work at The Washington Post.

In reviewing the state of publishing technologies with these technology providers, we found a number of important factors worth taking into consideration. Here's a quick briefing on the key observations and developments from our conversations with news technology vendors:

- **Functionality gap:** Most technology providers see a significant gap between the full capabilities of their technologies and the realities of how they are installed and used. These include key technology tools not installed, activated or configured and those not properly configured. This applies especially to metadata tools and workflow functions. They also see frequent examples where fully implemented systems are not used properly, or at all, due to inadequate training of users or inattention to enforcing workflow requirements.
- **Lack of awareness:** One offshoot of this issue is a lack of awareness of the full capabilities of their systems within newsrooms, for tools they would like to utilize that they simply don't know are already in their system. While largely unintentional, vendors report this can be caused by inadequate communication with users, and also by the sheer complexity of modern news publishing systems. We heard a number of examples of this during our interviews.

Technology

- **Preservation awareness:** One factor that emerged from discussions with tech providers was a surprisingly high level of awareness of the need for content preservation capabilities. While vendors all noted a general decline over the past decade in customers' ability to afford general news archiving systems, with some customers eliminating these as in-house systems in favor of relying on their web CMS or third-party services, most seemed fully aware that the need has not declined.
- **DAM functionality:** One apparent trend that seems to be developing is the increased adoption of technologies for Digital Asset Management (DAM). While not universal, a number of current popular CMS providers are seeing the increasing need for DAM capabilities, as part of the drive toward more multi-channel digital publishing. To meet this need, we learned of tech providers building or expanding DAM-like offerings to handle at least some parts of current digital content, especially video. These include Brightspot's MediaDesk, the Field59 system from TownNews' BLOX platform, Arc XP's VideoCenter, Avid Access and the Digital Collections offering from Stibo DX. These are in addition to existing standalone DAM providers such as SCC MediaServer and MerlinOne.
- **Digital Collections:** The most ambitious of the new DAM offerings from publishing vendors is the recent purchase and integration of Digital Collections's DCX system into the CUE Publishing platform offered by Stibo DX. This combination, now called CUE DAM, offers the only example we know of in which a publisher can tap the full functionality of a tightly-integrated sophisticated preservation system, developed over decades, directly within the larger publishing interface used by newsroom staff.

Here's a list of all the major systems used by each of the news organizations we interviewed, and the names or brands of those systems, including internal names for custom systems that were developed in-house. (Note: this is as complete as we could make it, but in some cases there are additional major systems not listed.)

Figure 22: What are the major newsroom systems in use and what functions/roles, in addition to archiving, do they serve?

NEWS MEDIA ORGANIZATION	SYSTEM NAME OR BRAND	PURPOSE OR CURRENT ROLE / FUNCTION	ADDITIONAL ROLES / FUNCTIONS
The Associated Press	ECR (custom)	CMS/distribution	archiving
Baltimore Afro-American	WordPress	web publishing	archiving
	Archon (ArchivesSpace)	archiving	
BBC	Forge (custom)	web publishing	
	BBC Images (custom)	archiving	
	MirrorWeb	archiving	
	Vizrt	broadcasting	video
	Scisys	broadcasting	
	PIP (custom)	video	
	Jupiter (custom)	video	
The Boston Globe	Methode	print publishing	archiving
	Arc XP	web publishing	
	VideoCenter (Arc XP)	video	
Chicago Sun-Times	Chorus	web publishing	archiving, video
	MerlinOne	archiving	
	YouTube	video	

Technology

NEWS MEDIA ORGANIZATION	SYSTEM NAME OR BRAND	PURPOSE OR CURRENT ROLE / FUNCTION	ADDITIONAL ROLES / FUNCTIONS
Chicago Tribune	Arc XP	web publishing	
	CUE Print (NewsGate)	print publishing	
	VideoCenter (Arc XP)	video	
	SCC MediaServer	archiving	
CNN	MIRA (custom)	archiving	
	Name unavailable	web publishing	
Columbia Missourian	BLOX	web publishing	
	BLOX Total CMS	print publishing	
	MerlinOne	archiving	
The Dallas Morning News	CUE Print (NewsGate)	print publishing	
	Arc XP	web publishing	
	Preservica	archiving	
	MerlinOne	archiving	
GBH	Nav-X (custom)	broadcasting	
	MARS (custom)	archiving	
	MLA (custom)	archiving	
	Core Publisher/Grove (custom)	web publishing	
KBIA	Core Publisher/Grove (custom)	web publishing	
	AudioVault	broadcasting	
KOMU	BLOX	web publishing	
	Ruby Shore	web publishing	
	BitCentral	broadcasting	
KWMU	Core Publisher/Grove (custom)	web publishing	
	AudioVault	broadcasting	archiving
Los Angeles Times	Graphene	web publishing	
	CUE Print (NewsGate)	print publishing	
	MediaDesk	video	archiving
	BRS (custom)	archiving	
McClatchy Corp.	CUE Web	web publishing	archiving
	CUE Print (NewsGate)	print publishing	
	Brightcove	video	
	SCC MediaServer	archiving	
Newsy	Primestream	OTT	
	Dashboard (custom)	OTT	
	Vizrt	OTT	
NPR	Core Publisher/Grove (custom)	web publishing	
	NewsFlex	broadcasting	
	Seamus (custom)	broadcasting	
	Artemis	archiving	
PBS NewsHour	WordPress	web publishing	
	Avid	broadcasting	archiving, video
	MLS (custom)	archiving	
QCity Metro	WordPress	web publishing	archiving
	Nikon Snapbridge	archiving	
Quincy Media	Avid	broadcasting	archiving
	Naviga	print publishing	
	BLOX	web publishing	
	WordPress	web publishing	

Technology

NEWS MEDIA ORGANIZATION	SYSTEM NAME OR BRAND	PURPOSE OR CURRENT ROLE / FUNCTION	ADDITIONAL ROLES / FUNCTIONS
St. Louis Post Dispatch	BLOX	web publishing	
	CMS	print publishing	
	Field59	video	
	SCC MediaServer	archiving	
Stars and Stripes	Polopoly	web publishing	
	Prestige	print publishing	
	Spring CM	archiving	
	CUE Web	web publishing	
	CUE Print (NewsGate)	print publishing	
Vox Magazine	BLOX	web publishing	
	BLOX Total CMS	print publishing	
	MerlinOne	archiving	
Vox Media	Chorus	web publishing	archiving

Note: table shows major systems in use at the time of interviews November 2019 to September 2020, and do not reflect changes since. Source: Interview responses to parts of Questions 15 and 16; see Appendix B for full text of questions.

Metadata is the key: without good metadata content is effectively lost

It's a paradox within a paradox. Not only is digital news content far more fragile and threatened than is generally believed in this increasingly digital world. But in the age of powerful search engines, it can also be surprisingly hard to find when you go looking for something. The primary reason: missing or inadequate metadata.

While search engines and search platforms have proliferated to the point of dominance in modern news technology environments, by themselves they cannot overcome the limitations of content that lacks good metadata. Search results can't tell, for example, whether a news story is breaking news or a personality profile unless it is marked as such in the metadata. They can't tell whether a photo or video was taken by a news photographer or a family member or a PR office unless the metadata shows who shot it. They can't tell who owns the photo, video, audio or graphic, when it was created, where it was created and much, much more about it without good metadata. And it can't tell, even though it's right there in your system, whether you actually have the rights to publish that image or video, or not, and in what channels or products.

In other words, the metadata that's associated with news lies at the very heart of the content preservation and reuse process. Since search provides access and access lends purpose to preservation, it's no exaggeration to say that content metadata is the secret ingredient, the key, in the digital news preservation lock.

For this reason, we tried in this research project to understand as deeply as possible the current practices, workflows, tools and policies that impact the degree and quality of metadata in the news content we examined. We reviewed the ways metadata is created and accumulated during news reporting, editing and publishing or broadcasting; the metadata structures and standards in use, the policies that exist for how news staff should add and improve metadata throughout the new cycle. And the success of these factors in the overall outcome, the metadata that's now saved with digital news content created every day.

The results are discouraging. What we found is generally weak, limited, fragmented metadata and weak or nonexistent practices in the newsrooms we interviewed, especially for former daily newspaper or legacy news operations that once had much stronger efforts. Newsrooms that once

Technology

ran well-thought-out workflows and good quality control that built up decades of highly valuable metadata for news content preservation have seen these dissolve as staff disappeared through layoffs, workloads multiplied, and preservation concerns were pushed well down the priority list in the struggle merely to survive.

This is not a recipe for good metadata. One factor is the multiplicity of different CMS platforms and other technologies. In most cases we examined, each CMS used its own metadata structures, with little to no overlap with others outside of standards such as IPTC fields for photos. In very few cases did we find integrations designed to synchronize these separate and often conflicting metadata sets.

Nevertheless, in our interviews we also learned about some very strong systems and metadata practices, primarily at the large public broadcasting operations such as GBH and National Public Radio, where strong preservation leadership has built some of the most extensive and highly effective operations in the news industry, resulting in often outstanding content metadata. In addition, The Associated Press has one of the most extensive metadata sets anywhere for news content. The AP is also a pioneer in this area, with digital systems, linkages and descriptive metadata going back nearly 20 years.

High-quality preservation models were not limited to the U.S. We also learned about strong metadata operations at the BBC in the United Kingdom and at a highly effective television preservation organization near Amsterdam, the Netherlands Institute for Sound and Vision.

We'll examine these in more detail, but first in this section we present our main findings on metadata in current news systems and processes, followed by the results of a limited quality analysis using the SPOT preservation standard. Then we'll share growing techniques for automated descriptive metadata, some experiments in artificial intelligence applied to this area, the special challenges faced by multi-newsroom companies and developing trends in the structure of news content.

Finding 8

News metadata is often haphazard, inconsistent

Our interviews with news organizations showed that most currently operate under metadata policies and procedures that are often broad, non-existent or unenforced. Few of the newsrooms interviewed had written policies on metadata. Many of those we interviewed were either unaware of the existence of written policies or documentation metadata or other functions. Training for reporters and photographers in this essential area has diminished over the years as a result of cost-cutting efforts.

We also found that most of the web CMS platforms use metadata structures designed largely, or solely for web publishing. For example, most web CMS platforms we examined assigned categories or nodes to stories, such as Local News, Business or Sports. These are primarily used for website navigation, to determine location on the site, and are often automated based on the type of reporter writing the story (business reporter, sports reporter, etc.), an increasingly unreliable association as news staff numbers decline and reporters work on multiple news topics that switch frequently.

In some web CMS platforms, certain types of metadata are stripped out when imported, for example with photos or videos. When published online, this leaves such content orphaned without the ability to track or trace origin, authorship, licensing rights and other factors.





Technology

Of the web CMS platforms we learned about, few enforced metadata entry as a condition of import to the database. This differed by content type with text most often requiring minimum data entry while none was required of other content types. The key reason for this, we observed: most CMSs are designed for ease of use and speed to facilitate the fastest possible news publishing. This is an understandable top priority, but one that works against preservation needs in not capturing at least basic subject information known at the start of the content creation process, the ideal time to do so. And for platforms where mandatory metadata entry is possible, in many cases it was not configured when the news organization initially installed the system, also for speed and ease of use.

Here is the data for our findings regarding the structure of news content in publishing systems and the metadata systems and workflows we encountered in the interviews.

- Nearly all of the systems analyzed use unique identifiers for locating and retrieving all content, 18 of 24. Four sites rely on file names for identifying at least part of their content. In addition, one site does not use unique identifiers, and one site uses unique identifiers for part of their content. Source: Interview responses to Question 20; see Appendix B for full text of questions.
- Industry standard metadata structures are used fully or substantially by 5 of the 24 sites; another 17 used them partially; one newsroom used no standard metadata; and one site did not respond to the question. Interview responses to Question 19.
- Content linkages that interconnect different elements in a story are used partially by 17 of the 24 sites. These kinds of linkages, which allow users and systems to access all parts of a story from any single element, are used fully in only three of the 24 sites. Five do not link content elements at all. The ability to connect, or reconnect, all elements of a story is essential to the ability to recreate, to re-render the story in a way similar to how it was originally published or broadcast. Interview responses to Question 18.
- Most of the newsrooms we interviewed actively tailor descriptive metadata for use by search engines, especially Google Search. Of the 24, 15 actively engage in the search engine optimization (SEO) process in which reporters and editors add factual terms and phrases intended to be found and indexed by search engines. In most large newsrooms today, there are designated individuals or a team whose job includes keeping abreast of changes that happen continuously in search engines and processes used by Google, social media and other platforms. Other newsrooms we interviewed do this partially or not at all. Interview responses to Question 23.






Figure 23: Is content uniquely identified? How is this done?

	SITES THAT DO EACH	
Yes: unique IDs	18	
Yes: file names or similar	4	
No	1	
Partial	1	

Source: Interview responses to Question 20; see Appendix B for full text of questions.




Technology

Figure 24: Does any of the metadata used in your organization follow or utilize any of the archival specification standards currently in use or under consideration, such as IPTC, Dublin Core, PBCore, PREMIS, etc.?

SITES THAT USE STANDARD METADATA	
Partial	17 
Substantially	4 
No	1 
No response	1 
Yes	1 





Source: Interview responses to Question 19; see Appendix B for full text of questions.

Figure 25: Do your publishing and preservation systems use, and preserve, structures or metadata that connect all parts of a published story? Is there a way, for example, to find not only the text of a story, but also images, videos, graphics, social media, podcasts, broadcasts, print and other channel representations for each story?

SITES THAT USE LINKING STRUCTURES	
Partial	17 
No	5 
Yes	3 

Source: Interview responses to Question 18; see Appendix B for full text of questions.

Figure 26: Do you tailor metadata to any specific search engines, platforms or browsing/consumption environments?

SITES THAT TAILOR SEO FOR SEARCH ENGINES	
Yes	15 
No	5 
No response	2 
Partial	2 

Source: Interview responses to Question 18, 19, 20, 23; see Appendix B for full text of questions.

Meaning, impact, observations: Descriptive metadata, information that helps clarify what's the content is about, as opposed to technical or administrative metadata, is an especially weak area for most of the news organizations we interviewed. Part of the challenge here is that journalists, as the creators of news content, are logically an ideal source of metadata. In recognition of this, nearly all the newsrooms we visited expect reporters and photographers to add tags such as words or short phrases that describe their stories or images as a part of their regular workflow. Basic metadata such as this often includes entries for date published, byline, multiple headlines; slug (a short, coded name for the story); and filename. The amount of metadata, type of information and whether it is required or not varies from one newsroom to another.

Another complication related to gathering consistent and accurate metadata is that there are any number of ways for content to enter the CMS or DAM of a news organization. For example, text content may be originally created in Microsoft Word or Google Docs and copied and pasted into a CMS, an approach that would likely obscure or erase altogether the actual creation date, history of version changes and other metadata about the origins and history of an article. Likewise, photographs can be added to a CMS without information about who took them—whether staff, freelance, wire service—or what rights the news organization may have to publish them, leaving the door open for possible copyright infringement or other legal vulnerabilities. Although this information can be embedded in IPTC metadata fields, some web CMS platforms do not read these fields, or read only some.

For text stories, this process can be facilitated by functions in the CMS or other software that provide “pick lists” or controlled vocabularies that limit or guide the selection of terms that might apply to

Technology

content. Unfortunately, setting up metadata tagging systems in a CMS isn't always a top priority given the range of demands in contemporary newsrooms, leaving the choice of metadata terms up to the individual. This results in a situation where it can be difficult or impossible to effectively search for an article or set of articles about a particular topic because the metadata describing them is inconsistent. Allowing each individual to decide which words to use when tagging stories and photos contributes to this issue. Well-structured metadata following taxonomies and using controlled vocabularies can help remedy such problems.

Although ubiquitous and enormous in quantity, text content does have the advantage of lending itself to computerized mining and analysis. In this realm, The Associated Press (AP) has become a leader in what they call "automated classifications services." As one of the largest news agencies in the world, the non-profit cooperative is published by more than 1,300 news organizations. It handles an estimated 2,000 individual news stories each day from operations around the world.³⁵ Providing efficient access – primarily to news outlets, not the general public – to the millions of articles in the AP text archives is a key competitive advantage. Much of the data science that powers the delivery of news to AP customers relies on the creation and application of good quality metadata of all kinds, including descriptive terms. The AP's Metadata Technology Group, with a staff of eight, oversees the development of library science functions such as schemas and classifications which power the heart of effective content delivery. To date the team of metadata experts at AP have found that they cannot rely exclusively on machine learning or artificial intelligence to correctly organize and apply descriptive metadata to stories. Instead, they rely on a rules-based auto classification system, computer algorithms guided by human-made taxonomies, to determine which categories the stories are sorted into.

Interestingly, one of the practices that enhances AP's ability to use automated metadata tagging processes is the pervasive use of The Associated Press Stylebook, which as of 2020 is in its 55th edition. Most newsrooms consider the Stylebook to be the gold standard for news writing. The consistent use of capitalization, abbreviation, punctuation, spelling, numerals and many other common writing issues improves the ability of AP's computer programs to correctly identify the meaning of stories and apply appropriate descriptive metadata to such content.

Technology can assist in the quest for better metadata

AP isn't unique in its automated approach to metadata enrichment. Many modern CMS platforms include features that analyze the meaning of a story and either select metadata or provide users with a list of suggested terms to apply. Often this process is facilitated by the use of software services based on natural language processing (NLP) and machine learning technologies, which helps computers interpret human language.

At the BBC, one particular application of artificial intelligence, speech-to-text conversion, has transformed the exploration of audio and video archives and helped overcome some weaknesses in previous archival practices. This has allowed the UK's public broadcaster to dramatically expand the ability to search content in their vast radio and TV archives. For example, the casual mention of the name of the house band playing in the background of a BBC program about hippie culture from 1964 led to the discovery of the first known video ever of a Grateful Dead performance in the UK. Ideally,

³⁵ "The Associated Press," The Associated Press, accessed February 23, 2021, <https://www.ap.org/en-us>

Technology

sufficient metadata would have provided that information. Without speech-to-text it is unlikely that footage would ever have been found. Another unexpected treasure that emerged from the BBC archives due to speech-to-text is footage from a 30-minute audio tape that was marked simply “news” and the date. The tape’s descriptive metadata indicated it was about a football (soccer) match, which was only partially accurate. The first 20 minutes of the tape was indeed a match, but the rest of the tape revealed Dwight D. Eisenhower reacting to the announcement of the Marshall Plan, a truly historic moment. Apparently, the first portion of the tape had been reused because tape was expensive. Missing appropriate metadata to begin with and lacking this technological assist, this priceless footage would probably have been lost to history.

Digital media present complex metadata issues

Digital media pose another potential snag in the quest for more descriptive news metadata. Photos, audio and video accumulate metadata in a variety of ways, starting with the capture devices involved. Many cameras used by contemporary news photographers have the ability to capture both still images and high-quality video footage. Many of these newer cameras include features that allow photographers to preload them with a great deal of metadata so that general information gets automatically embedded into each image file at the time of creation. Photographers can even attach voice notes to specific images to provide detailed information or to apprise editors of special circumstances. Another potentially useful feature—GPS tracking and tagging—is available on most cameras, allowing image metadata to provide a set of coordinates describing longitude and latitude, or location for each image. Although these practices would likely improve the quality of metadata in many cases, we did not see evidence of their use in the newsrooms we talked to or visited.



Researcher Edward McCain, left, learns about photo metadata practices from Bill Greene, director of photography at The Boston Globe.

At most of the newsrooms we contacted still image metadata was added through a process prior to ingestion and entirely outside of the CMS. This procedure takes place after images are captured on a camera and written to a media card. The images are downloaded to a computer, often a laptop on location at the site of a news story. Descriptive and other metadata is then written directly into the image files using an application such as Camera Bits’ Photo Mechanic or Adobe Lightroom. Once this metadata is embedded into the digital photographs, the metadata travels with the image and so can be harvested while it is being ingested into the CMS.

Although common, this type of workflow may be problematic, in our observations, because it relies almost entirely on the individual photographers to create metadata for sometimes large sets of images, often under deadline pressure. In most cases, photographers do not have the benefit of a

Technology

preset taxonomy, controlled vocabulary or pick list to guide this process, resulting in inconsistent and unstructured tagging. Afterwards, this lack of structure for image metadata can make it difficult to reliably search for content. In many cases, the photo editors we spoke with indicated that when searching for images they relied on their own personal knowledge of which photographer was assigned to a particular event in a given time frame instead of relying upon descriptive metadata.

Much of our information is anecdotal. During the research team's onsite visits we reviewed the outtakes of photo shoots with photo editors or other staff, viewing dozens or hundreds of images. Our observations showed that scant metadata was attached to many of the images.

Images selected for publication may have more complete metadata, but the rest of the images did not. This makes it difficult if not impossible to retrieve alternative images for future use. Photo department protocols call for the application of "bulk" metadata, basic information about the assignment, at minimum. Given the sheer numbers of images created and flowing in and through editing systems, DAMs and CMSs, quality control to enforce metadata enrichment is difficult to achieve.

Another problem we observed involving photo metadata is more difficult to diagnose. In some cases, the creation date, an absolutely critical piece of information for the origin of any digital object, but especially for a news photograph, was missing or incorrect. When viewed in one news organization's DAM, the IPTC creation date clearly did not match the date indicated in the IPTC caption field.

Most images use the International Press Telecommunications Council (IPTC) Standard developed in the early 1990s.³⁶ The IPTC Standard includes dozens of predefined fields in groups such as Image Content, Image Rights, Event and Location, Status, Contact Info, Licensing etc. This is in addition to the camera's EXIF metadata, which captures technical information about each image in great detail, including shutter speed, focal length, aperture, ISO, and much more.

There seems to be less agreement about a metadata standard for video formats. For example, The Associated Press says it does not embed a lot of metadata in its video content, although it has plans to do so in the future. In recognition of this, the IPTC has defined a universal metadata schema for video called the IPTC Video Metadata Hub, in an effort to fill this gap.

Ironically, given the time and energy required to create and embed photo and video metadata, in many cases some or all of that information is deliberately removed by some web CMS platforms on import. This means that the published photo or video is, in effect, orphaned, since there is no easy way to determine its source, copyright or ownership. Our discussions with CMS providers and news organizations about the reason for stripping out metadata focused largely on privacy and safety issues for the subjects in the images. Instead of doing this on import, however, this could be done during the rendering process, leaving the original images with their metadata for later reference and reuse.

³⁶ "IPTC Photo Metadata Standard," IPTC, accessed February 23, 2021, <https://iptc.org/standards/photo-metadata/iptc-standard/>.

Technology

Although we did not encounter the use of artificial intelligence for adding metadata to images in the newsrooms we visited, the availability and capabilities of such technology are growing. For example, during our visit to MerlinOne, sources at the DAM provider said that they see AI features such as facial recognition and visual similarity as key developments. On the AI front more broadly, Google Cloud’s Vision API offers the ability to assign labels to images and classify them into millions of predefined categories. Vision can also detect objects and faces and read printed and handwritten text.³⁷

Unique metadata challenges for news groups

As with many industries, intense competition and economic stresses have driven a trend toward consolidation. We spoke with several newsrooms that were part of media groups such as The McClatchy Company, with 30 newspapers, and Lee Enterprises, with 75 newspapers. Both of these large news enterprises are exploring ways to leverage their size in order to increase efficiency in the face of staff reductions. One way to do that is to create systems that allow for efficient content sharing. For both McClatchy and Lee, part of the solution was to deploy the same publishing systems in all their newsrooms and manage some publishing processes centrally. Several reported that they still struggle with variations in web navigation structures and other taxonomies that do not fully match across all newsrooms.

For example, one newsroom might categorize a story as “education” but another one might tag the same type of story as “K-12” and so on, several newsrooms told us. To make things even more complicated, these terms may be tied to the names of the various sections produced by a particular news organization at a certain time. Those section names almost always change over time, meaning that the same kinds of content—even at the same organization—may well require a number of search terms in order to find them all.

In one interview, we learned of an effort to standardize tags and web destinations. This was at Lee Enterprises, which began this in April of 2019 and was still working to align these values when we met with them in January of 2020. A partial sample of that list is below:

Figure 27: Partial list of standardized sections, tags and pages in use at Lee Enterprises

STANDARD SECTIONS CHANNEL NAME	STANDARD TAGS CHANNEL KEYWORD	STANDARD PAGES SECTION (IF APPLICABLE)
Agribusiness/ranching	agribusiness	/business*
Architecture	architecture	varies
Arts & Culture	arts-culture	/entertainment*
Entertainment	entertainment	/entertainment*
Food & Drink	food-drink	/entertainment/dining OR lifestyles/food-and-cooking
History	local-history	news/archives

The full list of keywords is available in Appendix D.

Newsrooms in some markets, such as St. Louis, found that they needed to expand the keyword list because they were the only Lee property that had professional baseball and pro soccer teams.

37 “Vision AI,” Google Cloud, accessed February 23, 2021, <https://cloud.google.com/vision>.

Technology

Structure of content is changing toward more granularity

One trend we encountered that's likely to impact preservation efforts in coming years is the changing structure of news content. To meet the increasingly multichannel needs of digital publishing and broadcasting, the change we observed is that content is becoming more and more granular.

For much of the history of news publishing, the central unit of news content was the story, the text that told the tale of one event or person or phenomenon. This method of organizing content was reflected in the structure of publishing systems from the introduction of technology into the newsroom. Other organizational units were relevant as well, depending on the interested group, including full newspaper pages and editions in the form of PDFs or microfilm images, and full recordings of news broadcasts, or news programs. These units are no longer as relevant, due to shifting listener/viewer consumption patterns toward individual stories of interest on digital channels, identified increasingly through search or social media.

The digital age itself introduced the need for content architectures in publishing systems that could group and manage collections of content objects that are related to one story into articles or packages of content. These could include a text story and photo, plus photos and URL embeds, graphics and more.

But the explosive growth in channels and outlets in the digital publishing era has fueled the need for even more flexible content structures that are no longer tied to news story text at all. These novel frameworks facilitate new forms such as the increasing number of ways we now see news stories told in largely visual presentations, through video, animations, interactive graphics and complex web presentations that lead readers step-by-step through a visual sequence that tells the story.

The impact of these changing needs is significant and pose an additional potential challenge to preservation. The impact was reflected in workflow and technology changes we encountered during our interviews:

- We talked with several radio and television broadcasters who are increasingly organizing their news content in smaller units than before, as collections of individual stories rather than one seamless broadcast. These are stored independently within content systems to be assembled or reassembled in any sequence needed. One example is NPR, which used to produce and archive news content for radio programs strictly on an episode-by-episode basis, usually one-hour. Now, with the increased demand for content on the web and in NPR's mobile apps, this is shifting more toward story-level segments, resulting in far more story segments to preserve than was the case with whole episodes.
- One radio station we interviewed has seen such strong growth in the audience for the news website — including recent months when their web audience exceeded broadcast listeners — that they changed their workflow from broadcast-first to web-first and are modifying systems to accommodate story units rather than full news.
- And one technology company we spoke with, Stibo DX, has redesigned the content architecture of their CUE Publishing platform to reduce the unit of content to a far more granular level. In CUE's new ContentStore architecture, each content element stands as an equal and independent entity, whether photo, video, social media embed, active graphic, map etc. This extends to text as well, with each paragraph of a text story managed as a separate element. Elements are pulled together in any number and type, in any sequence, in CUE's Storyline tool.

Technology

There are other examples as well of the impact of these changes, and more to come. It also became clear as we learned about these kinds of changes that they are certain to impact the preservation process by raising the level of complexity needed in systems that store this kind of news content. While it's not yet clear exactly how this will impact preservation efforts, these underlying changes may reinforce the idea of preserving the core news content elements and their relationship rather than saving snapshots of the expanding number of digital channels to which news is published.

When we look across the findings in this study some common threads emerge that are important to highlight. The interviews surfaced certain recurring factors that seemed especially important in their potential impact on news preservation or helped explain why news organizations are having trouble dealing with preservation in the digital era. One is the common experience of difficulties in switching systems.

Finding 9

System migrations often lead to lost content

Of all the technology issues we learned about during our interviews, one stands out as a nearly universal problem: that at least some parts of news content will be lost during a switch from one publishing system to another.

This is an issue cited by nearly every one of the 24 news organizations we met with or interviewed. The problems of content loss in this process range from minor errors with small amounts of data are lost, usually parts of metadata, to cases in which key parts of stories or even entire stories are missing. There are often major problems of broken links with web CMS data transfers, which can lead to stories actually existing in a new database, but not showing up on the web and difficult if not impossible to find in the new system.

This fact that this is so widespread makes technology transitions a potentially grave threat to the survival of news content, one that should be spotlighted and addressed in more comprehensive ways than it is currently.

Here are some examples of what we learned:

- A news librarian at one large metro newspaper who told us he sometimes has to refer to their print archive and use Google Search just to locate an article that's not appearing on their own website, because of data translation errors that occurred during the migration of existing news content from the previous publishing system to a new web CMS.
- At one large metro daily newspaper, an editor cited content lost in the 2012 transition to their current publishing systems, especially metadata and related content. "Yes there was some data lost. It was the richer part of the data," he said.
- One small digital news startup we spoke with discovered thousands of stories missing after switching from a custom-designed web CMS to WordPress. The lost stories included some of the most popular and important stories this group ever produced, including exclusive coverage of a trial that no other local media had.

Technology

Meaning, impact, observations: This problem occurs because of the difficulty of reliably mapping the content structure of one publishing system with another. For example, some systems use standard database tools with fully fielded data, others use internal markup within data containers that must be precisely handled to properly transfer. In other cases, custom routines for recording data are unknown or not fully documented. And lastly, the process of data transfer and interpretation can be resource-intensive and expensive, leading cash-strapped news organizations to limit or short-circuit the process.

Regardless of the cause or specific issues in a switch from one system to another, this issue can be among the most frustrating of all technical issues in publishing. Some comments we heard illustrate these concerns.

“When we first moved over to the new system ...and you look at the caption and it has all this weird formatting,” said a news librarian at a large metro newspaper we interviewed, speaking of a system transition that introduced garbage characters into parts of photo metadata that sometimes still gets published that way. “I know what to look for now, you know, the accents, the dashes, the ampersands, the copyright symbols and stuff like that. So, you look for that stuff. You take it out, but people under time constraints aren’t necessarily doing that.”

“We lost a lot of really good content,” the publisher and editor of a small digital news startup told us. “They just didn’t flow into the new WordPress CMS as we had hoped. I can tell you about some of the stuff that was near and dear to my heart. For that reason I feel like I’m wedded to WordPress from now on. We’ve put out a lot more good content. Not sure if I’m ready to risk that. So, I feel like I’m wedded to it now.”

At a midwestern daily newspaper, one editor shared concerns about the way websites change their display configurations frequently, affecting older news stories with elements that no longer display properly.

“What I’m trying to say is like, the site since 2014...we redesigned several times. And so the site doesn’t even look like the way it did back then. So, I can go back and pull up that ... story, but it won’t look the way it did back then.”

“So, you know, I asked, like, do we care how the story was first presented and what was attached to it at that point? Yes, we do.”

Practices Findings

When we look across the findings in this study a number of meta issues emerge in the practices and adaptations that have evolved to deal with the enormous financial stresses and tensions the news industry is facing. While not necessarily intentional, these kinds of practices do appear to drive significant changes in newsroom activities and help explain why news organizations are having trouble dealing with preservation in the digital era.

Finding 10

Financial stress on news industry displaces preservation

One problem we witnessed during our visits and interviews is the serious impact that years of financial stress has had on the news industry with its drive for greater efficiency. This affects all news media, but seems especially critical in the newspaper sector, where the old advertising-based business model has collapsed and the industry is struggling to find new models to support its essential role.

What's happened in the legacy/newspapers sector is that many newsrooms have been forced to focus narrowly on daily production requirements and demands for the next news cycle, to the exclusion of other needs. In our visits and discussions we heard from many newsrooms and their support groups that they have little choice but to focus almost entirely on the ever-expanding production needs of digital publishing and push aside all other considerations not directly tied to the now 24-hour news cycle.

This includes the needs of news preservation, along with other functions considered less essential. For example, most daily newspaper newsrooms no longer have a copy desk, instead relying on remaining reporters and editors to confirm their own facts. We also were told of deep cuts in support services, from admin to personnel to technology, the last a key challenge for news organizations struggling to stay abreast of rapidly expanding digital publishing needs.

Meaning, impact, observations: Financial stress has had a deep impact on news content preservation, from staffing to technology decisions to workflows, policies, monitoring and quality control. It's the key reason cited for the scarcity of news librarians in the newsrooms we visited; the rest let go in hopes of keeping more journalists on staff. At one midwestern newsroom, for example, the sole remaining news librarian informed us that the job would soon be eliminated.

It's the reason many cited for dropping their asset management systems and archive platforms and instead trying to rely primarily on a web CMS and outsourced services such as Newspapers.com and ProQuest. Not only have these moves saved the costs of the internal systems themselves, but also the expense of each system's technical support staff, many of whom are now gone from tech payrolls. At one newsroom we visited, for example the last full-time technology administrator for its print and photo archive system was recently let go.

Financial stress is the reason behind the examples that newsrooms showed us of workflows and policies that no longer work. In case after case, we saw examples of this issue, news content lacking critical metadata that had been required and utilized for many years.

In one newsroom, for example, an editor expressed deep frustration after pulling up news photos on screen with key metadata missing that they had long required, a concern the editor tied to losing staff who in the past would help enforce these standards. At another newsroom, we saw examples

Practices

of images transmitted from one publishing system to another that had lost key metadata, a problem attributed to a technical issue that staff said was reported but not yet resolved due to tech staffing issues. In another case, a news photo opened on-screen appeared to have short-circuited a long-standing workflow designed to ensure only one master version of the photo was preserved in the archive. The result: it was not clear if this was a duplicate or original image, a problem attributed to a rushed and overworked photography staff skipping key workflow steps.

Pressures also drive consolidation, centralization and standardization efforts

Financial stress is also the key reason that larger news companies with multiple outlets have focused so much effort on consolidating operations and systems, and in many cases content as well. These range from radio stations that broadcast similar programs produced in a central location to newspapers that operate centralized print design and production desks, including Tribune, Gannett, McClatchy, Hearst, Lee Enterprises and many others.

Our team had meetings and demonstrations with two such newspaper operations during our research, one for McClatchy based in Charlotte, another for Tribune based in Chicago, where this approach was first pioneered for the newspaper sector in the early 2000s. On both visits we saw highly effective central print production desks in which news content was first published in digital channels and then repackaged into print by each company's central design desk. We also learned about the centralized desk operation at Lee Enterprises during our visit with one of their newsrooms, the St. Louis Post-Dispatch.

All three companies used these centralization efforts to replace individual print design and production teams at each newspaper, 30 at McClatchy and 10 at Tribune, 11 prior to the Los Angeles Times 2018 sale to a local billionaire. Lee is still in the process of consolidating production at its 75 daily newspapers. As these desks show, the newspaper business model has changed so much, so fast over the past decade that it's become no longer sustainable to maintain local print production groups.

These moves to centralize and standardize news publishing systems and workflows have resulted in streamlined processes not only for print newspapers, which originate today as part of a born-digital stream, but also in the digital versions of print products, the so-called e-editions, or electronic editions, sometimes called e-paper. Some sites use their e-editions to offer expanded news content such as Lee and McClatchy, which published 86 extra pages for subscribers on a recent Monday, four times the size of the actual print edition. These e-editions offer a new opportunity to preserve born-digital content in a new form through the print-like e-edition, although not all of these companies were taking advantage of this by retaining all of this content. Many preserve only PDFs of these digital pages.

A similar centralization process has unfolded at National Public Radio in recent years, as we learned in interviews with the central office as well as two NPR member stations, KWMU in St. Louis and KBIA in Columbia, Missouri.

NPR has taken major steps to centralize and consolidate its content management systems to help bring local NPR affiliates into closer cooperation with the main newsroom. The purpose is to speed and simplify the process of sharing news content and programming, and also ease the technology and support burdens of non-profit radio stations that must sustain themselves primarily through local fundraising.

Practices

NPR's latest standardization effort is a new web CMS that is now being tested and rolled out to most local NPR affiliate radio stations across the country. The system, called Grove, is built on technology provided by one of the major industry vendors, Brightspot. This system will provide new preservation capabilities along with standardized workflows to speed the ability to publish and broadcast news programs from other stations and the central NPR newsrooms and studios.

Grove is replacing an earlier NPR system that's been in use for many years, called Core Publisher, which was custom built for use by NPR affiliates but was not often used by the central staff. The new system will be used the same way by all news and content staff, at local affiliates and the central office. This is another example of standardization providing a new avenue for simplified content preservation.

Finding 11

Migration to digital publishing incomplete, can mean lost content

Through our interviews and newsroom onsite visits we learned that not all media companies have successfully completed the shift to fully digital news operations and systems. One result is a problem we learned about at several sites where the news content for different channels was not synchronized between separate content management systems. This can lead to lost or incomplete content.

As the news industry transitioned over the past decade from older media channels such as print and analog radio and TV broadcasts to newer dominant digital channels such as the web, native apps, podcasts, social media and digital on-demand news, publishing systems and their workflows had to adapt to work in the news environments. The most common approach is a digital-first workflow in which news content is first created and published online for reader consumption through web browsers or native news apps.

Meaning, impact, observations: The problem arises because most news organizations require multiple, independent systems to handle different functions, often channel-specific functions. A web CMS for example, to publish content online, and often to auto-feed other digital channels such as native apps, platforms such as Apple News, Kindle, Google News, and more. These are usually separate from systems that handle legacy channels such as print publishing, or radio and TV broadcasting. Where these systems are not fully integrated with one another, content can become unsynchronized when changes are made in one channel platform and not reflected in others.

We heard many examples of this during our visits and interviews: at two midwestern newsrooms, for example, news stories still originate in the CMS used for print production, which automatically updates content to the web CMS when saved. However, any changes done in the web CMS are not reflected back in the print CMS, leaving that content incomplete, incorrect or outdated. To avoid this, some postpone digital publishing until print production has been completed.

"I see that happen like all the time on our sports game stories at night," said one newspaper editor. "We'll work it and then publish it online at 11 o'clock at night instead of, you know, 8:30 when the game might be over and people want to comment on it."

This can also result in content missing in archives or downstream systems such as online syndication services in cases where they are fed from the print system, after news stories are changed in the web

Practices

CMS and staff members do not take the time to manually update the print content. At one Midwest newsroom where web CMS content is auto updated to the print system, the solution is to break the link between the two in some situations.

“I don’t want them connected because you’re going to break all this design work I did,” said one editor. “So that’s when it usually comes to a head ... if they have not followed to the “T” what the template’s supposed to be, that’s where it messes up. And so, I’m not going to handcuff my (print) designers just for that reason.”

It’s a problem that stems in part from the increasingly complex technology environment that most large news organizations must deal with, systems acquired at different times under different leaders with changing priorities that don’t properly interact, systems with conflicting structures or workflow assumptions, systems that different teams favor, and so on. Although all of the news organizations we interviewed were aware of these issues, some painfully so, few have had the resources, the funding and staffing needed to remove all of these obstacles.

Finding 12

Relying solely on web CMS can be problematic for preservation

One of the key issues that emerged through our research is the potential for conflicts between preservation and some common web CMS functions for those that rely on their web CMS for news content preservation systems. This is not an immediately apparent or visible issue. But for any news organization that is relying largely on their web CMS, depending on how it is configured and used, we found this could be a potential problem area worth highlighting.

The reasons for this lie in the fundamental differences between the kinds of content needed for a web CMS, a production system designed primarily to build web pages, and the content needed for a news archive or preservation system.

Here are some examples of the differences:

- **Duplicates vs. original files:** Most web CMS platforms have little or no provision to ensure that content is not duplicated, such as multiple copies of a staff photo, video or graphic, possibly with different crops or duration for audio and video files. The reason? For web publishing purposes, there’s no reason to enforce uniqueness. One is usually as good as another in rendering the HTML for a web page. In contrast, news archives work best when there is only one photo, video or audio file, the original, and they utilize multiple techniques to avoid duplication.
- **Photo, video, audio size and quality:** Similarly for the web, the quality of still photos and video are often well below the color-depth, resolution, video frame rate or audio sampling frequency of the original digital content. Photos and video are optimized for web page reproduction, typically about 1,200 pixels wide. For example, until a few years ago McClatchy newsroom systems automatically modified and resampled any photos that were part of a news story package as they were sent to the web CMS for online publishing, changing the size, resolution and color space to optimize for the web at 1024 pixels on the long side. For staff photos from professional cameras especially, these kinds of processes are generally downsampling photos from much higher-quality originals, eliminating resolution and detail not needed for the web but potentially essential for later uses. If a news

Practices

organization is not keeping the original jpeg image file, they cannot be sure of the ability to feed a future digital channel that calls for higher resolution images than a current web page requires.

- **Metadata types, scope much different:** When we look at metadata for news stories sent to a web CMS for posting online, they tend to have little more metadata than what's needed to direct the story to the right location on the website, plus headlines, keywords and tags that can be picked up by search engines to help readers find the story. News archives need far more and are designed to capture wide varieties of metadata, including author, creation and publication dates, usage information about channels or publications where the story appears, ownership and origin, rights management and licensing data, and potentially large amounts of descriptive metadata about what's in the story, including what it's about, what types or genres or topics are related.
- **System performance needs differ:** For optimum system performance and load on a web CMS, it's best to keep content files as small as possible, to speed html rendering and multiple technical processes such as data transfers between different parts of the system such as for disk storage or CPU processing. In addition, it's best to periodically purge, or delete any content that's not needed. These operational needs are nearly the opposite of those for a news archive system, which is optimized for long-term preservation and designed to preserve full-sized original files, and ensure content not only is never deleted, but that any changes are tracked in detail.

Meaning, impact, observations: These differences may have little impact in cases where a news organization preserves original content in an independent archive platform, as a number of the organizations we interviewed do. So long as connecting linkages and specifications are retained for news story packages in this scenario, stories purged from a web CMS can always be re-created and re-rendered for the web in ways that are similar to the original publishing presentation.

Not so for the sites we interviewed that rely mostly or solely on their web CMS. In such cases, it's possible, maybe even likely, that the news content saved in the web CMS often has less metadata and lower visual or audio quality than the original content objects. And based on our interviews for this project, it's likely there will be multiple copies of many types of content, still photos especially, some or all without sufficient metadata to determine ownership rights or indications that an image was cropped from a larger original.

"I think a lot of the smaller sites have no archive... they might rely on the content management system to be an archive, which it is not," said one of the technology managers we interviewed for this project, who runs systems for a large metro daily newspaper. "That would be one of the key differences of a content system, is just, it's a workspace," he said. "A content system is a workspace. An archive is a history for reference.... They're completely two separate things."

When switching to their latest web CMS, the manager said they realized they could not bring over all the content they had accumulated over decades in their archive, it would be too much. Even now, with only seven years content, that volume of accumulated data has become an issue.

"You could never expect to store all of that," he said. "Now that conversation is reaching a real head because the amount of data that's stored in the content system makes upgrading the system difficult, makes backing up that system difficult. So, they're starting to see all these problems arise because content hasn't been purged from the (web CMS) system."

Practices

Finding 13

There's often nobody left to mind the archive store

During the course of our visits to newsrooms and numerous interviews online, one troubling development stood out as a strong indicator of the diminished state of news preservation: the virtual disappearance of news librarians and archivists from many media organizations.

This is a phenomenon that seems particularly acute for the daily newspaper segment of the market, but affects others as well, including radio and television stations. In contrast, public and government-owned media outlets continue to operate comparatively well-staffed preservation teams, as well as one large for-profit national and international media organization we interviewed. But across the legacy media sector, the absence of preservation or archive staff has become a significant issue.

From our observations and interviews we learned that, overall, there's been a dramatic reduction, in many cases wholesale disappearance, of news library and archive staffing devoted to content preservation over the past 5-10 years. The main reason for this has been cost savings efforts.

Here are some examples of the stark reductions we encountered in archive teams:

- Overall, there's been a dramatic reduction, in many cases wholesale disappearance, of news library and archive staffing devoted to content preservation over the past 5-10 years. The main reason for this has been cost savings efforts.

Here are some examples of the stark reductions we encountered in archive teams:

- At the Los Angeles Times, for example, there are now only four staff members engaged in archiving work where once there were 16 or more.
- At The Boston Globe, where the now-retired chief news librarian was so prominent in news investigations that her role was portrayed in the Oscar-winning film "Spotlight," her former 20-person staff is now down to one.
- At the St. Louis Post-Dispatch, which had a 22-person news library staff at its peak in the late 1990s, the newsroom now has only one.
- And at McClatchy, where each of its 30 newspapers once had news library staffs ranging from one or two to 10-12 each, there is now only one news librarian left across the entire company, at the Miami Herald.
- Of the news outlets we interviewed, only two out of three have any staff involved in news preservation (16 of 24). Of these, most have one person, some with part-time help, especially at legacy newspapers and groups. Eight news organizations had no staff whose job includes preservation work. There are many more if you look across entire media groups such as the 29 other daily newspapers with no news librarians at McClatchy alone. To clarify, this refers to work that was done for many years by archivists to ensure the quality of archived content; ensuring that it was complete, checking that photos or other visual media is included, along with captions or other text blocks, sidebar stories and linkages between these and related content that technology systems often miss. Depending on the outlet, this could also include work to preserve raw video, photo outtakes or other visual materials. Source: Interview responses to Question 4; see Appendix B for full text of questions.

Practices

- Of the news outlets we interviewed, only two out of three have any staff involved in news preservation (16 of 24). Of these, most have one person, some with part-time help, especially at legacy newspapers and groups. Eight news organizations had no staff whose job includes preservation work. There are many more if you look across entire media groups such as the 29 other daily newspapers with no news librarians at McClatchy alone. To clarify, this refers to work that was done for many years by archivists to ensure that it was complete, checking that photos or other visual media is included, along with captions or other text blocks, sidebar stories and linkages between these and related content that technology systems often miss. Depending on the outlet, this could also include work to preserve raw video, photo outtakes or other visual materials. Source: Interview responses to Question 4; see Appendix B for full text of questions.
- There's a very different picture in terms of trained archive staff only at the 11 public and government media organizations we interviewed and at The Associated Press, a private but non-profit news collaborative. Of the 24 news organizations we interviewed, there were a total of more than 74.75 preservation staff members. Most of these, 58.5, were at public and government media organizations and the AP. Of the 13 for-profit organizations that have some archive staff there were only 16.25 staff members in total. To the extent that this is representative of the entire for-profit segment of the news media, by far the largest segment, it represents a major decline from practices of the past.

In short, when it comes to the news content that will be counted on over the long term, there's almost nobody minding the archive store at legacy newspaper and broadcast media.

















Figure 28: Number of staff doing preservation work (sum of all sites)

TYPE OF NEWS OUTLET	
Public, non-profit, government	58.50
For-profit media	16.25
Total	74.75

Source: Interview responses to Question 4; see Appendix B for full text of questions.

Practices

Figure 29: Number of total staff defined, assigned in preservation work

NAME	TYPE	TOTAL STAFF	
The Associated Press	non-profit	8	
Baltimore Afro-American	private	1.25	
BBC	government	7	
The Boston Globe	private	1	
Chicago Sun-Times	private	0	
Chicago Tribune	private	1	
CNN	private	n/a	
Columbia Missourian	non-profit	1.5	
The Dallas Morning News	private	5	
GBH	non-profit	15	
KBIA	non-profit	0	
KOMU	non-profit	0	
KWMU	non-profit	0	
Los Angeles Times	private	4	
McClatchy Corp.	private	1	
Newsy	private	1	
NPR	non-profit	18	
PBS NewsHour	non-profit	6	
QCity Metro	private	0	
Quincy Media	private	0	
St. Louis Post Dispatch	private	1	
Stars and Stripes	government	3	
Vox Magazine	non-profit	0	
Vox Media	private	1	
Grand Total		74.75	

Source: Staff counts are from interview responses to Question 4; see Appendix B for full text of questions. News preservation groups differ in organization and roles so responsibilities may vary. For example, NPR's Research, Archives and Data Management group has a broad scope, the BBC's news preservation staff is part of a larger archive group, and CNN does have a significant preservation operation but was unable to share staff numbers.

Disappearance of news librarians had major impact on transition to digital preservation

Meaning, impact, observations: There are a number of potentially serious impacts that flow from this decline in preservation expertise in today's newsrooms. One is the impact on the quality of content that is fed to the archival systems and to external providers and other channels. Additionally, it seems clear that the lack of expertise contributes to industry difficulties in making the shift to new preservation workflows needed for digital content.

Unlike the past, when archivists worked to check the accuracy and completeness of news content before it was committed to archival systems and added critical descriptive metadata to help find it in the future, now these functions have largely disappeared at legacy news organizations. Instead, we saw many cases where newsrooms have been forced to switch workflows over to largely hands-off, automated technology functions. In some cases, we were told that management believed archive staff was no longer needed. This belief drove decisions to invest in new technologies in an effort to reap

Practices

cost savings, or to reduce news library staff rather than other newsroom staff such as reporters in the face of widespread revenue declines. Either way, the result has been a decline in content quality of archived material that news staff members who rely on these systems are aware of, but which is difficult if not impossible to measure.

The impact of the decline in staff expertise can be seen in other aspects of the news preservation problem as well. One is the absence of preservation expertise in newsroom staffing and technology decision-making. Without trained and knowledgeable specialists such as the news librarians of the past, there is no longer a voice advocating for the needs of news content preservation in many media management circles.

The decline and disappearance of news librarians may have also contributed to the slowness or inability of news preservation efforts to keep up with the shift to digital news publishing. With no advocates to guide changes and evolution in archiving practice, this may help explain why there is so little attention to saving digital news content at many newsrooms. It may also help explain one of our findings: why most newspaper publishers we interviewed still do an adequate job archiving print content, even if they have yet to adapt to the preservation needs of newer digital content types, such as social media or video.

This may also help explain the increased reliance we found among newsroom staff members on using external, third-party services such as Newspapers.com or NewsBank for daily news reporting background work, as well as for managing access by the public. In fact, some of the newsrooms that formerly managed their own archives for internal newsroom research told us their reporters and editors now rely almost exclusively on these external services. As noted above, the one most commonly cited by the newsrooms we interviewed is Newspapers.com.

Older media issues

The lack of expertise also affects the ability to maintain numerous types of older digital media that have not been preserved in a reliable archive system, and whose origin, content and underlying value is often known only to librarians who may no longer be on staff.

This is potentially devastating to informal content collections we saw and learned about at a number of news organizations we interviewed. Many sites acknowledge the existence of critical news content types that are stored informally in their newsroom on older or outdated digital media such as CD-ROMs, DVDs and a wide assortment of cartridge disk drives. This usually involved raw visual content such as photo outtakes, raw video or tapes of television and radio programs, both analog and digital.

At one major city newsroom, for example, we were shown a storage hallway off the main newsroom containing dozens of drawers full of CDs, DVDs and disk drives holding digital news photo outtakes and original published images from staff photographers going back into the 1990s. Most of these materials had not been uploaded to any archive system, existing only on these original, outdated media.

At another, a radio station, our host took us to a nearby storage room where he opened the doors to a cabinet showing shelf after shelf of stacked LTO digital audio tapes of broadcast content over many years, each with a handwritten label as the only indication of its contents. Many of these tapes constituted the only record of past radio broadcasts for this major-city radio station.

Mission Findings

It's not uncommon to see or hear of examples like this across the news industry, disks with original still photographs, video files or complex database projects sitting on a shelf or in a newsroom desk drawer, shelves of digital or analog VHS videotapes at local television stations, otherwise unpreserved, poorly documented and unknown except to a handful of staff.

In many cases newsrooms have partnered with local memory institutions such as public libraries and universities to help preserve these and older content materials. We learned of a number of such examples during our interviews:

- The Boston Globe, which partnered with Northeastern University when the news operation moved out of its huge Dorchester complex to leased downtown office space in 2017. The University library now houses a large Globe photo collection going back decades.
- The Baltimore Afro-American partnered with nearby Morgan State University, which now houses and preserves much of the historic news content, photos and other materials from the 128-year-old newspaper serving the region's Black community.
- The Raleigh News & Observer partnered with the state archive to take charge of many millions of non-digital photo negatives and prints, while the Charlotte Observer, also McClatchy, partnered with the local public library to preserve decades of photo negatives and prints.
- The Chicago History Museum, which took custody in recent years of a large set of news photos from the Chicago Sun-Times, which included some born-digital images as well as older images.

But as officials at these institutions report, they also face significant challenges in keeping pace with technology changes to enable them to adequately preserve digital news content and make it available for use.

Qualified staff scarce

On the flip side of the problem of declining library staffing at legacy news organizations, we also heard from some news outlets that are looking to hire preservation staff that they have a great deal of trouble finding qualified candidates.

The issue is not just finding those with requisite library and archival training, but also identifying candidates with a sound knowledge of modern digital technologies. This is the key holdup for two of the news outlets we talked with. Both cite this as a new, emerging field of expertise in digital preservation and an opportunity they would like to see universities work to address.

When we step back from the findings in this section, we see some additional patterns that tend to indicate how well a news organization is doing in preserving digital news content. These range from the role of preservation in a news organization's mission, the existence and nature of preservation policies, the alignment of technology with preservation needs and the organization's track record of commitment to news as a vital part of a community's public record.

Mission

Finding 14

Good preservation is linked strongly to mission and policy

Of all the factors at play in news preservation, the one most closely associated with good practices is whether or not news preservation is built-in and recognized as a part of the organization's mission. It also matters whether this is explicit, in the form of a mission statement or as part of a written policy, rather than something implied through current or past practices, handed down verbally and at risk of being lost.

Here are the key data points on this association:

- When asked about mission, 19 of the 24 news organizations we interviewed said yes, preservation is part of their mission, either formally or implicitly. Source: Interview responses to Question 2; see Appendix B for full text of questions.
- However, of these, only six reported that this was explicit, meaning written or formally acknowledged in some way as part of a mission for the organization, or at least for the team involved in preservation work. Source: Interview responses to Question 2.
- Most of the newsrooms where preservation is explicit are at public radio or TV stations. Source: Interview responses to Question 2.

Meaning, impact, observations: This last data point is the connecting link. What we found is that the news organizations that scored highest in our content analysis are, by and large, also the ones that have preservation baked into their mission, their culture. We found many different ways in which a preservation mindset manifests itself in the news organizations we talked with. But it's clear that those who are most explicit about the role of content preservation, the public news organizations, are also the ones doing the best job among the 24 we interviewed. In many cases their mission is clear, stated and publicly available.

For example:

- The most extensive mission-related document among news organizations we interviewed is the royal charter for the BBC, dating back to 1927 and modified many times since. It includes a section in the current version stating all content must be "kept safely to commonly accepted standards," and made publicly available without charge. (See the full history of the BBC Royal Charter here: <https://www.bbc.com/historyofthebbc/research/royal-charter>)
- Similarly, the U.S. public broadcasting news organizations we met with, National Public Radio, the Public Broadcasting System and various member stations of both groups, also call for content preservation. As one member of NPR's Research, Archives and Data Strategy team described it, their archive system's mission is "to improve and enhance NPR's archival database to support production and business needs, while ensuring long-term preservation and access to NPR produced content and serving as NPR's official record of broadcast."

Mission

- GBH, in addition to its own mission to “Improve, for all people, access to public media,” (<https://www.wgbh.org/foundation/who-we-are>) is involved in a number of preservation related efforts, most notably in the American Archive of Public Broadcasting (<https://americanarchive.org/>), which was designed to ensure “the long-term preservation and access to GBH’s vast archive of programming and original materials, including raw archival interviews, programs, and transcripts, which are all available through our Open Vault.”

Whatever form it takes, the fact that preservation is part of the news organization’s mission is one of the strongest indicators we found for news organizations that continue to rate news content preservation highly among their priorities.

Are there clear policies on preservation, including ways to monitor and follow-through?

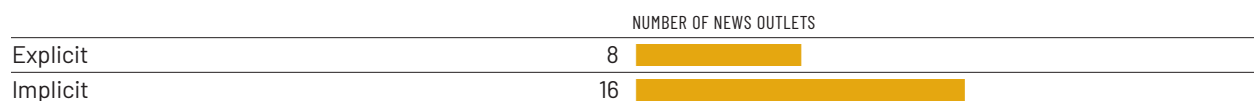
Closely allied with the mission is whether or not a news organization has a written policy regarding news content preservation. This includes clear written guidelines that spell out what content is preserved, how, and who has this responsibility.

Most of the news organizations we interviewed have no preservation policy at all, not written at least. Instead, as the findings above show, most have only generalized expectations that depend partly or entirely on technology and cost factors. Some, however, have clear and written policies, and it was these news organizations that scored highest in our rankings and findings.

Here are some examples of these policies:

- The BBC, for example, has a policy stating that it must preserve all final broadcast news content in English, both radio and TV. But, with a few exceptions, it does not preserve the versions it broadcasts in 37 other languages.
- Policies may specify that all parts of any news story are kept, not just some of the content elements. And that linkages connecting these parts are retained. The Associated Press, for example, saves all photos, graphics and video that were produced for each news story. Its policy also requires them to preserve all transmitted versions of news stories, from the initial breaking news alerts through multiple write-through versions that add information as it’s gathered, to the final version.
- A policy may also describe clear roles and responsibilities for who is assigned to do the work of news content preservation. At the Los Angeles Times, this work is done by a four-member team that is embedded in the newsroom.

Figure 30: Are your preservation selection guidelines implicit or explicit?



Source: Interview responses to Question 2; see Appendix B for full text of questions.

Mission

Are technology architectures aligned to preservation as well as publishing?

Also associated with successful preservation efforts are news organizations whose technology architecture is aligned to the needs of preservation, as well as production of the news. With the manic pace of change in technology this is not an easy thing to accomplish. But this research shows this alignment is essential if long-term content preservation is the goal.

Here are some of the key aspects of this technology alignment that we observed:

- One nearly essential element is whether or not the news organization has a preservation platform independent of publishing systems. Most publishing systems are not built to do the work of preservation. They lack many of the basic functions needed to ensure long-term content is saved, for example lacking the capabilities to manage large sets of metadata to meet preservation needs. Throughout our findings we saw that news organizations with independent systems for preservation, whether full archives or asset management systems doubling as archives, did better in saving digital news content than others. This ability to tailor metadata to your archiving needs is one of the key benefits of having a separate platform. It means you are not limited by the publishing needs of a website, mobile app or other digital channel.
- News organizations with the best preservation record have reliable systems of linkages between many or all of the content elements or content types of each story, along with the ability to track all usages of news content elements, and to track in detail any changes made to news content after it is committed to an archive.
- This also extends to other elements of a news organization's overall technology stack. Those that did better had strong technology fundamentals, including more robust disk storage systems with multiple levels of redundancy and backup to minimize or eliminate data loss, reliable monitoring systems and detailed business continuity plans that are regularly tested and exercised.

Good metadata is essential to good preservation

The importance of metadata cannot be overstated in the equation of preservation. Without strong metadata the very best collections of digital content can become essentially inaccessible. What's needed for good preservation is metadata that is extensive and detailed, tailored to the needs of preserving content for your news operation, and with strong practices to ensure its continuity over the long term.

These are the kinds of conditions we found at the news organizations doing the best job at preservation. Among the attributes we found at these news organizations are careful stewardship of their metadata and evidence of a deep understanding of the critical importance of going to extra lengths in planning a system transition to maintain almost any metadata that has been accumulated with past content. Here's an example.

At the Los Angeles Times, we learned much about their transition to a new publishing platform, Graphene, in recent years. The new ownership viewed this as an essential step in helping re-establish the newspaper's independence after decades of chain media ownership (Times-Mirror and later Tribune).

Mission

Despite the fact that it's been in place for a couple of years, however, the paper's news library and tech staff have not yet transferred the vast archived content built up over several decades to the new Graphene system. That's the plan, but it's not yet been done. The key reason: metadata.

The content is stored in the Times' BRS system, a version of the OpenText platform that was originally put in place in the 1990s, and which itself preserved metadata from an earlier system in the 1980s. Since then, BRS has been accumulating mostly text content. But the key issue is that BRS has large and custom-tailored metadata that expands the value of the content enormously. Its metadata is much more extensive than what's in Graphene at present, six headline fields, for example. The work of mapping metadata from one system to another is now being planned to ensure a data transfer that preserves these fields and others. Along with strong search functions, this kind of approach is one of the most effective ways to ensure the ability to discover and utilize content long into the future.

Finding 15

Track record of preservation matters



Among those doing preservation well we also learned they had a clear track record of focus on preservation, a willingness to invest in it over time, along with strong advocacy for preservation and the ability to build partnerships where needed to support and enhance these efforts. This tended to be associated with the longevity of a news organization.

These factors showed up in the answers to one of our most important questions, whether or not the news organization recognizes a responsibility to the public record of their community, however they define it.

Here's what we found:

- When asked if their company believes their news content should be saved as part of the public record, three out of four said yes, or 18 outlets.
Source: Interview responses to Question 6; see Appendix B for full text of questions.
- Most of the news organizations that said yes were the established, legacy outlets, 17 of the 18.
Source: Interview responses to Question 6.

Figure 31: Does your organization preserve news content for the public record?

Yes	18	
No	6	

Source: Interview responses to Question 6; see Appendix B for full text of questions.

This last point is important. The news organizations that did best were those that had been around a long time, have invested in preservation and maintained focus on content preservation throughout some of the major upheavals taking place in the industry, such as changes of ownership, business model and content publishing channels. Newer news organizations have not yet established a culture of preservation. It's apparent from our interviews that institutional longevity tends to encourage longer-term thinking, including the importance of preserving news content.

Mission

One of the best examples of this that we found is the Baltimore Afro-American. Although it currently relies on a web CMS for saving digital content, it continues to have an archive manager and has maintained a strong and consistent preservation mindset throughout its 128-year history. And it has a record of building institutional partnerships over time to help sustain its service to the community, including one with Morgan State University to house much of its collection. Although the commitment to preservation is not formally stated, it's clear from the newspaper's actions and management, who are descended from the paper's founder, that the stories of their readers' lives and history matter deeply.

The Afro has tapped into this in numerous ways to tell stories important to the Black community in the Baltimore area. One of the best is a recent one, a keepsake book published in 2020 spotlighting the Black Women's Suffrage movement that helped win the right to vote for women 100 years earlier. Titled "To The Front: Black Women and the Vote," this booklet mined the paper's rich text and photo archives to bring to life the moving story of how Baltimore's Black women stepped up in the early 1900s women's suffrage struggle to make sure their voices were heard.³⁸

³⁸ Savannah Wood, "New Magazine Celebrates Local Black Women Suffragists," *Afro* (blog), April 24, 2020, <https://afro.com/new-magazine-celebrates-local-black-women-suffragists/>.



Recommendations

Establishing good practices and systems for digital news preservation can be a daunting challenge, given the difficult financial challenges the industry is facing. But the need to act cannot be postponed to a better day. As the findings in this report demonstrate, digital news content is fragile and endangered, and the problems of unexpectedly losing content are happening every day. So, it's important to take any steps that your newsroom can to begin reducing and eventually resolving this problem.

In this section we recommend ways to begin tackling these issues. We offer a wide range of recommendations from technology systems and workflow changes to industry-wide actions that call for collaboration with other media outlets and sectors such as technology, government and academia. But most importantly, this section begins with measures that can be done right away by individual news organizations at any time, at little or no cost.

Through this research effort, we have come to believe that the potential benefits of preservation are real and significant over time for news organizations in better serving the news and information needs of readers, listeners and viewers in your community and by making the wealth of past content you have created more complete and more readily available than ever.

The industry's financial challenges may mean a slower process to carry out some of the recommendations described here that require significant investment. Our hope is that these recommendations can offer guidelines for more immediate steps, and others to be worked into your planning over time. The key takeaway? Don't wait, begin now.

The recommendations here are grouped into three sections for convenience, based on the degree of difficulty and/or cost:

1: Immediate actions: these are steps you can take now, at little or no cost, to begin the process of ensuring news content is preserved for your news organization and your community.

2: Medium-term actions: the steps outlined here are actions that will likely take longer to accomplish and may involve investments in new or changed technologies, staff or funding through grants or other sources.

3: Industry-wide actions: these are long-term steps that involve more than one newsroom to pursue solutions such as policy changes, institutional partnerships, actions by industry sub-groups or news associations, and some government actions.

Immediate Actions

These are steps any newsroom can take now, with no financial investment

As shown below, many of these are steps you can take in your newsroom without delay. The bottom line for addressing the problems of digital news preservation is the importance of acting. You may not be able to do everything you want to do right away but getting started now is the best path forward toward solving these issues.

Recommendation 1

Create a preservation policy, replace happenstance

We recommend that every news organization institute a written policy for news preservation. This need not be a long document, a page or two will suffice as a start. Outline the goals of the policy, including the content you want to preserve, for whom the content is being preserved, and what purposes you expect it to be used for – internal news research for example. Once that’s in place you may want to add to this, to specify other file formats, who should have access and guidelines on what kinds of changes can be made after content is published, who can do this and what permissions are needed, if any. For example, you should include your policy on unpublished news (see below), as this can directly affect your saved content.

You may also want to involve other parts of the organization beyond the newsroom, especially technology staff, along with those who develop new market opportunities from improved content archives. For technology, our research shows that having a written policy is a major step in the preservation process. If your newsroom is like most of the ones we interviewed, they likely have no preservation guidelines to follow and would welcome them. Once this is done, it’s important to communicate your preservation policy to everyone in the organization and arrange a mechanism to update and re-communicate any changes as this develops and improves over time.

Resources: To help get you started on this, we’ve listed a number of tools, websites, articles, books and other resources you can tap to help better understand preservation issues, and get you started. These include a simple two-page template for preservation policies, and some samples of policies and guidelines for some of the news organizations we interviewed.

- Digital archive policy template, Meta-Archive Cooperative:
https://metaarchive.org/wp-content/uploads/2017/03/ma_dp_policy_template.pdf
- Digital Preservation Step by Step, Orbis Cascade Alliance, a libraries group:
<https://orbiscascadeulc.github.io/digprezsteps/policy.html>
- BBC Archive policy, BBC Royal Charter, Section 69, page 43: “The BBC must make arrangements for the maintenance of an archive, or archives, of films, sound recordings, other recorded material and printed material which is representative of the sound and television programmes and films broadcast or otherwise distributed by the BBC.” available at this URL under “2017 Framework Agreement.”
https://www.bbc.co.uk/bbctrust/governance/regulatory_framework/charter_agreement.html

Immediate

- American Archives of Public Broadcasting, mission statement: “The American Archive of Public Broadcasting seeks to preserve and make accessible significant historical content created by public media, and to coordinate a national effort to save at-risk public media before its content is lost to posterity.” <https://americanarchive.org/about-the-american-archive/vision-and-mission>
- New York Public Radio preservation policy: <https://www.wnyc.org/collection/policy/>
- News archive articles at the Nieman Journalism Lab website: https://www.niemanlab.org/?s=news+archive&post_type=post
- History of the Vanderbilt University Television News Archive (PDF, no fee): <https://content.iospress.com/articles/information-services-and-use/isu200085>

Recommendation 2

Tap someone to be responsible for preservation

It’s important to tap someone in your newsroom to begin taking responsibility for content preservation. Most likely your newsroom no longer has a news librarian, or perhaps never did. Start the process now by assigning someone in your newsroom responsibility for beginning the process to ensure long-term preservation of your digital news content. Even if this is only part of their role to begin with, don’t wait. There’s a long road ahead that will take many months if not years, but this step is an essential prerequisite, and it cannot begin too soon. It won’t happen at all unless there’s someone to take ownership.

Ideally, it would be best to hire media archivists for this work, who could assist in creating digital news content preservation plans, manage metadata and workflow and much more. Our study shows that organizations with dedicated archival staff, even part time, are doing better in preserving digital news content.

Resources: Here are references and a reading list for that person once identified to help get them oriented and up to speed on this issue:

- CJR/Tow Center major article: A Public Record at Risk: The Dire State of News Archiving in the Digital Age: https://www.cjr.org/tow_center_reports/the-dire-state-of-news-archiving-in-the-digital-age.php
- “Missing Links: The Digital News Preservation Discontinuity,” Paper by Dorothy Carner, Edward McCain, and Frederick Zarndt, presented at the 80th IFLA General Conference and Assembly, August 13–15, 2014, Lyon, France, Dorothy Carner, Edward McCain, and Frederick Zarndt: <https://www.rjionline.org/stories/conference-paper-missing-links-the-digital-news-preservation> or https://www.ifla.org/files/assets/newspapers/Geneva_2014/s6-carner-en.pdf
- Levels of Preservation, NDSA, DLF: <https://ndsa.org/activities/levels-of-digital-preservation/>
- When an online news outlet goes out of business, its archives can disappear as well. The new battle over journalism’s digital legacy. CJR article 2018: https://www.cjr.org/special_report/microfilm-newspapers-media-digital.php
- American Archive of Public Broadcasting, collab between Library of Congress and WGBH (now GBH): <https://americanarchive.org/>

Immediate

- Archivemática, open-source digital content preservation software: <https://www.archivematica.org/en/>
- Preservica, cloud-based digital content repository service used by industry, libraries and academia, for monthly fee: <https://preservica.com/>
- DocumentingTheNow, a non-profit organization working to preserve social media, especially crowd-sourced material from public events: <https://www.docnow.io/>
- OpenArchive, a non-profit working to preserve mobile content using tools this organization has developed: <https://open-archive.org/>

Recommendation 3

Review metadata to ensure you have what's needed

Knowing and understanding your metadata is a critical step in improving content preservation for your unique, original content. A review will show what steps you can take to ensure you have clear metadata on content objects, including origin and ownership. This may call for changes in your existing technology, especially configuration changes to add or modify metadata fields needed with each type of content.

For example, one trend we noticed is that news content assignment functions are commonly done off-platform, not in any CMS but in tools such as Google Docs or Google Sheets, where it gets overwritten or discarded over time. But this information is a gold mine for preservation, because it shows the intentions, the planning for news that help identify the purpose, time and date and individuals involved in creating that story, photo or video.

If possible, work with your tech vendors to find a way to feed that info into metadata fields for your content, so it's always there, helping to ensure that you know where that photo or video came from, when it was shot, who did it, what the story was that generated that assignment. This is priceless info that most news organizations don't take advantage of. And if possible, be sure to do this also with data that's not in your production or publishing systems: the still photo outtakes, the original raw video and audio files. This one step can go a very long way toward ensuring that you know the full story behind all the content saved over time.

As part of this, check to make sure your CMS is utilizing one of the most common metadata standards in the industry: the IPTC photo and video metadata that's created by cameras from the moment images are captured, including timestamps and geolocation data. Typically, the IPTC fields are supplemented with captions and other information as image data goes through a news publishing workflow of editing and packaging before it goes live.

But we also learned that some web CMS systems either do not read the IPTC fields when imported or accept only a fraction of what is there. If that is your only system, this could mean your primary photo file will be missing critical metadata. It's OK if this data is stripped out in rendering the html for the public web page. But take the time to understand this process for your newsroom, to make sure that somewhere along the line the original image files with full metadata are stored.

See below for further information on this and other metadata standards.

Immediate

Resources: Here are some resources to help:

- The IPTC, a non-profit global standards body for the news media, created the so-called IPTC info on most modern still photos that used the JPEG compression standard. Here's their website: <https://iptc.org/>
- Video codec guide with details of common compression standards used with everyday video files, from the MDN WebDocs site run by the Mozilla Foundation, a non-profit that was one of the pioneers of Internet web browsers: https://developer.mozilla.org/en-US/docs/Web/Media/Formats/Video_codecs
- Levels of Preservation, NDSA, DLF: <https://ndsa.org/activities/levels-of-digital-preservation/>
- National Archives and Records Administration, Digital Preservation Strategy: <https://www.archives.gov/preservation/electronic-records/digital-preservation-strategy>
- PREMIS, Wikipedia page on PREMIS, Preservation Metadata Implementation Strategies website: https://en.wikipedia.org/wiki/Preservation_Metadata:_Implementation_Strategies
- PREMIS LOC page: <https://www.loc.gov/standards/premis/>

Recommendation 4

Establish a plan for handling “unpublishing” requests

Consider your newsroom’s plans for managing requests to remove, de-index or otherwise alter published content, a process sometimes called unpublishing news. With growing concerns about Internet privacy, the proliferation of reputation management services, and emerging legal restrictions such as the European Union’s data privacy legislation, this is a fairly urgent and rapidly growing area of concern for every news organization.

It’s not a question of whether this is happening. Chances are it has already happened. It’s typical that many people in newsroom or technology staffs could potentially unpublish without managers’ knowledge. The only way to control it is to have a plan and clearly communicate it to your entire organization. Communicating your policy publicly will support transparency and accountability.

Key elements of an unpublishing policy (Source: largely based on work by RJI Fellow Deborah Dwyer, see below) should cover these points:

- Map out how you want to deal with requests to unpublish news content. How is it done now? Is there a process? It’s critical that unpublishing not be left to the individual judgement of whomever a reader or reputation management agency happens to contact.
- Keep a record of requests, decisions and outcomes to remain accountable. Tracking is important to ensure policies are enacted equitably and you have the information to continuously improve your policy.
- Establish guidelines for what kinds of unpublishing requests will be considered, and just as importantly, what kinds of requests will NOT be considered.

Immediate

- Set clear expectations. Often, newsrooms cannot totally unpublish all instances of news content. For example, with print editions, their e-edition cousins and online research services, there may be significant limits on what can be changed retroactively.
- Steer clear of terms such as the “right to be forgotten,” which convey a promise to the public that most likely cannot be met.
- Set crystal clear guidelines on how the unpublishing requests will be managed and carried out. Will you convene a committee to review requests? Will the original story in your web CMS be edited to remove something? Will a whole story be set to be “de-indexed” by search engines? Similarly, make clear who will take the actual steps of unpublishing, what systems and tools they will use, and who is responsible for keeping track.
- Unpublishing is directly connected to what you choose to publish in the first place. Analyze your reporting philosophy, especially concerning minor crime coverage such as routinely publishing mugshots when no charges have yet been filed.

Resources: Here are some links to websites, projects, stories and background information on this issue:

- Original 2009 study on unpublishing issue by Kathy English, Toronto Star:
<https://unpublishingthenews.com/2009/01/21/2009-industry-report-the-long-tail-of-publishing/>
- Overview and the first of several recent very insightful articles by Deborah Dwyer as part of her Reynolds Journalism Institute Fellowship research on the unpublishing issue, 2020-2021, following up on her PhD thesis work at the University of North Carolina-Chapel Hill:
<https://www.rjionline.org/stories/facing-the-pressure-to-unpublish>
All stories by Dwyer on this subject: <https://www.rjionline.org/account/profile/4394>
- The Boston Globe story on their new Fresh Start unpublishing initiative:
<https://www.bostonglobe.com/2021/01/22/metro/boston-globe-launches-fresh-start-initiative-people-can-apply-update-or-anonymize-coverage-them-thats-online/>
- The New York Times story 1/23/21 on The Boston Globe’s Fresh Start program:
https://www.washingtonpost.com/lifestyle/media/old-arrest-boston-globe-fresh-start/2021/01/22/122cbd0c-5cd1-11eb-b8bd-ee36b1cd18bf_story.html
- Unpublishing news and resources <http://www.unpublishingthenews.com>

Immediate

Recommendation 5

Clarify content ownership in CMS, other systems

Clarify ownership and licensing for the content you publish now. While ownership information on past content may be difficult to establish or correct after-the-fact, there is no reason to wait to address any ownership uncertainties around content your newsroom and freelancers are currently creating and publishing.

We suggest you do some research on your company's existing policies, contracts and service agreements regarding any of the content you publish at present. For example, is there a clear expectation on who owns still photos or videos shot by your staff photographers? How about rights to news stories, graphics, audio and video from interviews or database journalism projects? There's no single answer to these questions, the key thing that matters is to ensure clarity for all involved.

This is especially an issue with content created by freelance news staff. It's important to be aware of the implications of the *Tasini v The New York Times* case of 2001, for example. It's critical that you take steps to ensure your newsroom is following the guidelines that came out of that case. And if you already have clear policy or guidelines, it may still be worth updating to ensure it covers today's digital realities, including all content types (reporters who shoot video, for example), social media and others.

Resources: Here are some links to help catch up on the details relating to the *Tasini* case:

- *Tasini* case on news content from freelancers, Wikipedia summary:
https://en.wikipedia.org/wiki/New_York_Times_Co._v._Tasini
- *Tasini* case excerpts from the Legal Information Institute at Cornell University:
<https://www.law.cornell.edu/supct/html/00-201.ZO.html>
- 2010 The New York Times story on developments in *Tasini* case:
<https://www.nytimes.com/2010/03/03/business/media/03bizcourt.html>

Recommendation 6

Run a self-assessment test on metadata, ownership, workflows

To check whether you may have issues to address related to the factors above, we suggest you conduct some simple tests, a self-assessment on how well your news organization is doing in ensuring preservation, identification and ownership of the unique content your news organization created.

Here are some sample questions to ask, and additional resources on this process:

- Self-test: Here's a test you can do any day to clarify what's going on in your news workflows and systems. Check whether you can identify, locate and retrieve the original version of every still photo and video your newsroom published today. Not multiple copies that may show up in your CMS, the resized and resampled versions, but the original, the high-resolution originals. Can you do that? And if so, does that content have sufficient metadata to clearly show your company owns that image, that video?

Immediate

- Going further on this, does your CMS include fields that clearly show whether content is staff-generated or comes from a free-lance reporter or photographer, or some other third-party source? If those fields are not available, consider taking steps to get them added, visible and editable in your content management systems, and add steps in the news workflow to ensure the fields are populated.
- Check to find out what news content is being saved, for how long, and what content is not being saved or is purged after a period of time. In today's newsrooms this is often a function left to technology staff, who in many cases have little to no guidelines to follow. In the absence of any other guidance, tech staff may default to deleting content or purging version stacks as needed simply to maintain system performance, a practice that can inadvertently work against good preservation practices.

Finding a way to preserve copies of the original full resolution, full color-depth, full-size news photographs and news videos, with full original metadata, is one of the best steps you can take for long-term preservation of this critical, irreplaceable content. It's great if you can get these files into a system with a database and redundant storage and fully indexed content. But even a simple system of drives can work to get this process started. Note: it's OK to save images in JPEG format rather than camera-raw image files, which are far larger, as JPEG is an accepted standard.

Resources: Here are some links to help you further assess your news content preservation practices:

- SPOT model for assessing risks to saved digital content, from the Corporation for National Research Initiatives, Reston, VA. SPOT model: <http://www.dlib.org/dlib/september12/vermaaten/09vermaaten.html>
- "Sustainable Economics for a Digital Planet: Ensuring Long-Term Access to Digital Information" was produced by the Blue Ribbon Task Force on Sustainable Digital Preservation and access in 2010. The information in it still holds up well as an overview of key digital preservation issues, including a section on "Commercially Owned Cultural Content." https://www.cs.rpi.edu/~bermaf/BRTF_Final_Report.pdf
- If you want to get deeply into preservation standards at some point, look into the high-level preservation standards framework and audit process, ISO 16363, or Primary Trustworthy Digital Repository Authorisation Body Ltd.: <http://www.iso16363.org/>
- National Archives Digital Preservation Strategy: <https://www.archives.gov/preservation/electronic-records/digital-preservation-strategy>
- For data journalists, consider the work of the Software Preservation network, a non-profit grant-funded organization which works to preserve software such as the off-platform code used in data journalism: <https://www.softwarepreservationnetwork.org/>
- Open Archival Information System (OAIS) Reference Model: An Introductory Guide (Second Edition) Digital Preservation Commission, 2014. <https://www.dpconline.org/docs/technology-watch-reports/1359-dpctw14-02/file>

Medium-term Actions

These will help your newsroom solve preservation problems for the long term

Recommendation 7

Services can help if web archiving is your goal

There is help out there for some news preservation efforts. It's possible that the Internet Archive or other similar tools and services could be tapped to preserve content, at least temporarily while you improve your own practices. Here's a summary of what some of these have to offer.

The Internet Archive is a privately funded and grant-supported organization based in San Francisco, which has a mission to provide "universal access to all knowledge." This includes news content, and the archive now estimates they have saved more than 500 billion news web pages. Anyone can upload content to the archive at no cost through its Save Page Now function (<https://web.archive.org/save/>). But of possibly greater interest to news organizations are several services they operate that could help fill gaps or supplement in-house preservation activity.

These include the Wayback Machine, the portal for accessing web pages from more than 200 million websites (<https://web.archive.org/>), and Archive-It, a subscription-based service the Internet Archive offers that will automatically save web news pages on a programmed schedule. The Archive-It service is the one most likely of interest to newsrooms that need help. While the Internet Archive may already be saving some of your news web pages into the Wayback Machine, this would only cover a small fraction of most news sites, and at a very low frequency, once a month, for example.

Other tools and services include an easy-to-use personalized clipping service called Save My News set up by Ben Welsh, a data journalist at the Los Angeles Times. Several other open-source tools are available for free from Welsh's PastPages.org site.

Also available is Conifer, a non-profit and grant-supported service that preserves an interactive and dynamic copy of web pages. This is descendent of a web archiving service started some years ago called WebRecorder, managed by Rhizome.

Resources: Here are some links on various tools and services that may be helpful in web and other content preservation:

- Wayback Machine by the Internet Archive, launched in 1996 and containing more than 150+ billion web captures: <https://web.archive.org/>
- Archive-It, a subscription service for preserving web pages that dates back to 2006. As a paid service, it can be customized to capture specific web pages on a regular schedule: <https://archive-it.org/>
- Past Pages, by Ben Welsh: <http://www.pastpages.org/>
- WebRecorder, Rhizome: <https://conifer.rhizome.org>
- NetarchiveSuite, web archiving software developed and maintained by The Royal Danish Library, and used for the Copenhagen, State and University libraries in Denmark, plus other government archives such as the National Library of France and of Austria: <https://sbforge.org/display/NAS/NetarchiveSuite>

There are also commercial paid services for web capture such as these:

- Authory, which helps journalists track their articles and automatically backs them up, at \$8 per month if paid yearly: <https://authory.com/>
- MirrorWeb, a UK company that offers this service commercially: <https://www.mirrorweb.com>

Medium-term

Recommendation 8

Talk with your tech vendors, they likely have improvements, new solutions

As explained in our findings, talks with top technology providers showed their hopes that customers would take better advantage of the full capability of their systems. This includes new tools and subsystems recently developed. Given the rushed and overworked nature of news organizations these days, it's possible, even likely that the companies providing your publishing systems have loads of new capabilities you are not using and may not even be aware of.

This action is one we recommend especially for newsroom leadership: Take the time to find out if there is helpful new or existing functionality available from the publishing systems you already own. There's likely to be at least some capabilities you're not taking advantage of. The tech providers our team talked with told us they are constantly working to improve their systems to keep up with the manic pace of change in this industry. It's one of their key survival strategies: evolve to meet customers' needs or risk being taken over by a competitor. The result is typically a rapid pace of software development that could help solve many of your problems.

Some may be available at little to no cost. Contact your tech staff to set up meetings with the tech companies to find this out. Work with your IT colleagues. Good IT departments will welcome your involvement as this helps them do a better job. Most are as overwhelmed as your newsroom, and they may not fully understand what your needs are today.

Based on the results of your self-tests above, or more in-depth analysis of preservation needs, you will likely see a need to modify technology configurations and workflows to add or change metadata applied to your news content. This may take some staff time and additional costs to map out and implement these changes with your tech providers. In the long term these kinds of improvements will be worthwhile. For example, if your news organization relies primarily on its web CMS to save original photo and video content, examine whether workflows could be changed to make these the highest resolution possible, as close as the system can handle to the original files.

We found that systems such as Stibo DX, Brightspot, MerlinOne and SCC provide highly sophisticated content and DAM tools and capabilities, particularly for creation and management of metadata, which are key to the preservation process. Yet those tools and capabilities are not always fully utilized by newsrooms.

One major caution: don't be easily sold on the benefits of moving to a different system, a new CMS. Our research shows that the small number of remaining tech providers all work hard to remain competitive, and possibly have what you need already. In addition, as with the media organizations themselves, where we learned that recently established news outlets have not yet built up good preservation practices, we were similarly surprised to learn that, in general, newer tech providers appear to have less functionality related to preservation than more established providers.

Lastly, as we outlined in our Findings, remember the dangers that occur when you switch from one CMS to another, in particular the content elements that frequently get lost in the conversion process, primarily metadata. System transitions are extremely difficult, problems are very common, successes without some lost content are rare to nonexistent. So, while a new system may ultimately be the right move, we strongly recommend caution.

Medium-term

Resources: To help you get started, here are the website home pages and development release pages for the vendors we talked with and some other top providers:

- TownNews site: <https://TownNews/>
 - Development releases: https://townnews.com/releases/software_releases/
- WordPress: <https://wordpress.com/>
 - WordPress documentation: <https://developer.wordpress.com/docs/>
- CUE Publishing: <https://www.cuepublishing.com/>
 - CUE DAM, formerly Digital Collections, is an integrated DAM system: <https://www.cuepublishing.com/solutions/2018-12-11/Digital-Asset-Management-1631.html>
 - CUE development releases: <https://www.cuepublishing.com/support/>
- Brightspot: <https://www.brightspot.com/>
 - Brightspot MediaDesk, an integrated DAM system: <https://www.brightspot.com/brightspot-dam>
 - Brightspot development releases: <https://docs.brightspot.com/4.2/en/releases.html>
- SCC Media Server: <http://www.sccmediaserver.com/>
 - SCC development news: <http://www.sccmediaserver.com/scc-news/index.html>
- MerlinOne: <https://merlinone.com/>
 - MerlinOne resources page: <https://merlinone.com/resources/>
- Other popular CMS providers:
 - Arc XP: <https://www.arcxp.com/>
 - Naviga (formerly NewsCycleSolutions): <https://www.navigaglobal.com/>
 - Eidos Media: <https://www.eidosmedia.com/platforms/>

Recommendation 9

For best preservation outcomes, consider asset or archive system

There's one major technology change we recommend that represents the ideal goal for news preservation: in order to properly manage your newsroom's digital assets, we recommend that you acquire a separate archive system that's independent of your publishing CMS technologies.

This will seem like a stretch for many news organizations. With revenue or funding so tight, this may feel like an unattainable goal, too remote. But after studying the forces and trends influencing the news industry for a year and half, it's clear to us that this step is the only way a news organization can gain full and permanent control over the priceless content assets they have created over the years at considerable cost and continue to create every day.

Medium-term

Many trends in technology, media markets and industry upheavals converge on this recommendation. A central one is this: every indication shows that the explosion of digital news and information channels is only going to continue and is likely to accelerate. More products of all kinds, more news websites, more interest-targeted sites and services and data feeds, more social media platforms, more digital devices and internet-enabled homes, appliances and vehicles...and the list goes on.

The more that digital channels grow, the more difficult it is, based on this research, to manage them all from a web Content Management System. The web, after all, is a publishing channel of its own, with specific needs for content in specific forms. As the CMS evolves into a multi-channel digital content distribution platform in addition to its original role of web publishing the system's conflicts with content preservation can grow.

The reasons for this are numerous: preservation systems need preservation metadata, for example, such as trackers to show how often, when and in which channels a content object has been published. Preservation systems also need to know who owns the content, what the usage terms are if third-party, descriptive info that's tailored to preservation and reuse, information on the origin of the content to ensure it's the original, and much more. These serve little to no purpose in a publishing system, only clogging a CMS database and slowing performance. (For more details see the Technology Findings section.)

What systems are we talking about? Here are some examples of systems that are designed for digital content archiving or digital asset management (Note: this is a sample, not an exhaustive list): Artemis, DigitalCollections DCX (now CUE DAM), SCC Media server, MerlinOne, Primestream, Scisys. These are systems that can perform the functions you need for preserving digital content and managing assets. While they vary significantly in specific capabilities from one to another, they are designed for the same general purpose, to preserve content over the long term, and provide the kinds of extensive metadata you need to ensure your ownership and know whatever you need to know about the origin and details of the content objects.

In addition, some of these kinds of systems are capable of integrating tightly into your publishing workflow. This is the approach we recommend: to integrate these directly into the publishing workflow in as automated a fashion as possible.

Here are some examples of how that could work, borrowed from best practices already in place at the media companies we interviewed:

- **Planning info:** Use this system to store news content planning info, or news budgets, and tie them to content objects as they are ingested. This provides highly useful data that holds key details on the origin of content.
- **Unique IDs:** Devise workflows that ensure unique IDs are assigned to every content object going to a publishing system and archive/DAM. This can be done either by routing first to the archive or DAM before going to a web CMS, for example, as a number of newsrooms already do. Or by using a unique identifier after import into a web CMS to send content to the archive/DAM, where it will be stored only if new. This will help ensure content is not duplicated in the preservation system, that there is only one original.

Medium-term

- **Usage data:** Create a mechanism after publication that sends info to the archive/DAM on each publishing instance. Route this usage data so it ties to every content object that was published, showing a full history of the original publication, and the times it was reused.
- **Automate metadata:** Automate as much descriptive metadata as possible. Capture all available info on authors, such as whether they are a sports or business reporter, videographer, etc. Wherever possible keep data on channel usages and topics/ labels/ navigation nodes in a website and other destinations. Use data mining engines to automatically apply descriptive terms; but don't rely only on this. The best descriptive metadata will always be that which thinking humans apply. An editor, for example, may know a given story is a profile of a news subject rather than a breaking news story, a useful tag if that's what you are looking for, one that some automated or artificial intelligence systems will have trouble discerning.
- **Scheduled transfer:** One challenge cited by a number of news organizations is the uncertainty of knowing when digital content is complete, since the non-stop news cycle can extend a story beyond a few hours or a day. One solution we learned about is an automated process used by a large broadcaster in which content is not sent to the archive until a week after it is published, at which point it is purged from the publishing system but can be recreated at any time. This allows nearly all errors and additions to be completed before content is sent to the archive, where changes are very closely tracked and, in some cases, require high-level permissions, after that ingest point. It also frees publishing systems to purge version stacks and other unneeded content. So long as package linkages are retained, connected to the archive, most web or digital content can be reassembled and rendered on demand.

As an offshoot of this, if you already have a DAM system that you're using for only one or two content types, still photos and video, for example, consider adapting it to cover more content types, or all content so it works as an archive. By and large the systems listed above are capable of handling all content types.

The bottom line is this: your web publishing system already has too much to do and the list is growing. By planning toward a system architecture that relocates preservation functions to a more suitable system, this will free up publishing systems to do their most important tasks, and not do double duty for preservation as well, a role they are not designed to perform and do not do well.

Lastly, the key advantage of having a separate archive or content repository is this: no matter how digital publishing channels change over time, coming and going with their respective tech platforms as news consuming patterns change, having a solid, permanent repository of all your content is the best way to position your news organization to serve any market or channel that may emerge.

Long-term Actions for the Industry

These may be the hardest but could make the most difference

Recommendation 10

Collaborate to define best methods, structure for saving news

One of the central goals of this research was to begin laying out a roadmap for news content preservation that could help the news industry rebuild its critical role as a primary source of unique digital news and information for each community, students, researchers in academia and other levels of society.

Our intent was to find out how digital news content is being preserved now and present what we found, along with best practices and possible solutions to address the issues. As one early step in this process we believe it would serve all involved to establish a shared definition of what news content is most important to preserve for communities and for larger social needs, and how this could be done.

This means a clear and achievable definition of news archiving. As evident in our research, concepts of preservation vary widely from one news organization to another, and depend on resources, business model, ownership and other factors. Answers to the question, “What should be preserved and how,” along with clarity on why certain materials are worthy of long-term preservation, would help establish a baseline for news organizations to use as a goal, even if that goal takes years to achieve. Here are some examples of what such a definition or policy might include:

- **Text:** What text content to save, at what point in the process from initial idea to final published version? Should this include the news planning or news budget data that describes the original intent? Should it include reader comments? What about audience metrics on how well that story performed? Which of these matters and how much?
- **Visual content:** What visual content elements matter most, the final published photographs or that plus the full outtakes as well? The final edited video or audio file? Or that plus the raw footage shot, the raw audio tape? Why? Who would these serve?
- **Data journalism:** What’s most important to preserve on data journalism projects? The final map, table or active graphic embedded in a web page? What about the original data sources used to build that map or table?
- **Social media:** One of the most difficult puzzles of all is the question of how and what to preserve in social media, a huge factor in modern news publishing and broadcasting, responsible for much of the audience traffic news content currently receives. Should the industry attempt to preserve every post, just a link to the post, or leave this to the platforms, to Twitter and Facebook, Instagram and others?
- **Business model:** And possibly the most difficult question of all, and most important to a news industry struggling daily with survival: how can all of this be done in a way that serves our communities while also compensating the news organizations that create the news in the first place? Most of the news media is still operating on a for-profit business model, one they argue ensures the greatest degree of independence, the highest incentives to serve the news-consuming public, one that, on balance over time, has served a proper role as a check on other parts of society. Whatever shared definition and process is created must take this key factor into account.

Long-term

In addition, as part of this process there could be an important role for establishing a standard metadata model to be used in news archiving, working perhaps from some of the existing models already developed by digital content preservation experts. In addition to the general benefits of such a standard, it could also help make it easier for various archive systems to communicate with each other to find and share news content as needed, on demand, potentially expanding these markets to more news organizations and to researchers and the public.

Accomplishing these objectives calls for a collaborative effort among stakeholders, including news media representatives, public libraries, universities and research institutions, business interests and others interested in the outcome. This could be done through existing organizations working together. It's beyond the scope of this report to recommend a specific process. But the organizations behind this research project would likely be interested in such an endeavor. The primary purpose of this entry is to call attention to the need to come to agreement, to establish a baseline, a consensus on what matters most to preserve and how this can be done in a way that will serve not only industry but the greater public and scholarly audience as well.

Recommendation 11

Advocate with tech companies for better preservation capabilities

While our research showed that there are some promising developments in the media technology field regarding content preservation, the larger finding still dominates, that most publishing systems, and the common ways they are used, do not take long-term preservation into account. We believe this can change through concerted action by the industry in advocating for it with technology providers. We believe it should change.

Today's CMS platforms go a long way toward basic preservation functions, but much more is needed, including features to ensure long-term integrity of original content objects and more granular metadata on content source, authorship, publication history, rights and more.

In addition, collective advocacy could be considered industry-wide, potentially including other technology players such as Amazon AWS and Google, along with industry-specific services provided by companies such as NewsBank, Newspapers.com and ProQuest to assist with needs for more comprehensive digital news preservation.

Despite the difficulties, we believe these kinds of changes should not be left to another day. Content is created every day and lost every day. So, there is urgency needed to encourage technology providers to create or enhance solutions that work.

This can be done through individual and collaborative efforts by news organizations in their normal interactions with technology providers and also through industry associations. User groups for specific technology companies are also a good forum for raising these kinds of issues.

Long-term

Recommendation 12

Advocate for industry sharing on benefits for digital archiving

In addition to establishing how newsrooms vary widely in their practices for digital news preservation, this research also offered insights into the significant degree to which preservation benefits vary from one news organization to another.

These benefits can range across the spectrum from financial, in the form of new revenue, to less tangible but nonetheless real benefits in engagement with readers and viewers, traffic to news content and overall service to the community. In other words, some news organizations are taking better advantage of their past content than others. And it seems fair to believe that benefits should be more widespread, more common.

One reason is a lack of awareness across the industry of what is possible in taking better advantage of past content, especially in the digital age, and what kind of value these could bring, especially in new revenue. Because most news companies consider revenue information private, this kind of data is not well known. But there are ways this could be done without jeopardizing private and potentially competitive information.

To resolve this we recommend the industry begin a process to better share information on the value of past news content. This would include information on the kinds of services and workflows now in use to tap into past content, but also a way to share how well this kind of content works to drive additional traffic to news websites and other channels. And to include details, shared anonymously as needed, on how these services contribute financially to the bottom line.

It's our belief that better leveraging of past content is a significant untapped potential source of revenue. What's needed is the data to prove it.

One recent study on this topic is a related research effort by RJI Fellow Neil Mara (also a member of this project team), that serves as a companion to this report.³⁹ In it, Mara shares many useful approaches by newsrooms across the country to share past news content in ways that help enlighten today's news for readers and serve their communities. One key missing element of this is information on audience metrics or financial benefits from these endeavors, which none of the news organizations were prepared to share.

If the industry can find a way to better share these practices and their benefits, including financial ones, this can help prove the advantages for news companies of the value of investing in news preservation, for better systems and workflows, staffing and services to readers, viewers and listeners.

³⁹ Neil Mara, "News Archives: The Untapped Resource" (Columbia, MO: Donald W. Reynolds Journalism Institute, University of Missouri, forthcoming), <https://hdl.handle.net/10355/80861>.

Long-term

Recommendation 13

Work with others to build long-term institutional partnerships for preservation

Looking ahead to the long-term future of news content as a key part of the public record, we believe one of the primary pillars is the need to establish substantial partnerships with institutions in each community, along with regional and national institutions that also have a strong interest in the success of these efforts.

Our research showed that this is one of the most important factors in preservation. News organizations with strong institutional partners also had the most effective, sustained news preservation activities.

We came across many examples of this in our research. Newspapers, for example, that worked out agreements with nearby universities to house large collections of analog and digital news content, including The Boston Globe with Northeastern University, the Baltimore Afro-American with Morgan State University and CNN's relationship with the Vanderbilt University TV News Archive, which has been recording CNN's programs off-the-air continuously since 1994.

We also learned about partnerships involving other community institutions, such as the collection of millions of images from the Chicago Sun-Times at the Chicago History Museum.⁴⁰ And the announcement last year that the Ebony and Jet Magazine photo archive, a priceless collection of 20th Century images of Black Americans, was purchased for the Smithsonian National Museum of African American History and Culture and the Getty Research Institute, funded by the Mellon Foundation.⁴¹

Libraries are also a major player in these efforts. We learned about numerous examples such as the collaboration between WGBH and the Library of Congress that created the American Archive of Public Broadcasting, a landmark effort to make public broadcasting programs from across the country available to the public through their website. And the Center for Research Libraries' many efforts, ranging from its enormous collection of print news content for loan to major universities, to the national academic site licenses negotiated by CRL for online access by U.S. colleges and universities to content from The New York Times, The Wall Street Journal, and The Washington Post.

Local public libraries offer a key preservation opportunity for news organizations across the country. There are already many examples of news organizations working with their local library to house and preserve news content. One well-known example is that of the Rocky Mountain News, which donated its entire collection to the Denver Public Library when the newspaper closed in 2009.^{42,43} There's no need to wait for such moments, however. One idea is to make such arrangements in advance of any crisis, to set up a kind of "estate plan," with a local library or university to take over and preserve content whenever such steps might be needed.

40 Millions of Moments: The Chicago Sun-Times Photo Collection," Chicago History Museum, accessed March 5, 2021, <https://www.chicagohistory.org/exhibition/millions-of-moments-the-chicago-sun-times-photo-collection/>.

41 The National Museum of African American History and Culture Will Acquire a Significant Portion of the Archive of Ebony and Jet Magazines," National Museum of African American History and Culture, July 25, 2019, <https://nmaahc.si.edu/about/news/national-museum-african-american-history-and-culture-will-acquire-significant-portion>.

42 "The Rocky Mountain News at the Denver Public Library," Denver Public Library, October 7, 2015, <https://history.denverlibrary.org/rocky-mountain-news-denver-public-library>.

43 Edward McCain, "How the Denver Public Library Ended Up Owning the Rocky Mountain News Archive," Donald W. Reynolds Journalism Institute, December 17, 2014, <https://www.rjionline.org/stories/how-the-denver-public-library-ended-up-owning-the-rocky-mountain-news>.

Long-term

Whatever the need, we believe local libraries are a natural fit for partnerships with news organizations, and there's a strong future in such efforts. Libraries are located in almost every community in the country with a news organization. They also have physical branches in many neighborhoods and virtual services available to all. And they have a unique role as respected institutions that offer a central forum for community information resources.

It's important, however, for media and other stakeholders to recognize their role in advocating for resources that libraries need to properly manage digital news content and make it available to the public, rather than merely house it.

The final leg of the preservation structure we envision are improvements in the role of government in the news archiving process. There are numerous models of public-private partnerships with governments to emulate and expand. One example we learned about is the Netherlands Institute for Sound and Vision, near Amsterdam, which preserves news content from TV and radio stations across the country and provides useful services to those broadcasters. In the United Kingdom, the BBC's dominant role in TV and radio is backed up by a strong news archive operation. The BBC Archive preserves content from all BBC TV and radio channels, going back a century to early radio news bulletins in the 1920s. The group works to preserve modern digital news content including raw video and audio files along with finished programs, distributed on broadcast channels and its website, one of the highest-traffic news websites in the world. The BBC is the only news organization we interviewed that operates a service to capture and preserve its web pages. This material is all made available to the public, a key part of its mission, through the British Library and the British Film Institute.

The U.S. Library of Congress provides access to its collection of news content in multiple forms, with some going back to the seventeenth century but most dating after 1874, when the first mandatory deposit laws were enacted.⁴⁴ The library runs many other preservation programs, including the National Audio-Visual Conservation Center in Culpepper, Virginia, a massive facility with more than 90 miles of shelving for millions of film, television and audio programs in analog and digital form.

One of the newest Library of Congress preservation programs we learned about is the effort to preserve pages of online news websites. This program, done in cooperation with the Internet Archive, captures selected sets of news content from news websites chosen by a team of nearly 300 recommending officers.

But we also learned that the Library faces significant funding issues. We believe it would serve the news industry well for members to advocate for the much-needed resources to enable the Library of Congress to expand these and other programs to better serve as a resource for the public and scholarly needs for authentic, persistent digital news content.

⁴⁴ Hansen and Paul, *Future-Proofing the News*, 27.

Long-term

Recommendation 14

Work with universities to create or expand programs, training for digital archivists

One surprising finding in this study was that news organizations that told us they wanted to hire additional archival staff, but had difficulty finding qualified candidates. This can be addressed.

We recommend that universities consider creating degree programs that focus on the specific skills needed to handle the rapidly changing work of news archiving in the digital era, or add to existing programs to cover these needs.

Digital publishing has changed so rapidly that hiring managers are having difficulty finding candidates who are fully digitally literate. This is especially difficult in areas such as video data and publishing processes, social media, data journalism and ways in which news tech functions operate to influence preservation, such as the Amazon web services or similar functions now often employed for data manipulation for news content.

It's also important that candidates be able to function in broader roles that blend many of these skills, involving not just archiving but data analysis, research, training and strategic functions for technology management.

"Some of it is the creative use of technology, thinking on the fly. How do we do this? What can I come up with? What can I use Python for to write a script that will automate some of these processes?" said one manager in an interview. "And I'm not sure that anyone in the news industry is thinking about these things. I'm having a difficult time finding anyone who understands the scope. So, I think a curriculum would be fantastic, but also an applied sense ... it has to be applied to real work, to real projects in real time and challenges we're currently facing. I'm finding incredible gaps in graduate education right now in those areas."

One resource in this area may be graduate degree programs in Information Science at colleges and universities. Many School of Information programs now offer advanced programs including master's degrees or graduate certificates in digital preservation or digital curation.



Digital News as Historic Record

From the Stone Age onward, even the most rudimentary attempts by humans to record their experiences have proven that there is an inevitable trade-off between the longevity of the medium used and the ability to share it with others. For example, we know that pictures carved on rock during the Stone Age often last tens of thousands of years, but obscured by their remote locations, relatively few people will ever bear witness to them. As time passed and language emerged and progressed, clay tablets enabled a much easier inscription process than carving on stone. The fact that tablets were portable made information easier to record and share, but over time they were also more subject to breakage and loss than pictures on rock.

Technology progressed and people found writing with ink on papyrus or paper to be faster and less laborious than pressing a stylus into clay. Lighter and more flexible than clods of earth, paper could be rolled into scrolls or cut into pages for books, which opened the door to wider distribution of knowledge. The downside was that paper could burn easily and was subject to any number of other threats including, but not limited to vermin, humidity, bright light and rough handling. In recent times, electronic and digital communication media have succeeded paper for many purposes, providing instant access to vast amounts of information, including online news content. As has been the case with humankind's employment of all other media historically, digital delivery of news content means that it has never been easier to read, listen to or view this kind of information—nor has it been easier to lose it.

It also seems that for almost as long as humans have kept some sort of historic record, there has been an effort to keep it from disappearing. Clay tablets hardened in a kiln were among the first known archives. For example, in what was northern Mesopotamia, now northern Iraq, the great library of Ashurbanipal featured structures constructed to house and organize more than ten thousand tablets, dating back to the 7th century BC.⁴⁵

Of all memory institutions in the public consciousness, perhaps none is as well-known as the fabled Library of Alexandria, often portrayed as the greatest collection of knowledge ever gathered in the ancient world. Much, perhaps most, of the fascination with the Library of Alexandria derives from the popular legend that this magnificent structure and its vast numbers of scrolls and books were destroyed by one cataclysmic inferno in an act of war—a demonstration of the fragility of knowledge in a barbaric world. Current scholarship denotes this version of events as a “collection of myths and legends.”⁴⁶ The real story is apparently both less dramatic and more complicated, “a cautionary tale of

⁴⁵ Richard Ovenden, *Burning the Books: A History of the Deliberate Destruction of Knowledge* (Cambridge, MA: Belknap Press, 2020), 20.

⁴⁶ Ovenden, 31.

the danger of creeping decline, through the underfunding, low prioritization and general disregard for the institutions that preserve and share knowledge.”⁴⁷

In other words, the Great Library of Alexandria fell to neglect, not fire. One possible lesson that we might take away from this legend is that human complacency may be the most dangerous enemy of archives. If we wish to preserve the knowledge of our time, including digital news content, humankind will need to be proactive and ever vigilant. We will need to continue and enhance digital preservation activities for the long-term public good.

“While the default for physical artifacts is to persist (or deteriorate in slow increments), the default for electronic objects is to become inaccessible unless someone takes an immediate pro-active role to save them. Thus, we can discover and study 3,000 year old cave paintings and pottery... but we’re unable to even decipher any of the contents of an electronic file on an 8-inch floppy disk from only 20 years ago.”⁴⁸ Howard Besser

There are many ways that archives can serve the public good, including a circumstance that resonates with current events as this report was prepared and written: the SARS-CoV-2 pandemic. From the early days in the spread of COVID-19, journalists were digging into news archives to glean relevant information from similar public health threats. One of the most obvious comparisons involved the flu pandemic of 1918, which killed at least 50 million people in every part of the world.

What can we learn from the pandemic experience of over 100 years ago? The data—much of it from news reports of the time—provides evidence that public health policies can have a profound effect on the health of entire cities. For example, according to newspaper reports, in 1918 officials in Philadelphia refused to cancel a war-bonds parade despite strong warnings from doctors. “Within a week of the parade, more than 45,000 people in Philadelphia were infected with influenza, as the entire city, from schools to pool halls, ground to a halt.... Within six weeks, more than 12,000 Philadelphians were dead.”⁴⁹ Around the same time, St. Louis cancelled a similar parade, closed schools and discouraged large social gatherings. As a result, the death toll in St. Louis was kept under 700.

Humankind can benefit greatly from the lessons of history, but only if a clear, accurate record is passed from one generation to the next. If a deadly pandemic strikes again in 100 years, will journalists, historians and researchers in the year 2121 have the chance to benefit from our experience today? The answer to that question is not obvious. There currently is no clear pathway from the systems holding born-digital news content today to some version of publicly-accessible archives of the future. It does seem apparent from this study that, despite many risks and challenges, contemporary news content is still valued enough to not be deleted. The window of opportunity for long-term preservation is still open for a great deal of born-digital news content, but that opportunity will not last indefinitely.

47 Ovenden, 36.

48 Howard Besser, “Longevity of Electronic Art,” February 2001, <http://besser.tsoa.nyu.edu/howard/Papers/Tmp/elect-art%20longevity.html>.

49 Meagan Flynn, “What Happens If Parades Aren’t Canceled During Pandemics? Philadelphia Found Out in 1918, with Disastrous Results.” *The Washington Post*, March 12, 2020, <https://www.washingtonpost.com/nation/2020/03/12/pandemic-parade-flu-coronavirus/>.

It is encouraging that the principles and practices of digital preservation continue to advance. This study has been an effort to increase knowledge about the workflows, technologies and policies of modern newsrooms. It seems clear that a better understanding of how news enterprises work in the digital age is critical to identifying key findings that point toward potential pathways for progress in saving the born-digital news record. Our team of researchers is grateful to the many people who shared their time and knowledge with us as we explored the complex processes behind producing and distributing digital news content. Without exception, in all the news organizations and memory institutions we visited, we found champions for preserving born-digital journalism. It is our hope that this report will encourage and assist them and engage other such champions to avoid complacency, to pass the frangible journalistic record of our time, in ones and zeros, into the hands of generations to come, as past generations have done for us.

Adobe Creative Cloud: a collection of more than 20 cloud-hosted desktop and mobile apps and services for photography, design, video, web and others, including Photoshop, Illustrator, InDesign, and Premier. Formerly known as Adobe Creative Suite.

Airtable: a cloud-hosted spreadsheet-database hybrid.

Akamai: a Content Delivery Network (CDN) is a service offering geographically distributed proxy servers to serve web pages and streaming content to users.

Arc XP: formerly Arc Publishing, digital publishing software created by The Washington Post, hosted in AWS. Key modules are: Page Builder, WebSked, Composer, Photo Center, Video Center and Bandito.

AudioVault: a broadcast-related audio content management system used at NPR and other radio stations.

Avid: a software company that produces video and audio editing software, including MediaCentral, formerly iNEWS, and Avid Production Suite.

AWS, Amazon Web Services: a cloud computing environment offering services ranging from servers to storage. The most frequently used service is S3, enterprise cloud storage.

BitCentral: a software company providing content management, workflow and broadcast controls systems, primarily for video/broadcast video, including Core News.

BLOX CMS, TCMS, BLOX Total CMS: a news publishing content management software from TownNews, including the BLOX web CMS, TCMS print production system, Field59 video management system, and related tools and modules for multichannel digital, print and broadcast products.

CCI, CCI Europe: the former name of the Danish company now called Stibo DX.

Chartbeat: hosted web analytics software used to track customer behavior on news websites and in apps, newsletter, podcasts and other channels.

Chorus: a content management and publishing platform built by VOX Media, which uses it for hundreds of company-owned websites and began offering this publicly to other media companies in 2018.

CLIR, Council on Library and Information Resources: a nonprofit organization that promotes research, teaching and learning with libraries, cultural institutions and higher education, and runs the Digital Library Federation (DLF).

CMS, Content Management System: a generic term for content management software used to manage content for publishing, most often used to refer to web publishing software, but also often for multiple digital or print channels.

Colo, Co-location: a data center facility in which a business can rent physical space for its own servers, storage and related computing hardware in a secure facility with redundant electrical power and other features designed to minimize downtime.

Content type: a term commonly used to refer to categories of digital content files stored in a publishing system, such as text, photo, video, graphic, audio, or to a content package or structure containing these, customized for various digital channels such as social media, website, mobile app, blog, OTT service, IoT device, print or other product.

Core Publisher: an internally developed web news publishing and content management platform used by most National Public Radio stations, which are transitioning to a new system called Grove.

CoreNews: a broadcast news content management system for broadcast video news production from BitCentral.

CrowdTangle: internet monitoring software that publishers use to track the popularity of news content posted through social media channels, including detection of content going “viral.” Now owned by Facebook, which bought the company in 2016.

CRL, Center for Research Libraries: an international consortium of university, college, and independent research libraries that acquires, preserves and distributes traditional and digital materials from around the globe, used by member institutions for research, learning and teaching.

CUE Publishing: a suite of content management and news publishing software systems from Stibo DX, including CUE Web, the web CMS component; CUE Print, the print publishing system of CUE, formerly NewsGate, itself a multi-channel system for edition-based digital or print products; CUE DAM, a digital asset management and archival system, formerly DC-X from Digital Collections, a Hamburg, Germany company recently acquired by Stibo DX, which is based in Aarhus, Denmark.

DAM, Digital Asset Management: a generic name for software systems used to store, organize, manage, access and distribute file-based digital objects. A DAM may integrate with a CMS or other production tools, but keeps an original, unchanged version of the file to guard against change or loss; also, the practice of managing digital assets.

Data architecture: this term usually refers to the structure of data in a content management or publishing system. It can refer to the physical (storage media) or database structure (schema), but for workflow analysis purposes refers to the object model for data.

DC, Digital Collections: a digital asset management system, developed by Digital Collections. DC was bought by CCI/Stibo DX in 2019 and is now being integrated with CUE Publishing as the CUE DAM platform.

Deck: a newspaper term for a subordinate headline or short article summary that accompanies the main headline of an article.

Designated Community: digital preservation experts define a Designated Community as an identified group of potential consumers or users who should be able to understand a particular set of information. May be composed of multiple user communities.

Django: a website development framework built on the Python programming language. It is free and open source.

Document Cloud: an open-source, cloud-based file hosting service to help journalists share, analyze, annotate and, ultimately, publish source documents to the web.

Drupal: a free and open-source web content management framework written in PHP.

DTMH, Dodging the Memory Hole: a series of forums engaging journalists, scholars, technologists and digital preservation experts to address the preservation of born-digital news content, sponsored by the Donald W. Reynolds Journalism Institute and the University of Missouri Libraries. Areas identified for action include awareness, legal framework, policies, resources, standards, practices and technology.

E-edition, e-paper: two common names for the digitally-delivered replica of the daily newspaper, usually accessed by readers over the web, via links sent by email every morning to subscribers. Recently, e-editions have begun including expanded content not included in print editions.

ENPS: radio news production and content management software created by The Associated Press and used by many radio and TV organizations.

Escenic: former name for web CMS originally developed by Vizrt, an Oslo company that primarily supplies graphics systems to the broadcast TV industry. CCI Europe bought Escenic in 2013. It is now CUE Web, provided by Stibo DX.

ESRI: mapping and analytics software company.

Exif, exchangeable image file format metadata: the part of a still image (photo) file that contains info for each frame, such as camera type/model, aperture setting, shutter speed, color space, resolution, focal length, size and geolocation info.

Factiva: a subscription-based news information and research service, owned by Dow Jones & Co.

Field59: a video content management system for news media companies offered by TownNews.

FTP / ftp, File Transfer Protocol: a standard method of moving digital files from one computer to another. One of the original elements of TCP/IP, the core of what eventually became the Internet.

GBH: the current name for the Boston-based public television and radio station originally named WGBH, the largest in the U.S. GBH changed its name in August 2020, dropping the “W” designation used for broadcast operations to emphasize its increasingly digital nature.

Google Analytics: an internet traffic and audience service used to track customer behavior on news websites and in apps, newsletter, podcasts and other channels, owned by Google.

Google Docs: a suite of web-hosted office productivity applications for text editing, spreadsheets, slide presentations and other content types. Google is owned by Alphabet Inc.

Graphene: a customized Content Management System created by the LA Times, built on the Brightspot commercial platform.

GUID, Globally Unique Identifier: Also UUIDs or Universal Unique Identifiers: technically they are 128-bit unique reference numbers used in computing which are highly unlikely to repeat when generated despite there being no central GUID authority to ensure uniqueness.

IPTC, International Press Telecommunications Council: a consortium which develops standards for the news industry. Its best-known standard is the so-called IPTC metadata that is embedded in photo files, including JPEG and several other image file formats.

ISO 9000: an International Organization for Standardization Series certification. A set of international standards on quality management and quality assurance.

Jira: a project and ticket management system used for most software development work, especially development groups using the Agile method. Owned by Atlassian.

LAMP, or LAMP stack: a combination of software often used to run websites, web services and related technologies. Its components consist of Linux, Apache, MySQL and PHP. Linux is the server OS, Apache is the open-source web server, MySQL is the open-source relational database, and PHP is the open-source web server-side scripting language.

Lexis-Nexis: a subscription-based news information and legal research service, owned by RELX Group.

LOC, Library of Congress: the largest library in the world, with millions of books, recordings, photographs, newspapers, maps and manuscripts in its collections. The Library is the main research arm of the U.S. Congress and the home of the U.S. Copyright Office.

Lightroom: photo editing and organizing software from Adobe, now part of the Adobe Creative Suite.

LTO, Linear Tape-Open: a magnetic tape data storage technology originally developed in the late 1990s as an open standards alternative to the proprietary magnetic tape formats that were available at the time.

MAM, Media Asset Management: similar to a digital asset management system, but with features that are optimized for video production workflows.

Media Desk: the digital asset management system that is part of the Brightspot software suite.

MediaCentral: the workflow software that is part of the Avid video and audio suite of systems.

MerlinOne: digital asset management software that provides planning, assignment and production workflow tools in addition to full asset management functionality.

Metadata: a set of data that describes and gives information about other data, usually represented as fields in a database or additional information that is part of a file format, such as the IPTC fields included in digital photo files.

NDIIPP, National Digital Information Infrastructure Preservation Program: a preservation effort by the Library of Congress from 2000–2018 to archive digital information, develop strategies for and promote digital preservation. The organization is no longer active. The National Digital Stewardship Alliance, hosted by the Digital Library Federation, has assumed a similar role since 2016.

NewsBank: a news database service that contracts with publishers to resell access to archives of hundreds of media publications.

NewsFlex: NPR's story workflow and content management system for audio files.

NewsGate: multi-channel content management software that was recently rebranded as CUE Print and used primarily for print or edition-based digital products as part of the CUE Publishing suite from Stibo DX.

Newspapers.com: a news database service that contracts with publishers to resell access to archives of hundreds of media publications.

NLP, Natural Language Processing: software systems able to analyze and understand written or spoken communication presented in a natural, human manner.

OAIS, Open Archival Information Systems Reference Model: a plan for architectures, standards, and protocols for system design, metadata requirements, assessment, and other issues central to digital preservation.

Open Calais: the original name for a natural language processing system that automatically analyzed written content and applied metadata tag identifiers for subjects and entities, from Thomas Reuters, now part of its Intelligent Tagging Service.

OTT, Over The Top: a streaming media service offered directly to viewers via the Internet. OTT bypasses cable, broadcast, and satellite television platforms, the companies that traditionally act as a controller or distributor of such content.

P2P: the name used for in-house developed content management tools used at the Chicago Tribune, Tribune Corp. and the LA Times news organizations.

Parse.ly: web analytics software that tracks customer behavior on news websites and in apps, newsletter, podcasts and other digital channels.

PARS International: a content licensing agency.

Permalink: a permanent static hyperlink that is intended to remain unchanged for many years into the future, yielding a URL that is less susceptible to link rot; a type of persistent identifier.

Photo Mechanic: photo browsing, metadata editing and workflow automation software designed to run on individual computers, from Camera Bits. In common use by photographers.

Polopoly: former name of the web content management system from Atex, now part of its Atex Content Engine (ACE) system.

Preservation repository: a set of software components designed to ensure access, viability, security, usability and discoverability of content for the long term.

PressReader: an E-edition provider to the news publishing industry.

ProQuest: a news database service that contracts with publishers to resell access to archives of hundreds of media publications, along with scholarly publications and other resources.

RAW: an imaging format which contains the full original digital photograph data captured by digital cameras, comparable to a digital negative. These mostly proprietary formats vary by camera manufacturer and sometimes by camera model. RAW camera files can provide the highest possible resolution for original digital images

Ruby Shore: a commercial web content management software and website development company.

S3: the most commonly used type of cloud storage provided by Amazon Web Services; often used for the primary enterprise storage solution for business applications.

SAAS, Software As A Service: a technology model in which customers pay for access to use the central system hosted and run by the provider. The only thing they own is the license to use the software for the specified period. There are no components installed on-premises.

SAN, Storage Area Network: refers to high-speed disk storage systems purchased by companies that want data on-premises.

SCC MediaServer: digital asset management software from Software Construction Company (SCC) that provides planning, assignment and production workflow tools in addition to full asset management functionality.

Schema: the structure or layout of a database, usually referring to the fields in one or more database tables, their type and definition, and the relationships between tables.

SEO, Search Engine Optimization: the process of customizing content and metadata to maximize the capabilities and features of major search engines, primarily Google Search.

Slack: a cloud-based proprietary instant messaging platform.

SPOT, Simple Property-Oriented Threat Model for Risk Assessment: a digital preservation strategy that defines six essential properties of successful digital preservation: availability, identity, persistence, renderability, understandability, and authenticity. The SPOT model focuses on preservation outcomes rather than specific causes of a threat.

Stibo DX: the new name for the software development company that provides the CUE Publishing multi-channel news platform. Formerly CCI Europe. Stibo DX is part of the Stibo Group, run by a 200-year-old foundation with a Danish royal charter from 1794.

Taxonomy: a scheme of classification, usually involving a hierarchy of nodes such as subjects, topics or relationships.

TCMS, BLOX Total CMS: component system in the TownNews suite of software that publishes to print and multiple digital channels.

Tech stack: the term used to refer to all of the technology hardware and software systems and services used to support a single application, including the computer (laptop or data center server), the operating system (Linux, MacOS, Windows), and the application software (web browsers, word processing application or spreadsheets), plus any virtualization layers (Docker containers, virtual machines) that are involved in enabling the application.

Technology Lifecycle: usually used to refer to the timeline from the development of technology tools or systems to the implementation, usage period and end-of-life, the point where the technology is no longer used or supported by the originating organization.

TownNews: a news publishing software company that provides hosted news publishing and broadcasting systems, including the BLOX web CMS, BLOX Total CMS and related tools and systems. Based in Moline, IL, owned by Lee Enterprises.

Trax: photo assignment and tracking software available from MerlinOne.

WordPress: a web content management system written in the PHP programming language and paired with MySQL, an open-source database management system, and the Apache web server. WordPress, which originated as a blogging platform, is now the most popular web CMS on the internet. It is available as a free version of the software or through tiered pricing and hosting options.

Bibliography

- Abernathy, Penelope Muse. "News Deserts and Ghost Newspapers: Will Local News Survive?" Center for Innovation and Sustainability in Local Media, Hussman School of Journalism and Media, University of North Carolina at Chapel Hill, 2020. https://www.usnewsdeserts.com/wp-content/uploads/2020/06/2020_News_Deserts_and_Ghost_Newspapers.pdf.
- Alverson, Jessica, Kalev Leetaru, Victoria McCargar, Kayla Ondracek, James Simon, and Bernard Reilly. "Preserving News in the Digital Environment: Mapping the Newspaper Industry in Transition." Center for Research Libraries, April 27, 2011. https://www.crl.edu/sites/default/files/d6/attachments/pages/LCreport_final.pdf.
- American Archive of Public Broadcasting. "American Archive of Public Broadcasting Permanent Entity Grant." Accessed February 23, 2021. <https://americanarchive.org/about-the-american-archive/projects/permanent-entity>.
- American Library Association. "Digital Millennium Copyright Act," January 24, 2019. <http://www.ala.org/advocacy/copyright/dmca>.
- The Associated Press. "The Associated Press." Accessed February 23, 2021. <https://www.ap.org/en-us/>.
- Besser, Howard. "Longevity of Digital Art," February 2001. <http://besser.tsoa.nyu.edu/howard/Papers/Tmp/elect-art%20longevity.html>.
- Blue Ribbon Task Force on Sustainable Digital Preservation and Access. "Sustainable Economics for a Digital Planet: Ensuring Long-Term Access to Digital Information," February 2010. https://www.cs.rpi.edu/~bermaf/BRTF_Final_Report.pdf.
- Breeding, Marshall. "Building a Digital Library of Television News." *Computers in Libraries*, June 2003. <https://librarytechnology.org/document/10346>.
- Center for Cooperative Media. "Center for Cooperative Media." Accessed February 24, 2021. <https://centerforcooperativemedia.org/>.
- Chicago History Museum. "Millions of Moments: The Chicago Sun-Times Photo Collection." Accessed March 5, 2021. <https://www.chicagohistory.org/exhibition/millions-of-moments-the-chicago-sun-times-photo-collection/>.
- Denver Public Library. "The Rocky Mountain News at the Denver Public Library," October 7, 2015. <https://history.denverlibrary.org/rocky-mountain-news-denver-public-library>.
- Dictionary of Archives Terminology. "Archives." Accessed February 18, 2021. <https://dictionary.archivists.org/entry/archives.html>.
- Dublin Core Metadata Initiative. "DCMI: Home." Accessed February 24, 2021. <https://dublincore.org/>.
- Flynn, Meagan. "What Happens If Parades Aren't Canceled During Pandemics? Philadelphia Found Out in 1918, with Disastrous Results." *The Washington Post*, March 12, 2020. <https://www.washingtonpost.com/nation/2020/03/12/pandemic-parade-flu-coronavirus/>.
- The GDELT Project. "The GDELT Project." Accessed February 23, 2021. <https://www.gdelproject.org/>.
- Giardina, Carolyn. "Industry Scrambling After Japan Earthquake, Tsunami Lead to Tape Shortage." *The Hollywood Reporter*, March 20, 2011. <https://www.hollywoodreporter.com/news/industry-scrambling-japan-earthquake-tsunami-169456>.

- Google Cloud. "Vision AI." Accessed February 23, 2021. <https://cloud.google.com/vision>.
- Halvarsson, Edith. "Introduction to Digital Preservation: What Is Digital Preservation?" Oxford LibGuides, February 12, 2021. <https://ox.libguides.com/digitalpreservation/whatisdp>.
- Hansen, Kathleen A., and Nora Paul. *Future-Proofing the News: Preserving the First Draft of History*. Lanham: Rowman & Littlefield, 2017.
- International Consortium of Investigative Journalists. "International Consortium of Investigative Journalists." Accessed February 24, 2021. <http://www.icij.org/>.
- Internet Archive. "Internet Archive." Accessed February 23, 2021. <https://archive.org/>.
- International Press Telecommunications Council. "IPTC Photo Metadata Standard." Accessed February 23, 2021. <https://iptc.org/standards/photo-metadata/iptc-standard/>.
- "ISO 16363:2012: Space Data and Information Transfer Systems – Audit and Certification of Trustworthy Digital Repositories," February 2012. <https://www.iso.org/standard/56510.html>.
- Krakov, Gary. "Radio Is Going Digital." NBC News, March 11, 2003. <https://www.nbcnews.com/id/wbna3078252>.
- Lavoie, Brian. "The Open Archival Information System (OAIS) Reference Model: An Introductory Guide (Second Edition)." Digital Preservation Coalition, 2014. <https://www.dpconline.org/docs/technology-watch-reports/1359-dpctw14-02/file>.
- Library of Congress. "Chronicling America." Accessed February 18, 2021. <https://chroniclingamerica.loc.gov/>.
- Mara, Neil. "News Archives: The Untapped Resource." Columbia, MO: Donald W. Reynolds Journalism Institute, University of Missouri, forthcoming. <https://hdl.handle.net/10355/80861>.
- McCain, Edward. "How the Denver Public Library Ended Up Owning the Rocky Mountain News Archive." Donald W. Reynolds Journalism Institute, December 17, 2014. <https://www.rjionline.org/stories/how-the-denver-public-library-ended-up-owning-the-rocky-mountain-news>.
- National Endowment for the Humanities. "National Digital Newspaper Program." Accessed February 23, 2021. <https://www.neh.gov/grants/preservation/national-digital-newspaper-program>.
- National Endowment for the Humanities. "U.S. Newspaper Program." Accessed February 23, 2021. <https://www.neh.gov/us-newspaper-program>.
- National Museum of African American History and Culture. "The National Museum of African American History and Culture Will Acquire a Significant Portion of the Archive of Ebony and Jet Magazines," July 25, 2019. <https://nmaahc.si.edu/about/news/national-museum-african-american-history-and-culture-will-acquire-significant-portion>.
- Ovenden, Richard. *Burning the Books: A History of the Deliberate Destruction of Knowledge*. Cambridge, MA: Belknap Press, 2020.
- PBCore. "PBCore Metadata Standard." Accessed February 24, 2021. <https://pbcore.org/>.

- "Reference Model for an Open Archival Information System (OAIS)." Recommendation for Space Data System Practices. The Consultative Committee for Space Data Systems, June 2012. <https://public.ccsds.org/Pubs/650x0m2.pdf>.
- Reilly, Bernard. "The Library and the Newsstand: Thoughts on the Economics of News Preservation." *Journal of Library Administration* 46, no. 2 (2007): 79–85. https://doi.org/10.1300/J111v46n02_06.
- Ringel, Sharon, and Angela Woodall. "A Public Record at Risk: The Dire State of News Archiving in the Digital Age." Tow Center for Digital Journalism, Columbia University, 2019. <https://academiccommons.columbia.edu/doi/10.7916/d8-7cqr-q308>.
- "Sustaining Public Media Archives: Summary of Sustainability Discussion Hosted by CLIR and WGBH." WGBH and the Council on Library and Information Resources, November 2017. <https://clir.wordpress.clir.org/wp-content/uploads/sites/6/2016/09/Sustaining-Public-Media-Archives.pdf>.
- University of Missouri Office of Research and Economic Development. "Institutional Review Board." Accessed February 19, 2021. <https://research.missouri.edu/irb/>.
- Vermaaten, Sally, Brian Lavoie, and Caplan, Priscilla. "Identifying Threats to Successful Digital Preservation: The SPOT Model for Risk Assessment." *D-Lib Magazine* 18, no. 9/10 (October 2012). <https://doi.org/10.1045/september2012-vermaaten>.
- Victoria McCargar Consulting, and Victoria McCargar. "Missouri J-School and the 'Backstory.'" Report. Victoria McCargar Consulting, 2008. <https://mospace.umsystem.edu/xmlui/handle/10355/45033>.
- Waters, Donald, and John Garrett. "Preserving Digital Information: Report of the Task Force on Archiving of Digital Information." Commission on Preservation and Access and The Research Libraries Group, May 1, 1996. <https://clir.wordpress.clir.org/wp-content/uploads/sites/6/pub63watersgarrett.pdf>.
- Wood, Savannah. "New Magazine Celebrates Local Black Women Suffragists." *Afro* (blog), April 24, 2020. <https://afro.com/new-magazine-celebrates-local-black-women-suffragists/>.

Appendix A

This appendix shows the initial set of interview questions used in this project research, prior to the revisions prompted by the pandemic. Interviews up to May 2020 used these questions, interviews after this point used the revised questions in Appendix B below.

Introduction and consent to participate in research

The University of Missouri Libraries and the Donald W. Reynolds Journalism Institute (RJI) are conducting research through interviews to help us understand how born-digital news content is being preserved and used at contemporary news outlets. The information you provide will guide us as we seek to define the technology, workflows and policies necessary to keep digital journalism from disappearing forever.

You have been selected for this interview because of your status as an employee of a news organization. Your participation is voluntary, and you can stop at any time. The risks associated with participating in this interview are minimal.

This project is funded by a grant from The Andrew W. Mellon Foundation.

If you have questions or concerns about the interview, or if you feel under any pressure to enroll or to continue to participate in this study, please contact the University of Missouri's Institutional Review Board (which is a group of people who review the research studies to protect participants' rights), 489 McReynolds Hall or at (573) 882-3181 or the principal investigator, Edward McCain, Digital Curator of Journalism, 218 RJI or at (573) 882-8049.

Cover sheet questions

Today's date is [month-day-year] and we are located at [name of organization, city and state].
Let's go around the room and introduce ourselves by giving our name, position and a very quick description of what we do in our organization.

Content types questions

How does your news organization deliver your products?
What are the primary distribution channels for your original content?
What are the secondary distribution channels?

Web/Mobile

- Organization site
- Native (mobile) apps
- Other

Social media (what unique content is posted directly vs what is just linked to a site)

- Facebook
- Twitter
- Instagram
- YouTube
- Blogs
- Other

Video

- Broadcast television
- OTT

Audio

- Broadcast radio
- Podcast
- Smart speakers

Print (including PDFs and e-Prints)

- Scripts (for newscast)
- Newspapers
- Periodicals
 - Magazines

Other

- Aggregators such as LexisNexis, Factiva, NewsBank, ProQuest
- Reader access to previously published content

Workflows questions

Describe the people, technology, processes media formats and metadata you use throughout each of these processes:

Media formats include:

- Text
- Graphics
- Layout design
- Photo
- Audio
- Video
- Tables
- Interactive applications
 - Databases
 - Other

Planning

- People - Processes - Media formats - Metadata

Creating (Third party licensed content?)

- People - Processes - Media formats - Metadata

Editing

- People - Processes - Media formats - Metadata

Producing/Packaging

- People - Processes - Media formats - Metadata

Publishing/Broadcasting/Distributing

- People - Processes - Media formats - Metadata

Analytics

- People - Processes - Media formats - Metadata

Archiving/Preserving

- People - Processes - Media formats - Metadata

Reusing

- People - Processes - Media formats - Metadata

If a reporter wants to access content that is not-yet-published/published/unpublished, how do they get that information?

Tech stack questions

What CMS(s) are you currently using?

What are the functions of each CMS (web, print, social, etc.)?

- Why did you select this particular CMS?

How is the CMS deployed (local/cloud/both)?**What technologies are used for the CMS?**

- Does the CMS store and keep data from third-party systems, such as geo- or data-mining services?
 - DocumentCloud?
 - Google Maps?
 - ArcGIS?
 - Other?

How many objects are stored in the CMS?**What is the amount/size/footprint of the content in the CMS?What metadata do you collect?**

- How is it added (automatically/manually)?

Is there a backup system/process?

For broadcast, what production suite(s) are you using?

What technologies are you using for your production suite?

- Why did you select this particular production system?

How is your storage technology deployed (onsite/cloud/both)?**How many objects are held in your production suite?****What is the amount/size/footprint of the content in the production suite?****What metadata do you collect?**

- How is it added (automatically/manually)?

Is there a backup system/process?

Do you use a DAMS/MAMS for news content?

What kinds of assets do you keep in your DAMS/MAMS?

- Text
- Photo
- Audio
- Video
- Other

How is your storage technology deployed (onsite/cloud/both)?

What technologies are you utilizing for your DAMS/MAMS?

- Why did you select this particular DAMS/MAMS?

How many objects are stored in your DAMS/MAMS?

What is the amount/size/footprint of your DAMS/MAMS data?

What metadata do you collect?

- How is it added (automatically/manually)?

Is there a backup system/process?

How is the news content stored?

What technologies are you using to store news content?

- Why did you select this particular storage technology?

How is your storage technology deployed (onsite/cloud/both)?

How many objects are you currently storing?

What is the amount/size/footprint of the news content you store?

What external objects do you currently store?

What metadata do you collect?

- How is it added (automatically/manually)?

How often do you back up your news content?

How many copies do you keep?

Do you test your backups and, if so, how often?

What formats do you use to store content?

- JPEG
- WAV
- ProRes
- Other

Approximately how much content do you have in your systems now?

How many stories?

How many digital objects?

How many years?

How many Terabytes/Petabytes?

What audience analytics tool(s) are you currently using?

What are the functions of your analytics tools?

- Why did you select this particular analytics tool?

How is the analytics tool deployed (local/cloud/both)?

What technologies are used for analytics?

What is the amount/size/footprint of the content in analytics?

Is there a backup system/process?

- Do you keep copies of analytics data on your own storage facilities?

Tech lifecycle questions

Where is your organization in implementation of news production/distribution/archiving technology?

Any current or planned migrations? (Hardware, Storage, software?)

What is your system selection process?

Do you have internal product development? Please describe.

Data architecture questions

Describe your approach and standards for data architecture

What taxonomies do you use?

For example:

- Term list
- Hierarchies
- Thesauri

What schemas do you use?

For example:

- IPTC
- PBCore
- Schema.org
- Dublin Core (DCMI)
- Other - describe

Is the metadata for SEO or for internal searches?

How do you tag rights for third-party content?

Policies questions

What are your policies for the following?

Please indicate if they are written/formal policies or just accepted practice:

What is your selection policy?

How do you decide what to add to your archive/library?

- What types of content are important to keep?
- What formats do you retain?
 - Production formats:
 - Master formats:
 - Do you run into problems capturing particular kinds of resources or certain formats?
- Written/formal or accepted practice/informal?
 - If written, who writes the policy?
 - How is policy enforced?

Which people within your organization are primarily responsible for selecting which assets to keep? (Reporters, editors, photographers, photo editors, others?)

What is your retention policy?

How do you decide what to keep in your archive/library?

How long is news content kept after publication/broadcast?

- Does the retention period vary?
- Is news content intentionally deleted (how often? Why? Who decides?)

Which people within your organization are primarily responsible for ensuring the preservation of particular assets? (Reporters, editors, photographers, photo editors, others?)

Are there written policies around retention of stories/data?

- Who writes the policy?
- What priorities are used to determine what gets kept for later access?
- How are policies communicated?

How successful has your org been at keeping news content accessible for the long term?

Which of your preservation practices are most successful? Which pose the most challenges?

- What would you need to mitigate challenges?
- Have you unintentionally lost any stored news content?
 - How much?
 - What happened?
 - Did anything change as a result?

How many copies do you make of stored collections (redundancies)?

Number of copies?

How often backed up?

Are backup collections tested? What are those policies/procedures?

Are there different policies for:

- Production formats?
- Master formats?
- Analytics data?

Describe other elements of your preservation strategy:

Do you use analytic/business information to make decisions about what to keep?

This appendix shows the revised set of interview questions used in this project research, following project changes prompted by the pandemic. Interviews after May 2020 used these questions, interviews before this point used the original questions shown in Appendix A above.

Digital News Preservation Research Project

A research project of the University of Missouri Libraries and the Donald W. Reynolds Journalism Institute, supported by a grant from The Andrew W. Mellon Foundation.

Introduction: This document outlines the questions we are seeking to discuss with you and your colleagues about issues related to preserving born-digital news content at contemporary news outlets. The shift to digital publishing has revealed a number of new challenges and hurdles that make it far more difficult to ensure long-term access and preservation of news than it was in the past.

We understand that the purposes and practices of preservation can vary significantly across the news industry. This study is focused on understanding these variations. The information you provide will guide us as we seek to understand current publishing practices and define the technology, workflows and policies necessary to keep digital journalism from disappearing.

The questions and topics that follow are grouped for discussion with different parts of a news organization, system provider or preservation-related institution. We anticipate that only some of them will apply to each interviewee. All are included to help provide overall context about this project.

We plan to record the interviews for internal purposes only to ensure accuracy and assist with analysis. Interviews are voluntary and all information we request is subject to your permission.

Questions for interview subjects

A: Organization and management: These questions are aimed at understanding your organization's plans, intentions, policies and responsibilities when it comes to preserving published news content.

Interview subjects: this information might best be provided by a newsroom manager or high-level editor who understands the organization's overall goals, plans and workflow processes.

- 1:** Does your organization preserve news content? What is preserved, and for whom? For what purposes?
- 2:** Does your news organization include long-term preservation of news content, and public access, as part of its overall mission? Is this expressed in written goals/aspirations, or is this implied?
- 3:** Who are the primary users of any saved content that your organization has collected?
- 4:** What roles within the organization are part of any content preservation process, if any? Or are you relying on an external organization for this function?
- 5:** What are the key challenges the organization faces in archiving and preserving news content? Have these changed in any significant way with the shift to digital publishing? How, or in what ways have they changed?
- 6:** Does your news organization recognize or accept any responsibility to preserve the content you have created for the long-term future, for the public record?

B: Content / Collections: These questions are aimed at identifying the news content produced at your organization, the channels to which it is published, what content is preserved, and the process for those selections.

Interview subjects: this information might best be provided by a news librarian/archivist, if any; a newsroom or technology manager, product owner or high-level staff members who is knowledgeable about content preservation functions, systems and processes in your organization.

7: What is the primary news content the organization produces?

8: What are the primary distribution channels for this content?
Please provide at least basic information on all that apply.

Examples:

- Web publishing
- Broadcast TV/Radio channel(s)
- TV OTT
- Native mobile apps
- Newspaper (print)
- Social Media (please identify which platforms have curated content)
- Third-party platforms with prearranged or negotiated feeds, such as Apple News, Kindle, etc.
- Podcast(s)
- Smart speakers
- Internet of Things (IoT)
- Magazine, other periodical
- Documentaries
- Others (please specify)

9: What are the secondary distribution channels for this content? Please provide at least basic information on all that apply. Please see the list above for examples.

10: Please describe the content and collections accumulated to date. How far back does this material extend? Are there any gaps or limits for specific date ranges, or segments, either in content saved or in metadata for that content? Please describe briefly how and why these variations occurred.

11: When deciding what content to preserve, what if any guidelines or policies are followed in that selection? In other words, what is saved and why?

12: What file types or formats are preserved?

13: Does your organization bring in and publish content from non-staff generated sources, such as freelancers, wire services or stock image/footage houses? Please identify these sources and your rights to retain and preserve that content long term, if any?

14: Does your organization distribute published content to third-party syndication services for distribution to the public, institutions or for other purposes? What are your licensing arrangements for these services?

C: Systems, Search and Metadata: These questions are aimed at identifying and understanding the technology systems used for publishing and preservation by your organization, along with metadata used in the process, and access to your content.

Interview subjects: this information might best be provided by a news librarian/archivist, if any; a newsroom or technology manager, product owner or high-level staff members who is knowledgeable about content preservation functions, systems and processes.

15: What Content Management Systems (CMS) or publishing systems are used? What role do they serve?

16: Do you use any Digital Asset Management Systems (DAMs), or Media Asset Management Systems (MAMs)? What role do they serve?

17: What metadata is used in the publishing process and how is it applied, edited or removed through the workflow? What metadata is preserved, and how is it preserved? Is anything added in a preservation step or process?

18: Do your publishing and preservation systems use, and preserve, structures or metadata that connect all parts of a published story? Is there a way, for example, to find not only the text of a story, but also images, videos, graphics, social media, podcasts, broadcasts, print and other channel representations for each story?

19: Does any of the metadata used in your organization follow or utilize any of the archival specification standards currently in use or under consideration, such as IPTC, Dublin Core, PB Core, PREMIS, etc.?

20: Is content uniquely identified in a way that carries across systems? How is this done?

21: Who has access to preserved content inside your organization? Are there any internal limits on access? How about unpublished content, meaning content that was previously published but is no longer public?

22: Who has access outside your organization, including readers/customers, the public, students and researchers? How do they access your content?

23: How is the metadata used in searching for and retrieving content within your organization, and outside your organization? Do you tailor metadata to any specific search engines, platforms or browsing/consumption environments?

24: What other collection or content management systems are in use? What role do they serve?

D: Storage and Technology Management: This group of questions is about the extent of preserved content, the systems and techniques used in data storage and management.

Interview subjects: this information might best be provided by a technology manager or high-level technical staff members who are knowledgeable about publishing and preservation systems.

25: Approximately how much content is currently retained in systems now (unique copies)?

- Stories, images, videos, audio clips, etc.
- Other digital objects
- Date ranges
- TB/PB

26: How many copies are there of stored collections? How long do you keep content?

27: What is your data storage technology and how does it ensure data integrity?

28: Do your storage systems meet any of the archival standards for data storage, such as ISO 16363, ISO 9000 or others?

29: What file types or specifications are accepted, supported, or required for content to be preserved (formats, file types, etc.).

30: What are the key drivers of recent technology or system changes? Are there any planned technology changes which will or could impact preservation?

31: Does the organization have a disaster plan and/or a business continuity plan? Please describe it briefly.

E: System Vendor Questions: This group of questions is aimed at understanding the full capabilities of technology systems used in the news publishing industry, including capabilities related to news preservation.

Interview subjects: this info might best be provided by high-level managers, sales support staff and product owners knowledgeable about your systems and common customer implementations and workflows.

32: What systems do you offer the news publishing industry? Please describe their functions, capabilities and workflows, and if possible, provide a demo.

33: Please outline content types supported, your internal data or file formats, and data file formats your systems can handle for input/output.

34: What is the current tech stack for your systems, including database management and backup processes?

35: What metadata standards, taxonomies, data mining, tagging processes or other data architecture standards/approaches do your systems use?

36: Do your system offerings include preservation or archival capabilities? We are interested in any digital asset management (DAM or MAM) capabilities, including usage tracking, rights management, etc. Also, what tools or techniques are used to ensure data integrity?

37: What is the intended use of the system with regard to long-term archiving and preservation? Is it intended to support long-term preservation? Could it be used to support long-term preservation?

38: Do your systems primarily serve the news publishing industry, or other industries? Is your company focused on and financially supported by news organizations primarily, or through other industries? Please clarify if others.






For additional information or if you have questions, please contact the principal researcher for this project, Edward McCain, Digital Curator of Journalism, 218 RJI, (573) 882-8049; or the University of Missouri's Institutional Review Board, 489 McReynolds Hall, (573) 882-3181, which oversees university research in order to assure best practices.

Interpreting the SPOT Model for news organizations

This is the text of definitions we used for the SPOT model analysis in this research, which was based on descriptions and details of content provided by news organizations that were interviewed. These definitions are derived from the original model.

Using a process guided by AVP, the research team applied the SPOT Model to all of the news organizations that we interviewed, assigning a numerical range from 0–4 for each property. This allowed us to assess and compare organizations in many different ways, identifying those doing a better job of ensuring content integrity and longevity. We applied the scoring guide using the following interpretation.

SPOT Model Scoring Guide

SCORE	INTERPRETATION
0 	Organization is not addressing this property at all
1 	Organization is minimally addressing this property
2 	Organization is partially addressing this property (e.g., for some content types and not others)
3 	Organization is fairly comprehensively addressing this property
4 	Organization has fully addressed this property for all content

Each member of the team was instructed to score each of the 24 news organizations on their preservation of born-digital news content according to the six SPOT properties using the Scoring Guide above. Partial scores, such as 2.5, were not allowed. Participants were advised to use a loose or general interpretation of each SPOT property when assigning a score. In addition, team members were asked to provide a short rationale for each score.

To facilitate our research team’s assessment process, we provided lists of key factors to consider when attempting to evaluate each organization’s performance for a given SPOT property.

The foremost quality we considered, which was factored into the rankings for all six properties, was very simply the completeness of content from a news story. The more content objects from the original “package” that were preserved, the higher the potential ranking. Thus, if all objects, including accompanying photographs, video, and datasets, if any, from a story were preserved, the ranking could be higher, perhaps a 3 or a 4. If only the text objects were captured from a package, the ranking would be limited to 2, with possible exceptions.

The first SPOT property, “**availability**” assesses whether a digital object is available for long-term use. This is the most basic and essential quality for maintaining any prospect of preservation. For born-digital news content the team interpreted this to mean the presence of the package in a CMS, DAM or MAM system, rather than necessarily a dedicated preservation repository. In evaluating news organizations for this property, the research team focused on the following questions:

- How much of the content from the original story is present in the system?
- Is the content captured/stored?
- Is the content in a centralized location?
- Is content accessible to a designated community (internal users of the news organization)?

When applying the second SPOT property, “**identity**,” the property of being referenceable or independently identifiable, distinct from other content, the team assigned ratings based on these questions:

- How much of the content from the original story is present in the system?
- Is the content independently identifiable?
- Is sufficient metadata applied to uniquely identify the object?

The third property, “**persistence**,” is the property that the bit sequences comprising a digital object continue to exist in a usable/processable state and are retrievable/processable from the medium on which they are stored. This property is only possible if the storage environment is robust and secure. The key questions posed for assessing Persistence were:

- How much of the content from the original story is present in the system?
- Is the storage environment current/robust?
- Is the data corrupted/at risk of being corrupted?

Note was made that Amazon’s AWS, a storage solution widely used by news organizations studied, was acceptable (suggesting a maximum rating of three), but higher scores should be given for additional levels of backup, including local or cloud approaches.

Property number four is “**renderability**,” when a digital object is able to be used in a way that retains all of its significant characteristics, the qualities that are essential to its accessibility and meaning over time. Importantly, the determination of which characteristics are considered to be significant is made by the designated community, internal users of news organizations. For example, if reporters or editors need to see the news content as published, that rendering becomes a significant characteristic and the ability to produce content closer to the original version would be rated higher than plain text, preliminary drafts, or such raw materials of the story as unedited video or audio recorded in the field. The team’s suggested questions for consideration when rating news organizations for renderability were:

- How much of the content from the original story is present in the system?
- Can the content be rendered in a way that could be considered acceptable to stakeholders (internal users from the news organization itself)?
- Are content objects saved in a form/format that can be read using today’s software and hardware?
- Do linkages between original content objects allow reconstruction of the original content package, though not original appearance or user experience?
- Do linkages enable the original content package and original appearance or user experience to be re-rendered, using either original or emulated stack. (receives highest rating)

“Understandability,” the fifth property, requires associating enough supplementary information with archived digital content such that the content can be appropriately interpreted and understood by its intended users. Understandability is concerned with issues associated with complex content. If the content is not complex, then it will be easier to render and understand in the future, but if it needs additional metadata to be understood, as is the case with interactive graphics, the organization must preserve that metadata or the full environment required to enable such use or viewing of the content. Preserving the interactive environment and tools to understand complex content in the future is a frequent challenge for news organizations. Important questions for rating understandability include:

- How much of the content from the original story is present in the system?
- Can the entirety of the content be independently understood?
- Is there sufficient metadata to ensure that each piece of media that is part of the story package is independently understandable?
- Is the full picture of the story and all of its components understandable within their original context (i.e., what section of the publication it was in, what program was it aired on, what date, etc.)?

Major threats to understandability can stem from the repository’s inattention to characteristics of the content that its community of current and future users require, failure to preserve the metadata required to maintain functionality of the content, and ill-considered retention policies.

The sixth and final property, **“authenticity”** is the property that a digital object is, either as a bitstream or in its rendered form, what it purports to be. The quality of authenticity necessitates a complete rendering of the original story. Alternatively, if loss or change has taken place, this property requires a careful record of version changes after the original publication. In news organizations, the content is rarely static and may undergo many changes over time. If an organization has a methodology for tracking changes post-publication, it scored higher. Essential questions asked by the team in order to rate the authenticity of content in newsroom systems include:

- How much of the content from the original story is present in the system?
- Is there a record of versioning/changes after original publication or broadcast?
- Are there policies around which versions are saved, and for how long, and when (higher score for this)?
- Does the record of changes apply to all components of the story?

Note was made that it is generally not practical to keep all versions of a video story due to space/cost limitations.

Appendix D

This is the list of subject keywords used by Lee Enterprises newsrooms to tag news stories at the time of the interview in January 2020. This is an example of a shared controlled vocabulary that facilitates identifying, locating and sharing content.

STANDARD SECTIONS Channel Name	STANDARD TAGS Channel Keyword	STANDARD PAGES Section (if applicable)
Agribusiness/ranching	agribusiness	/business
Architecture	architecture	varies
Arts & culture	arts-culture	/entertainment
Entertainment	entertainment	/entertainment
Food & drink	food-drink	/entertainment/dining OR lifestyles/food-and-cooking
History	local history	news/archives
Home & garden	home-garden	/lifestyles/home-and-garden
K-12 education	k-12-education	/news/local/education
Local environment	environment	/news/local/environment
Local government	local-government	/news/local/govt-and-politics
Local health care	health-care	/lifestyles/health-med-fit OR business
Local personalities	personalities	varies
Local politics	local-politics	/news/local/govt-and-politics
Major college sports	college-sports	/sports/college
Major industry	major-industry	/business
Medical	medical	/lifestyles/health-med-fit OR business
Military	military	varies
Nation World	nation-world	/news/national OR /news/world
Opinion & commentary	commentary	/opinion
Pets & animals	pets-animals	/lifestyles/pets
Prep sports	prep-sports	/sports/high-school
Pro sports/baseball	pro-baseball	/sports/baseball/professional STL only
Public safety/courts	public-safety	/news/local/crime-and-courts
Recreation & outdoors	recreation-outdoors	lifestyles/recreation hike-bike
Rivers and lakes	rivers-lakes	/news/local/environment
Science & technology	science-technology	/business/technology OR /lifestyles/technology
Sense of place	local-places	varies
Sports & fandom	fandom-sports	varies
State	state-and-regional	/news/state-and-regional
State government	state-and-regional	/news/local/govt-and-politics
Outdoors	state-government out-doors	/outdoors fish-hunt
Travel - Go and see	travel	/travel
Universities/colleges	college-education	/news/local/education
US/Mexico border	us-mexico	varies
Weather	weather	/weather

There currently is no clear pathway from the systems holding born-digital news content today to some version of publicly accessible archives of the future. It does seem apparent from this study that, despite many risks and challenges, contemporary news content is still valued enough to not be deleted. The window of opportunity for long-term preservation is still open for a great deal of born-digital news content, but that opportunity will not last indefinitely.



University of Missouri

University Libraries

Donald W. Reynolds Journalism Institute