Hanna Soltys

DTMH 2016 Scholarship Project

December 2016

For: Edward McCain, Digital Curator of Journalism, Reynolds Journalism Institute

*Scholarship Project Introduction*

As part of the scholarship to attend the “Dodging the Memory Hole” forum at the University of California-Los Angeles in October 2016, winners had to complete a research project in conjunction with their studies at the forum. For this project, I sought out to better understand today’s landscape of preserving digital news, the scope of archives for this type of news medium, and identify potential solutions to help garner better results for the field.

This topic was one of great interest to me as I completed an undergraduate program in Journalism and am now in my second year of a Masters program for Library Sciences, archives concentration. Seeing these two worlds coming together opened my eyes for potential job opportunities as well as the immediate need to begin problem solving for these specific types of consumed news.

After attending DTMH 2016, I truly see the power of collaboration among various industries and levels of experience. With journalism teachers, reporters, students, archivists, and librarians around me during this conference, I saw the power firsthand of different types of minds solving for the same problem (mainly during group work).

My project focus will begin with a look at the landscape of digital news, why digital news archiving solutions aren’t a one-size fits all due to dynamic content and social media, and lastly look at solutions from other industries and entities working to save digital work. Widening the scope from digital news archives allows for greater discussion on preservation of digital materials and potentially identify what can be modified for the digital news media world.

## *Today's Landscape*

When beginning to look at archival work within the digital and digital news realms, it's important to first clarify what exactly constitutes "digital." For purposes here, digital preservation and archiving techniques will solely be centered around those materials created as born digital files with interactions happening online.

For digital preservation to truly be effective, we must take a step back and recall the integrity and authenticity of our records. As a repository or archive, you must prove with evidence, the record is what it claims to be and has not been corrupted. With digital files easily manipulated and overwritten, it's a greater challenge than one anticipates, and should be used when drafting guidelines and protocols as this is the main goal.[1]

Also important to note before beginning a deep dive into the preservation practices is that the general landscape of digital preservation talk is scattered. We know it's important to do, yet we're not quite sure how exactly to get the work done. Or even who should be responsible for preserving and collecting. This can be attributed to the many facets and characteristics of the digital materials we are creating.

A common preservation method used today is a bit passive and not as effective as we need it to be. Quick of digital material online uses web crawlers to capture content, though due to permissions, restrictions, and web crawling durations, we have a vast and wide unarchived web. These types of crawls also often skip over any embedded properties from computer languages such as JavaScript and Flash. Given the dynamic content of today's digital news, web crawls are quickly showing the inefficiency to help preserve digital news.[2]

*What is Digital News?*

Not all digital news is created equal, requiring a look beyond a one-size-fits-all solution you often see in other archival practices of physical materials. Digital news today consists of apps, infographics, social media postings, among others, in addition to the news entity's own website. All of these different platforms give us a multitude of file formats. In a presentation from Tim Gollins, the Head of Digital Archiving at the National Records of Scotland, specifically about file format types, he cautioned getting too caught up in the race to find the ever-lasting file type.

> "Useful file formats are dictated by a populous use. JPEGs aren't going to go out of style tomorrow because we're all using them. Don't spend money predicting the future."[3]

Gollins point is one not often heard throughout the archives industry in the US as most cite file formats as a barrier for long-term digital preservation. That said, there are other formats we should perhaps be more worried about when looking at digital news preservation.

In a survey conducted in Maine by Jennifer E. Moore and Jennifer L. Bonnet, we gain an understanding of how newspaper, librarians, archivists, and researchers assess and see the preservation landscape of digital news. When discussing born digital material, the survey results showed that Twitter and Facebook were the top content creations, with news articles and online video following behind. Though when looking at content that was archived among these entities, born-digital news articles, blogs, and infographics were the top materials. And often, social media was not considered a high priority for archiving. Though if we're making the most of our content on these sites, why aren't we concerned about how to preserve it? With stories happening on various sites and platforms, we have different user experiences, yet we're only deeming a handful of experiences worthy of preservation.[4]

As many know, using social media platforms presents its own challenges we can't solve due to permissions and restrictions. This type of barrier prevents capturing, accessing, and sharing, even from personal/corporate accounts. One could attempt to transfer data from a site, such as Twitter, though the Library of Congress should serve as a precaution. After many years with the large data file, the project of archiving all of the tweets has not gone according to plan.

While it would be easy to take the focus off of social media platforms with these barriers, today's world of reporting and consumption means we will lose stories. Often, stories happening here are ones not happening in the same type of manner on organization websites. For example, the #JeSuisCharlie hashtag which served not only as awareness for the Paris attack, but also served as a form of solidarity for free speech, was continually This hashtag, and in turn, the posts accompanying it, began to shape a news story all on its own.[5]

Other news content often overlooked when archiving are news apps. Meredith Broussard, assistant professor at New York University, has published articles and spoken at conferences directly cautioning and providing solutions of what we can do with these dynamic pieces of content currently being ignored in archiving practices.

In Broussard's, "Preserving news apps present huge challenges", these types of news stories contain many, many layers of code for an interactive experience, rendering screenshots non-effective as a preservation method. She discusses how HTML static pages (one for each layer of a news application) and the use of a national news app registry can help determine what should be preserved and how we can preserve it.

An immediate issue with this approach is the amount of static HTML pages given that these types of applications have thousands of layers. While the Internet archive preserves static web pages, we must look to preserve the dynamism and images within the text.[6]

## Storage Options

Even when we figure out how to archive and preserve all of our dynamic content, we then must look to storage. In the survey by Moore and Bonnet previously discussed, most kept their archives on servers, the cloud, or a third-party service. With various storage options available, we have the ability to find the one directly aligning with our institutions' needs, workflows, and more importantly, our materials/records and accession needs.[4]

Storage containers aren't the only thing we should worry about with our digital archives placement. Content disappearing overnight is of great concern, especially when turning over media archives to cultural institutions or public entities. The *Milwaukee Journal Sentinel* saw this first hand in October 2016. After donating their digital files to Google News Archive, the collection disappeared due to a new system implementation. While collaboration is important within the field to generate solutions, we all must too be responsible for creating and maintaining our own systems simply to ensure we truly are the owners of our creations.[7]

One possible method to help alleviate some of the worry with storage could be to use a repository designated to a certain type of dynamic content. Broussard mentions the use of creating an app repository to help with preservation of these dynamic pieces.

Using the Rhizome ArtBase repository, we see how to capture and ensure dynamic content is viewed in the proper context with specific hardware and software notations. Requiring this level of metadata could allow us to archive and preserve pieces as long as the technology is still functioning. In addition, using a repository would allow for a standardization throughout digital media archiving practices, ensuring the budgets and urgency of archiving remains top of mind and integrated into workflows.[6]

*Projects and Initiatives*

Every conference you go to, report or journal you're reading, you see the importance collaboration plays not only within our designated fields, but also from other types of professionals. Preserving digital-born files and the challenges is not specific to news media. You see this playing out in records management of corporations, digital tapes at film studios, 3D blueprint architecture drawings, the world today demands a digital life and in turn, we've created quite the digital life that's a bit harder to archive. Because of this, we'll now look at various projects, initiatives, and solutions being implemented to help preserve digital news. Whether these methods will be successful and a long-term solution is still up in the air, though at the very least, we have a place to begin.

Wanting to record the sentiments of the Irish population during various events throughout the year by following hashtags, The Social Repository of Ireland set out to find ways to record and preserve these social/news media events and developed a framework to help preserve the sentiments and feelings happening in such a social manner. The collected tweets are limited, though the Repository acknowledges this. Instead, they look at the framework they developed (a four-step process) and have found ways to preserve tweets with relevant metadata, making accession easy for future use.[5]

The academia world and in particular, their journals, is a great place to look for ideas when developing preservation and accession practices. At the University of Michigan Library for example, they archive in the HathiTrust system using an mPach toolset. Adding this toolset to the system gives creators (journal editors) an integrated workstream of publishing and preserving. When publishing born-digital material in this system, metadata is created as a byproduct and creators have the opportunity to convert materials into other archival languages including XML and Journal Article Tag Suite (JATS). Directly tying preservation into the workstream allows for easy migration once the proper facets are in place.[8]

It's hard not to draw comparisons to the work these journal editors are doing and the work our media editors and writers are doing. Developing a workflow into a set content management system that a news media organization shows the power in collaborating with technologists and software programs to create an adaptation toolkit for preservation use.

In a funded study from the National Endowment for the Humanities, the Chronicles in Preservation project focused on digital newspaper preservation efforts across academia using three distributed digital preservation (DDP) systems. From this project, the "Guidelines for Digital Newspaper Preservation Readiness", an analysis of DDP systems, and interoperability tool. From this project alone, digital media outlets have a starting point to become familiar with terminology and technology available within the archiving and preservation field. While academia newspapers might have varying levels of data and materials for preservation, there is work being done to look at formats, metadata packaging, and organization of digital newspapers that can be applied, no matter the institution's size.[9]

Broadcast too is making efforts for their digital archives and how best to preserve for future use, despite their infancy in development. These archives are taking cues from digitized text, image, and music archives to determine best practices and release of the digital broadcast archives.

These types of archives are also looking at ways to educate users of the materials. What good is preserving if it's not accessible and useful? With interactive workshops, the BBC archives is able to determine what user needs may be and how to tweak metadata to reflect user behavior and thinking to help with identification. One dynamic piece of metadata they are including in their materials are transcripts and time codes. The hope is that these transcripts would be able to find key words within a transcript and in turn, the time code, making the searching process for specific material much easier.[10]

Keeping with the BBC, BBC Scotland has drastically improved their archives and findability casting aside the card catalogue and opting for a dynamic desktop interface in the form of a database. With archive services placed in siloed entities throughout the organization with each repository featuring its own organization and cataloguing of materials, the database was able to bring various departments together without compromising the user. This database system, INFAX, allowed cross-pollination among departments while enhancing the user experience with interoperability.[11]

These are a mere representation of the experimentation going on with born-digital files. In most literature, another concern with preservation is to whom the responsibility falls to. We've seen in various instances where it's the responsibility of the creator, the archivist, a third-party, a librarian, etc. No matter the title running the archive initiative, we can all agree on the need to get the work done, and to do so with the support of those outside of your immediate department and staff.

*Conclusion*

At the end of the day, Tim Gollins said it best, "We keep the bits safe. And then we make the bits useable." While the preservation of digital news materials will continue to morph and shift as technology and behaviors evolve, it's important to not lose sight of the need to begin some sort of preservation, even if it's not a legacy solution. Given the vast and rapid way we develop digital news; it will be easy to fall behind if we continue to let the days pass by without determining some sort of answer(s).

## References

[1]Delaney, B., & de Jong, A. (2015). Media Archives and Digital Preservation: Overcoming Cultural Barriers. *New Review of Information Networking*, *20*(1/2), 73-89.

[2]Huurdeman, H. C., Kamps, J., Samar, T., Vries, A. P., Ben-David, A., & Rogers, R. A. (2015). Lost but not forgotten: finding pages on the unarchived web. *International Journal On Digital Libraries*, (3-4), 247. doi:10.1007/s00799-015-0153-3

[3]Gollins, Tim. "Digital Archiving for University of Wisconsin-Milwaukee." 2016. Presentation.

[4]Moore, J. E., & Bonnet, J. L. (2015). Survey finds differences on preserving born-digital news. *Newspaper Research Journal*, *36*(3), 348-362.

[5]Harrower, N., & Heravi, B. R. (2015). How to Archive an Event: Reflections on the Social Repository of Ireland. *New Review of Information Networking*, *20*(1/2), 104-116.

[6]Broussard, M. (2015). Preserving news apps present huge challenges. *Newspaper Research Journal*, *36*(3), 299-313. doi:10.1177/0739532915600742

[7]Warburton, B. (2016). Newspaper archive in limbo. *Library Journal*, (16). 20.

[8]Hawkins, K. S. (2015). Automated Creation of Analytic Catalog Records for Born-Digital Journal Articles. *Serials Librarian*, *68*(1-4), 299-306.

[9]Skinner, K., Schultz, M., & Zarndt, F. (2013). Chronicles in Preservation: Preserving Digital News and Newspapers. *Preservation, Digital Technology & Culture*, *42*(4), 199-203.

[10]Smith, J., Hammond, K., & Revill, G. (2016). Earth in vision: pathfinding in the BBC's archive of environmental broadcasting. *Historian (02651076)*, (129), 18-23.

[11]Wilson, J. j. (2016). From librarian to media manager: looking after BBC Scotland's archive. *Indexer*, *34*(3), 47-53.