News Media Section

Describing dynamic data journalism: developing a survey of news applications

Katherine Boss
Libraries, New York University, New York, USA.
Katherine.boss@nyu.edu

Meredith Broussard
Arthur L. Carter Journalism Institute, New York University, New York, USA.
merbroussard@nyu.edu

Abstract:

Preserving dynamic, born-digital data journalism requires more than web scraping, as news stories today are built from much more than just text and images. Data journalism projects like news applications, or "news apps," are composed of a database, the data in the database, the graphical interface that appears in the browser, accompanying text, and often images, videos, audio, and other multimedia components. Existing Internet archiving methods are not sufficient to save these data journalism projects for the future.

This paper summarizes the context and history of news apps archiving, and describes the development of a survey of news applications. This survey will be used to create a working list of news organizations that are producing data journalism in the United States and a better sense of how and where these projects are currently being stored. The results of the survey will inform recommendations and processes for archiving dynamic data journalism.

Keywords: data journalism, digital archiving, computational journalism, Internet archiving

Introduction

The history of newspaper archiving is filled with loss. For many centuries, the main threats to archives were natural disasters like fires or floods, or man-made disasters like war. But today's news, much of it published exclusively online, faces extinction of a different sort. It is the crisis of the born-digital age. Thousands of articles are published on the Internet every minute, yet their average lifespan has been estimated at 75-100 days (Rein, 2004). Born-

digital news has become the fragile ephemera of our time. Some of these articles, published by legacy organizations like The Washington Post or the Los Angeles Times, have established relationships with information vendors, and a version of the online article is distributed and archived through those channels. But a tremendous and growing volume of news content is being published by organizations with no established archiving system and no deliberate plans to save their web content for the future. Many of these influential startups, like Buzzfeed News, Vice, The Huffington Post, and Vox, have no plan for their back files should the newsroom be purchased or shut down.

The end of the website Gawker.com, a born-digital gossip magazine created in 2002, is a recent cautionary tale. Gawker was a website that very much reflected a time and place in early Internet history, and it wielded an outsized influence on media and society, particularly in New York City (Farhad, 2016). Gawker folded in the summer of 2016, after legal troubles forced the company into bankruptcy, and Gawker.com was bought by Univision for $135 million dollars (Smith, 2016). The archives included, crucially, not just the articles but also the comments and the dialogue that readers had with the authors. Yet no advance plans had been made for Gawker's archives, and the abruptness of the bankruptcy and the sale, compounded by the controversial nature of the site and its implosion, made it even more difficult to save at the last minute. The status of the Gawker archives, as of this writing, is still not clear. Informal conversations have suggested that the Internet Archive was able to save their entire back file of posts (Kathleen A. Hansen & Paul, 2017, p. 184), but a check of the holdings in July 2017 showed that archive.org had more than 16,000 snapshots of www.gawker.com dating back to 2003, with substantial gaps in coverage (no snapshots were taken for the entire month of February 2008, for instance). Quality assurance of each post was hit or miss, and comments were also archived only intermittently ("Summary of gawker.com," 2017). These snapshots, incomplete though they may be, are an invaluable collection given that Gawker is no longer in production.

As the Gawker example illustrates, librarians and digital archivists are able to copy and save much born-digital news content through web archiving technologies, with little to no assistance from the media company or news organization itself. Archive-It, the web crawler behind the Internet Archive's Wayback Machine, is a core technology in this effort, and digital archivists have been crawling and saving webpages since the early 1990s. There are many organizations today scrambling to increase the volume and quality of these born-digital news archives, including the Internet Archive, the Library of Congress, the National Digital Stewardship Alliance, the Educopia Institute, and the Donald W. Reynolds Journalism Institute at the University of Missouri (Cain, 2003; K. A. Hansen & Paul, 2015; McCain, 2015; Skinner & Schultz, 2014). These efforts are the thin binary line saving Internet history from oblivion.

However, as the Gawker example also illustrates, Internet archiving cannot be the sole strategy for saving born-digital content. Journalists and media organizations need to begin planning for posterity and preserving their born-digital legacy in archives. The urgency of this moment is even more apparent in regard to data journalism, as many of these stories have advanced to the point that they cannot be captured or archived by any current web archiving technology. News applications, or "news apps," fall into this category. They are some of the most innovative pieces of journalism being produced today. Examples include the visually exciting "Snow Fall" story by John Branch of The New York Times, the database-driven "Dollars for Docs" tool by the team at ProPublica, or the many apps produced by organizations like FiveThirtyEight that are devoted specifically to data journalism (Branch,

2012; Groeger, Ornstein, Tigas, & Jones, 2010; Silver, 2016). The subset of news apps that query a database are too complicated to be saved in any holistic way, and as such, they are disappearing.

It is imperative that librarians, scholars, journalists, and media companies collaborate to stem this loss before more dynamic digital journalism disappears into the metaphoric "memory hole" (McCain, 2015). But before beginning any large-scale project in saving these stories, a better sense of the number and nature of news apps being produced will help to inform best practices. This paper describes the development of a news apps survey, the first survey of its kind, which will be used to gather data on the way news apps are built and stored. Preliminary results from the survey pre-test are described here. The results of the final survey will be used to develop recommendations for tools to capture, archive, and preserve these vital data journalism stories so that they may be discoverable and accessible to future generations.

Newspaper Archiving in the United States

In the past 250 years, the United States has faced a number of major challenges in trying to preserve its journalism. In the 17th and 18th century, when paper was expensive and public libraries were extremely rare, maintaining a newspaper archive was often financially untenable. Many early publishers sold or pulped their archives to make money or to stay in business. Even the Associated Press lacked the foresight for early archiving, and retained no records of dispatches from 1846 to the 1930s (Kathleen A. Hansen & Paul, 2017). Efforts to preserve newspapers were few and far between, and when collections were developed or maintained, they were threatened by man-made and natural disasters. The U.S. Library of Congress made one such early effort to document and preserve newspapers, but during the War of 1812 the Library of Congress was set afire in a battle with the British, and all of the Library's books and newspapers were destroyed. Thomas Jefferson attempted to rebuild the Library from his personal collection, but "another fire in 1851 destroyed much of that reconstituted collection, including the newspapers" (2017, p. 17).

Newspaper archiving changed dramatically in the late 1920s with the commercialization of microform, a process of photographing each page of a newspaper, turning it into a microreproduction, and preserving those microreproduction images on more durable film ("Microform," 2014). But this migration of newspaper archives from print to microfilm faced other losses of quality. When newspapers were cut from their bindings and photographed to be microfilmed, quality assurance in the microfilming process was unreliable, and could often result in blurred text, warped text, gutter shadows, and general degradation; loss of the original color was also common, as black and white microfilm was far cheaper than color (Baker, Nicholson, 2000). Scholars and researchers have lamented these losses, large and small, for decades.

An important second migration of newspaper archives is currently underway. Microform readers, and people able to navigate them, are quickly becoming a thing of the past, and are being supplanted by digitized PDF archives. This process, ushered in by the Internet and the PDF/A file format, makes news and newspapers more accessible and discoverable to a global audience. But the process of digitizing print papers or converting microfilm archives to PDFs has brought with it challenges of quality assurance similar to those encountered during the transition from print to microform (degradation of text, loss of color). Digitization is also costly and time consuming. Fortunately, governmental and non-governmental organizations

have seen the need to fund this important work, and are leading the way in opening the knowledge these newspaper archives contain to the world.

Now as then, newspapers are undergoing a rapid and continuous evolution of journalism production, publishing, and distribution. History shows that these shifts result in a hole in the cultural record, as the new media format precipitates a process by which to archive and preserve it. This phenomenon has already begun with data journalism, but quick action could stem this loss before it grows larger.

Background of Data Journalism and News Apps

Most academic work on data journalism thus far has focused on defining the field and examining how it is produced. What is today called data journalism is the inheritor of what was called computer assisted reporting, or CAR, in the 1980s. CAR evolved from what was, in the 1970s, called "precision reporting," or applying the tools of quantitative social science to journalism (Meyer, 2002). In the past decade, the terms "data journalism" or "data-driven journalism" have become more prevalent than computer assisted reporting, a reflection of the advances in computing from machines that took up an entire room to machines that fit in our pockets (Coddington, 2015). Another interesting categorization of the field is Usher's (2016) description of interactive journalism: "a visual presentation of storytelling through code for multilayered, tactile user control for the purpose of news and information." For the purposes of this research, the differences between these categories are superficial; computer assisted reporting, data journalism, and data-driven journalism will all be examined. All need preserving, though the methods may vary slightly.

A news app is a type of data journalism artifact that incorporates a database, and is presented online and accessed via a web browser, in order to add context to the user's experience of a single story (Boss & Broussard, 2017). There is currently no canonical estimate of the number of organizations producing news applications, or the total number of news applications that have been published. However, some international and national surveys have been done in the United States, the United Kingdom, Norway, and Sweden that give an emerging picture of the volume and nature of this work being produced globally (Appelgren & Nygren, 2014; Fink & Anderson, 2015; Heravi, 2017; Howard, 2014; Knight, 2015; Stavelin, 2012).

In describing the parameters of a news application, the work of Young, Hermida, and Fulda is also helpful; their analysis of award-winning data journalism projects in Canada categorized these data journalism projects (2017). The team developed a series of semantic operations based on research from the information visualization community (Boy, Detienne, & Fekete, 2015; Yi, Kang, & Stasko, 2007) that could be performed when interacting with data journalism stories. This list of semantic operations included the ability of readers to:

- Inspect – specifics of the data
- Connect – click on one element and connect with similar or related elements
- Select – mark something as interesting
- Filter – show information conditionally
- Abstract/Elaborate – show more or less detail, usually on a map
- Explore – "show specifics based on a user's input/query
- Reconfigure – show a different arrangement
- Narrate – show a different section

(Young et al., 2017, p. 6)

A few of these semantic operations stand out as being necessary for database-reliant news applications: the ability to filter, explore, and reconfigure the data in the news story.

Other analyses of data journalism projects served only to demonstrate the imperative of archiving news applications. For example, a recent study that sought to assess the form, impact, and content of data journalism in the United Kingdom examined only data journalism that appeared *in print*, "…because visualisations are not available in archive form, and online sites are either inaccessible to trawling software… or contain little more than the print publication" (Knight, 2015, p. 59). A casual mention of a fundamental problem. Knight continued to sort the data elements she examined into the following categories: textual analysis, timeline, static map, dynamic map, graph, infographic, table of figures, list of numbers, or numerical pullquote. News applications were not included as a type of data journalism, but this is not surprising, as news applications do not exist in print format, so Knight would not have encountered any in her content analysis. Despite this glaring limitation, Knight went on to conclude, "the data journalism found in this study is largely superficial, institutionally sourced and non-remarkable" (p. 70). This painfully arrived at summarization is evidence of a much larger problem: the lack of any substantial archiving system for news applications and interactive journalism presents a clear problem for the future of research not just in journalism, but in all fields the rely on news and newspaper analysis. If researchers cannot access this content in a systematic way, it will go unexamined.

Archiving News Applications

The academic literature on archiving news apps is relatively new, and is addressed in the authors' previous work (Boss & Broussard, 2017; Broussard, 2015a, 2015b). These issues have gone largely unknown to researchers and were only felt within the data journalism community for many years. The problems were first discussed by news app developers who had the nostalgia and sense of posterity to take note that some of their projects, a mere five years after development, were already disappearing or being rendered inaccessible due to software obsolescence. In the first formal description of the challenges of archiving these projects, Broussard outlined several prominent examples and identified the thorny nature of the problem: news applications have multiple components, including a database, the data in the database, the graphical interface that appears in the browser, accompanying text, and often images, videos, audio, and other multimedia components (Broussard, 2015a, p. 300). If the news application is to be preserved, all of these components must be saved and reassembled in the same way, so as to ensure the look, feel, and functionality of the story can be recreated.

In their comprehensive book on the history of news and newspaper archiving in the United States, Kathleen Hansen and Nora Paul also touched on this issue in Future-Proofing the News. In the chapter on digital news, Hansen and Paul lament the loss of several important dynamic data journalism projects, including the Philadelphia Inquirer's 1997 story, "Blackhawk Down" ("Philadelphia Online | Blackhawk Down," 1997). "Although the site still exists, many of the multimedia features no longer function because the software that rendered the video, audio, and animations is long out of date" (2017, p. 194). A similar fate befell an early interactive from The Washington Post that focused on fatal medical helicopter

crashes. The text of the story accompanying the infographic is still available, but "none of the multimedia maps, interactive crash timelines, or trend infographics work" (2017, p. 194).

So how should journalists, librarians, and digital archivists approach the task of archiving dynamic data journalism? There are several options available, depending on the time and resources the institution is willing or able to devote to the effort. The optimal method would be to capture, archive, and preserve the news app in its entire software environment, a process that digital archivists refer to as emulation (Boss & Broussard, 2017). Several tools are in development that could facilitate this process and compress the files into a package small enough to allow them to be described and made discoverable in library catalogs. These involve the use of virtual machines that can encapsulate the operating system and all of the code necessary to recreate and run the news app, irrespective of the user's current operating system or software dependencies. Piccolo and Frampton (2016) included VirtualBox, XenProject, and VMWare (all open source) on their short list of recommended virtual machines for reproducing experiments or packing software environments, noting that there are dozens of others that could be considered.

One tool with a promising level of applicability to archiving news apps is the open source, computational reproducibility software ReproZip. Designed to make computational experiments reproducible across different platforms and over time, ReproZip could be customized for the specific needs of news apps (Chirigati, Rampin, Shasha, & Freire, 2016). The authors are currently investigating this option, but more work is needed on using virtual machines and reproducibility tools to pack and archive data journalism projects through emulation.

Developing a Survey of News Applications

As this is a new and developing area of research, an environmental scan and survey of news applications will help to inform which virtual machine or reproducibility software will be most applicable for news apps. As there was no published survey instrument that captured data on how news apps have been built and stored, it was necessary to develop an original survey to perform this environmental scan. A survey pretest was conducted in the fall of 2016 to refine the survey questions.

The survey pretest questions were based on the Performance Model Framework for the Preservation of a Software System (Matthews, Shaon, Bicarregui, & Jones, 2010), a framework previously identified as applicable to this research (Boss & Broussard, 2017). The framework included the following metadata categories needed to properly describe, archive, and preserve a software object: functionality, software composition, provenance and ownership, user interaction, software environment, software architecture, and operating performance. These broad categories were used as a starting point. The survey, created in Google Forms, was then distributed via email and on the listserv of the National Institute for Computer-Assisted Reporting (NICAR). It asked data journalists to describe the code, data, software libraries, and server environments of a news app they had created, as well as the proprietary and licensing information related to the data and editorial content.

Sixty responses were received from some of the major national and international organizations producing these stories, including the Los Angeles Times, The Washington Post, The Guardian, the Wall Street Journal, and ProPublica, among others. The responses revealed that the Performance Model Framework for the Preservation of a Software System

needed to be customized for specific systems, including news apps. For example, the framework makes reference to the location of binary files. Binary files are created when code is compiled, as in the C or Java programming languages. However, none of the news organizations surveyed in the pre-test were using programming languages that compile. Instead, they used interpreted languages such as Python or R. Thus, a better question to ask developers is where the source code is stored. The majority of respondents indicated that the source code is stored in a repository on Github or on the news organization's cloud server. The most common cloud service providers were Amazon Web Services (AWS) and Heroku. This is useful information because many computational reproducibility and virtual machine tools such as ReproZip are compatible with Github, AWS, and Heroku.

The survey pre-test also revealed several areas where language and the lack of shared meaning and terminology between archivists and data journalists presented problems. For example, the pretest included several questions related to if and where the news app was archived. Many of the open-ended responses to this question referred to GitHub and GitLab repositories, Amazon Web Services, or the organization's content management system. This indicated that the term "archive" has a different meaning among data journalists than among librarians. The wording and design of the question was altered to better collect data given this lack of shared terminology. However, the responses to this question also revealed that the code of most news apps is not backed up anywhere outside of the news organization's servers. This is worrisome, as it is quite common for news organizations to switch staff or switch service providers. Should a news developer leave the organization, and the service contract lapse, the organization's intellectual property will disappear. This finding reinforces the need for archiving processes and systems for data journalism in newsrooms.

Discussion and Future Directions

More advocacy and research is needed to find ways to capture, archive, and preserve news applications. There are many factors that complicate the effort to save news apps, including technical, legal, financial, and logistical challenges (Boss & Broussard, 2017). But the problem also suffers from a lack of awareness. This is partly due to the fact that news apps are not a commonly recognized type of data journalism. However, there is evidence to suggest that news apps differ enough from other types of journalism to warrant a separate and internationally recognized media format, which would facilitate and accelerate conversations about how to save them. The work of Young, Hermida, and Fulda (2017) as well as that of Boy, Detienne, and Fekete (2015) could be considered in establishing such a media format, in that news applications allow users to perform certain semantic operations other news media formats do not; specifically, the ability to filter, explore, and reconfigure data in a news story.

A survey of news applications will also be helpful in furthering discussions of how libraries, digital archivists, journalists, and news organizations can work together to prioritize a path to save dynamic born-digital content. The survey will provide a clearer picture of the nature and number of organizations producing news apps, as well as how the apps are built and stored. The results will be used to inform recommendations and processes for an archive of dynamic data journalism projects.

References

Appelgren, E., & Nygren, G. (2014). Data Journalism in Sweden: Introducing new methods and genres of journalism into "old" organizations. Digital Journalism, 2(3), 394–405. https://doi.org/10.1080/21670811.2014.884344

Baker, Nicholson. (2000, July 24). Deadline. The New Yorker, 42–56.

Boss, K., & Broussard, M. (2017). Challenges of archiving and preserving born-digital news applications. IFLA Journal, 43(2), 150–157. https://doi.org/10.1177/0340035216686355

Boy, J., Detienne, F., & Fekete, J.-D. (2015). Storytelling in Information Visualizations: Does it Engage Users to Explore Data? (pp. 1449–1458). ACM Press. https://doi.org/10.1145/2702123.2702452

Branch, J. (2012). Snow Fall: The Avalanche at Tunnel Creek. Retrieved October 28, 2016, from http://www.nytimes.com/projects/2012/snow-fall/

Broussard, M. (2015a). Preserving news apps present huge challenges. Newspaper Research Journal, 36(3), 299–313. https://doi.org/10.1177/0739532915600742

Broussard, M. (2015b, November 20). The Irony of Writing Online About Digital Preservation. The Atlantic. Retrieved from http://www.theatlantic.com/technology/archive/2015/11/the-irony-of-writing-about-digital-preservation/416184/

Cain, M. (2003). Being a library of record in a digital age. The Journal of Academic Librarianship, 29(6), 405–410. https://doi.org/10.1016/j.jal.2003.08.007

Chirigati, F., Rampin, R., Shasha, D., & Freire, J. (2016). ReproZip: Computational Reproducibility With Ease (pp. 2085–2088). Presented at the 2016 ACM SIGMOD International Conference on Management of Data (SIGMOD), San Francisco, USA. Retrieved from http://bigdata.poly.edu/~fchirigati/papers/reprozip-sigmod206.pdf

Coddington, M. (2015). Clarifying Journalism's Quantitative Turn: A typology for evaluating data journalism, computational journalism, and computer-assisted reporting. Digital Journalism, 3(3), 331–348. https://doi.org/10.1080/21670811.2014.976400

Farhad, M. (2016, August 24). Gawker's Gone. Long Live Gawker. Retrieved July 13, 2017, from https://www.nytimes.com/2016/08/25/technology/gawkers-gone-long-live-gawker.html

Fink, K., & Anderson, C. W. (2015). Data Journalism in the United States: Beyond the "usual suspects." Journalism Studies, 16(4), 467–481. https://doi.org/10.1080/1461670X.2014.939852

Groeger, L., Ornstein, C., Tigas, M., & Jones, R. G. (2010). Dollars for Docs. Retrieved December 12, 2015, from https://projects.propublica.org/docdollars/

Hansen, K. A., & Paul, N. (2015). Newspaper archives reveal major gaps in digital age. Newspaper Research Journal, 36(3), 290–298. https://doi.org/10.1177/0739532915600745

Hansen, K. A., & Paul, N. (2017). Future-Proofing the News: Preserving the First Draft of History. Lanham: Rowman & Littlefield Publishers.

Heravi, B. R. (2017). The State of Data Journalism Globally. Presented at the Data & Computational Journalism Conference, Dublin, Ireland.

Howard, A. (2014). The Art and Science of Data-Driven Journalism (pp. 1–145). Tow Center for Digital Journalism.

Johnston, L. (2014, March 11). Preserving News Apps | The Signal [webpage]. Retrieved July 16, 2017, from //blogs.loc.gov/thesignal/2014/03/preserving-news-apps/

Klein, S., & Fisher, T. (2014, March 18). Preserving interactive news projects with Newseum, OpenNews and Pop Up Archive | Knight Lab | Northwestern University. Retrieved March 27, 2014, from http://knightlab.northwestern.edu/2014/03/18/preserving-interactive-news-projects-with-newseum-opennews-and-pop-up-archive/

Knight, M. (2015). Data journalism in the UK: a preliminary analysis of form and content. Journal of Media Practice, 16, 55–72. https://doi.org/10.1080/14682753.2015.1015801

Matthews, B., Shaon, A., Bicarregui, J., & Jones, C. (2010). A Framework for Software Preservation. International Journal of Digital Curation, 5(1), 91–105. https://doi.org/10.2218/ijdc.v5i1.145

McCain, E. (2015). Plans to save born-digital news content examined. Newspaper Research Journal, 36(3), 337–347. https://doi.org/10.1177/0739532915600747

Meyer, P. (2002). Precision journalism: a reporter's introduction to social science methods (4th ed). Lanham, Md: Rowman & Littlefield Publishers.

Microform. (2014). Encyclopædia Britannica.

Philadelphia Online | Blackhawk Down. (1997). Retrieved June 16, 2017, from http://inquirer.philly.com/packages/somalia/sitemap.asp

Piccolo, S. R., & Frampton, M. B. (2016). Tools and techniques for computational reproducibility. GigaScience, 5(1), 1–13. https://doi.org/10.1186/s13742-016-0135-4

Rein, L. (2004, January 22). Brewster Kahle on the Internet Archive and People's Technology. Retrieved from http://www.openp2p.com/pub/a/p2p/2004/01/22/kahle.html

Silver, N. (2016). FiveThirtyEight. Retrieved April 5, 2016, from http://fivethirtyeight.com/

Skinner, K., & Schultz, M. (2014). Comparative Analysis for DDP Frameworks (p. 14). Educopia Institute. Retrieved from https://educopia.org/sites/educopia.org/files/deliverables/Comparative_Analysis_for_DDP_Frameworks.pdf

Smith, G. (2016, August 18). Gawker's Flagship Site Will Shut Down After Univision Deal. Retrieved July 13, 2017, from https://www.bloomberg.com/news/articles/2016-08-18/gawker-s-flagship-website-will-shut-down-after-univision-deal

Stavelin, E. (2012). Nyhetsapplikasjoner: Journalistikk Møter Programmering. In M. Eide, L. O. Larsen, & H. Sjøvaag (Eds.), Nytt På Nett Og Brett: Journalistikk i Forandring (pp. 107–125). Oslo: Universitetsforlaget.

Summary of gawker.com. (2017, July 13). Retrieved July 13, 2017, from https://web.archive.org/web/20080615000000*/gawker.com

Usher, N. (2016). Interactive journalism: hackers, data, and code. Urbana: University of Illinois Press.

Yi, J. S., Kang, Y. ah, & Stasko, J. (2007). Toward a Deeper Understanding of the Role of Interaction in Information Visualization. IEEE Transactions on Visualization and Computer Graphics, 13(6), 1224–1231. https://doi.org/10.1109/TVCG.2007.70515

Young, M. L., Hermida, A., & Fulda, J. (2017). What Makes for Great Data Journalism?: A content analysis of data journalism awards finalists 2012–2015. Journalism Practice, 1–21. https://doi.org/10.1080/17512786.2016.1270171