Challenges facing the preservation of born-digital news applications

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

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

Abstract:

Born-digital news content is increasingly becoming the format of the first draft of history. Archiving and preserving this history is of paramount importance to the future of scholarly research, but many technical, legal, financial, and logistical challenges stand in the way of these efforts. This is especially true for news applications, or interactive databases that comprise some of the most sophisticated journalism stories today, such as the "Dollars for Docs" database by ProPublica. Because news applications are standalone pieces of software, they cannot be preserved in the same way as text-based news stories, or fully captured by web archiving tools such as Archive-It. As such, news apps are currently disappearing. This paper will outline the various challenges facing the archiving and preservation of born-digital news applications, as well as outline suggestions for how to approach this important work, including the need for a technical and environmental report on the current status of news apps, as well as case studies in capturing, archiving and preserving this content.

Keywords: Born-digital news, News applications, Software preservation, Web archiving

Introduction

Archiving and preserving born-digital news is of paramount importance to the future of scholarly research. In the United States, initiatives to advance this work are being led by many archivists, researchers, and institutions, including 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; Hansen & Paul, 2015; McCain, 2015; "National Digital Newspaper Program | Library of Congress," 2015; Skinner & Schultz, 2014). However, none of these efforts have yet addressed the preservation of some of

today's most innovative data journalism projects: artifacts that news developers call a news app, which is short for "interactive news application."

What exactly is a news app? Scott Klein, senior editor for news development at ProPublica, gives the following definition in *The Data Journalism Handbook*: "A news application is a big interactive database that tells a news story. Think of it like you would any other piece of journalism. It just uses software instead of words and pictures" (2012). This is not an app that is downloaded to a phone, but rather a unique piece of software that is custom-built to tell a story. One example is ProPublica's "Dollars for Docs" project, which allows users to search a database to see if their own physician has received drug company money, and if so, how much (Groeger et al., 2015). These stories are unique in that they are interactive — users can query and explore the database. Because news apps like these are standalone pieces of software, they cannot be captured in the same way as text-based or multimedia-based news stories, and they are currently disappearing.

How many news apps are disappearing, and how many others are still in need of preservation? It isn't clear. Internet companies and digital media companies are bought, sold, consolidated, and bankrupted at a rapid rate. The media landscape will only get more complicated; the Pew Research Center estimates that there were 438 small digital news organizations in the US in 2013, most of which are digital-first startups (Pew Research Center Journalism Project, 2014). All of these organizations are producing journalism, and some of them are making news apps. This suggests the need for an environmental scan of interactive data journalism projects, and the development of selection criteria to identify priorities for capture and preservation (Broussard, 2015). Several informal efforts at such a registry have been started and abandoned (Han, 2014); there is a need to build upon these efforts and create a thorough analysis of the nature and number of news apps.

There is also the complicated and time-consuming task of determining best practices for caputing and archiving news apps for long-term access. Most born-digital news preservation efforts are focused on static news objects such as images and text, which comprise the vast majority of the born-digital news content being produced today. However, the question of how to archive and preserve dynamic digital news objects such as news apps has not yet been answered (Harris, 2013; Klein & Fisher, 2014). This is due in part to the unique design of these objects, which are built on complex software that is often connected to a database reliant upon various other systems to display and function fully. These software packages may be external and reside on different servers, further complicating matters (Waite, 2013). Preserving a news app is thus a markedly different exercise than archiving and preserving other news content on the web; it is an exercise in software preservation (Broussard, 2015).

This paper will outline some future projects, initiatives, and research needed for the successful capture, archiving and preservation of news apps. This includes the need for a report on the current status of news apps, consensus on a preservation framework, the establishment of best practices and techniques for approaching the work, metadata standards in describing the content, and more case studies in capturing, archiving, and preserving these data journalism projects.

A Report on the Status of News Apps

It is difficult to describe the number of news apps that have already disappeared or to identify those at greatest risk of disappearing, as there is currently no substantive effort to document the landscape of these projects. An initial environmental and technical analysis would be helpful in this regard, as it would collect information that is currently diffuse, as well as collect some basic descriptive information on the projects – how they are built and how they operate. This metadata would grant insights into the landscape of what needs to be archived and preserved.

The goals of this analysis are important to consider before beginning the research. One of the primary goals would be for preservation purposes. It would collect information needed by curators and archivists to archive and preserve the works. This would include descriptive information of the news app, the software architecture and environment, its creators, ownership and licensing information, and its archiving status.

Significant Properties of Dynamic Digital Objects

There are many challenges related to software preservation. The first relates to the fundamental question of what needs to be preserved. Unlike say a book or magazine, software packages do not have easily defined boundaries, and it can be difficult to determine just where the digital object begins and ends. Does one preserve just the source code, the binary executable version of the program, or the executable program as well as the software environment (hardware, operating system, programming languages and compilers, software libraries, etc) on which it was run? This difficulty has been identified in several case studies in preserving virtual worlds and mathematical, scientific, and e-science software (Matthews, McIlwrath, Giaretta, & Conway, 2008; McDonough & Olendorf, 2010). The answer to this question depends on several important factors and limitations, such as the human and fiscal resources available to devote to the project, and is generally determined on a case-by-case basis.

Depending on what boudaries have been established as constituting the “significant properties” of the news app, and on what curators and digital archivists have determined to be a feasible level of preservation, there are a few options for the preservation technique that might be used. If few resources can be devoted to the cause, then a bare-bones approach to saving news apps might decide, in some cases, that the database on which the news app is built could suffice as the object for preservation. In such cases, the boundaries of the digital object would be far more straight forward than if the entire look and feel of the news app were to be captured. The data could be described by the Data Document Initiative (DDI) metadata schema, which could be used to document the data as well as make it discoverable (Data Document Initiative Alliance, 2015). However, the downsides to this preservation approach are many. To capture and preserve just the data, and none of the interactive features of the news app, would be to sacrafice many of the components that make it desirable for preservation in the first place, including the user interface, data visualizations, and analysis. To supplement the preservation of the data set with web archive snapshots via Archive-It might make up for some of this loss. However, in the case of complex news apps that rely not only on a database but also on an external application programming interface (API) like the Google Maps API, this bare-bones preservation approach would strip the news app of most of its purpose and funtionality. Thus it does not seem plausible that this approach would be applicable in many instances.

Emulation vs. Migration

More robust techniques to digital preservation are still being debated in the information science community, and most notably include technical preservation, emulation, migration, and cultivation (Castagné, 2013). While each of these approaches have unique pros and cons, there are two techniques being used most commonly: emulation of the software environment, and migration of the software to current standards. While migration continues to be a successful strategy in providing long-term access to static digital objects, this approach has proved less successful in preserving dynamic digital objects:

> "Migration – the most widely used digital archiving strategy today – seeks to address this problem by changing the digital object in order to prepare it for access and rendering in future digital environments. Although this strategy applies to static digital objects such as images, text, sound and animation, it is not suitable for dynamic objects such as educational software or computer games. As a lot of digital material is becoming more advanced, relying solely on migration as a preservation strategy is risky and will certainly result in loss of authenticity and information" (Von Suchodoletz & Van der Hoeven, 2009, p. 147).

Consensus in the digital preservation community is leaning toward emulation as the preferred technique for the highest level of preservation for dynamic digital objects (Granger, 2000; Johnston, 2014; Von Suchodoletz & Van der Hoeven, 2009). However, as Anderson et al. have noted, the debate between emulation and migration is something of a false dichomoty, as emulators are also digital objects which will experience the same obscolescence as any other software over a long enough period of time. When a hardware paradigm shift occurs, as it inevitably will, "there are just two strategies that an emulation-based digital preservation approach can follow:1. Produce from scratch a new emulator (or emulators) to run on the latest hardware to provide access to digital objects dependent on previous platforms; or 2. Migrate the existing emulators" (Anderson, Delve, & Pinchbeck, 2010, p. 115). To produce new emulators from scratch would be time consuming and expensive, thus the authors recommend a hybrid emulation-migration technique of emulating the software to be preserved on an emulator designed with future migration of itself in mind. Advancements in this technique are being spearheaded by the European Commission's Keeping Emulation Environments Portable (KEEP) project ("European Commission : CORDIS : Projects & Results : Keeping Emulation Environments Portable," 2014). This hybrid approach seems most promising for long-term access to complicated software systems.

Legal issues also present a major obstacle in the preservation of news apps (Von Suchodoletz & Van der Hoeven, 2009, p. 151). Proprietary code, digital rights management, and intellectual property rights must all be taken into account when embarking on any digital preservation project. However, if emulation of the dynamic digital object is chosen as the preservation strategy, this can be advantageous from a legal perspective, as the creation of an emulator is often considered to be a form of reverse engineering. While this has generally been found to be a legal activity, "the ability to create an emulator can be limited by various laws, including copyright law, trade secret law, patent law, the Digital Millennium Copyright Act (DMCA), contract law, and the Electronic Communications Privacy Act (Electronic Frontier Foundation, 2010)" (McDonough & Olendorf, 2010, p. 54). Further work on establishing a legal pathway to the preservation of news apps is needed.

Adopting a Framework and Metadata Schema

Given that the field of software preservation is still an evolving one, it is also helpful to utilize a preservation framework to provide a broad overview of how to approach the work. Some frameworks are already emerging, most notably the Performance Model Framework for the Preservation of a Software System (Matthews, Shaon, Bicarregui, & Jones, 2010). Developed as a result of a 2008 study launched by JISC to determine the "significant properties" of software, the framework is built on case studies in mathematical, scientific, and e-science software preservation (Matthews et al., 2008). It identifies functionality, software composition, provenance and ownership, user interaction, software environment, software architecture, and operating performance as the general metadata categories needed to properly describe a software object (Matthews et al., 2010, pp. 98–101). This study and its recommendations provide an excellent outline for the needed categories in news app preservation.

A metadata schema should also be considered and adopted before beginning any work on news app preservation. A controlled vocabulary of description of the objects will allow for better discoverability, digital identification, and interoperability with various hardware and software systems (National Information Standards Organization (U.S.), 2004). Though a definitive schema for describing dynamic digital objects has not yet been agreed upon by the preservation community, there are several accepted standards that should be considered (additionally, complete consensus may not even be required; multiple schemas may be necessary to describe a spectrum of dynamic digital objects). These include the ISO Reference Model for an Open Archival Information System (OAIS), the Functional Requirements for Bibliographic Records (FRBR), and the Preservation Metadata: Implementation Strategies (PREMIS) (Consultative Committee for Space Data Systems, 2012; International Federation of Library Associations and Institutions, 1997; Library of Congress & PREMIS Editorial Committee, 2016).

Of these metadata schemas, PREMIS emerges as particularly applicable to news app preservation for a variety of reasons: it is complaint with the OAIS information model and more inclusive, and the newest release of PREMIS version 3 has been updated with the express goal of accounting for software as a primary object of preservation (Dappert, Peyrard, Chou, & Delve, 2013). Additionally, the Performance Model Framework for the Preservation of a Software System maps to OAIS terms, and since PREMIS is compatible with OAIS, the framework can be used and mapped to PREMIS v. 3 equivalent fields. Table 1 shows this relationship between the Performance Model Framework for the Preservation of a Software System, OAIS, PREMIS, and its applicability to news app preservation.

Table 1.

Framework Category (Matthews, et al., 2010)	Framework Description (Matthews, et al., 2010)	Examples (Matthews, et al., 2010)	Equivalent OAIS Terms	Equivalent PREMIS v.3 categories (select examples)
Functionality	- Description of the typical characteristics of software. - Useful for efficient discovery and accessibility of the software in	- Description of inputs and outputs - Description of operation and algorithms	- Descriptive Information	- Intellectual Entities or Object Entities (significantProperties, objectIdentifier,

	future	-·Description of the domain addressed		etc)
Software Composition	-Description of the components that constitute software -Useful for rebuilding and reusing the software in future -Detailed history of version changes and other significant changes that a software product has undergone facilitates verification of its authenticity	-A typical record: binary files, source code, user manuals and tutorials. -A more complete record: requirements and design documentation, test cases and harnesses, prototypes, formal proofs.	-Representation Information -Preservation Description Information (PDI)	- Intellectual Entities or Object Entities (environmentFuntion , environmentDesignation, etc)
Provenance and Ownership	-Different software components have different and complex licensing conditions. -Needs to be included in the preservation planning	-Software owner and licence information, e.g. Microsoft for MS Word®	-Provenance Information category of Preservation Description Information (PDI)	- Rights Entities (rightsStatement, licenseInformation, etc.) -Agent Entities (agentIdentifier, etc.)
User Interaction	-Description of expected mode of interaction between user and software -The 'Look and Feel' and the model of user interaction can play a significant role in the usability of the software and therefore should be considered among its *Significant Properties*.	-The inputs which a user enters through a keyboard, pointing device or other input devices, such as web cameras or speech devices -The outputs to screens, plotters, sound processors or other output devices	-Not comprehensively addressed in the OAIS – may be categorized as the Significant Properties of software	- Event Entities (eventDetail, eventOutcome, etc)
Software Environment	-Description of the environment that the correct operation of the software depends on -Dependencies between software environment related entities and history of changes made to them	-Hardware platform, operating system, programming languages and compilers, software libraries, other software products, and access to peripherals. -Binaries usually require an exact match of the environment to function	-Representation Information	-Intellectual Entities or Object Entities (relationshipType, relatedObjectIdentifier, etc)
Software Architecture	-Plays a significant part in the reproducibility of the original functionality and features of software	-Client/server, peer-to-peer, and Grid systems all require different forms of distributed system interaction which would require the configuration of hardware and software to be	-Representation Information	- Intellectual Entities (environmentFunction, etc)

		reproduced to reproduce the correct behaviour.		
Operating Performance	-The performance of the software with respect to its use of resources (as opposed to its performance in replaying its content) -Plays a significant part in the reproducible behaviour of software. -Contributes towards the information needed to measure the overall adequacy of software	-Speed of execution, data storage requirements. -In some circumstances, we may wish to replay the software at the original operating performance rather than a later improved performance. -A notable example of this is games software, which if reproduced at a modern processor's speed would be too fast for a human user to play.	-Not comprehensively addressed in the OAIS – may be categorized as the Significant Properties of software	-Intellectual Entities (environmentDesignationNote, etc)

If news apps are to be saved, it will have to be a group effort, with buy in from as many institutions as possible. It is conceivable, then, that this work might be performed by a variety of organizations with a range of resources and goals, including public and private universities, large and small cultural and memory institutions, public, non-profit, and commercial news content producers, and so on. Thus it is important to identify a variety of archiving and preservation approaches geared toward different scenarios and stakeholders. The National Digital Stewardship Alliance's description of digital preservation levels is helpful in this regard, as it incorporates considerations for the digital object's safe storage, fixity and data integrity, information security, metadata, and file format compatibility at four levels of increasing preservation status (Phillips, Bailey, Goethals, & Owens, 2013).

Recommendations and Areas for Further Inquiry

Many terabytes of born-digital news content will inevitably be lost to the black hole of technological obsolescence. News apps, which represent some of the most complex, original, and innovative journalism stories being produced today, should not be among them. In recent years news apps have exploded in production and popularity on websites devoted to data journalism, such as FiveThirtyEight, ProPublica, The UpShot, and Vox.com (ProPublica Inc., 2016; Silver, 2016; The New York Times Company, 2016; Vox Media, 2016). Legacy news publications like *The Guardian*, *The Washington Post*, and *The Wall Street Journal* have also dramatically increased publication of data-driven stories over the last five years, and the reach of these projects speaks to their role in shaping our culture (Howard, 2014). It is imperative that a strategy for saving interactive data journalism be established.

A thorough analysis of the nature and landscape of current news apps would be a useful first step in this endeavor. It would centralize some metadata on the projects and provide insights into strategies for capture, archiving, and preservation. The work of saving news apps is a difficult but not insurmountable task, and can benefit from the significant progress being made in the field of software preservation. In particular, newsroom and memory institutions like libraries and archives could investigate emulation as a

preservation strategy, and, in selecting an emulator, should choose one that would comply with the findings of the KEEP project to make the emulator itself migratable ("European Commission : CORDIS : Projects & Results : Keeping Emulation Environments Portable," 2014). The Performance Model Framework for the Preservation of a Software System should be considered as a viable approach in deciding the "significant properties" and boundaries of the digital object to be preserved (Matthews et al., 2010). Furthermore, the PREMIS version 3 metadata schema has the potential to be a fitting schema for thoroughly and accurately describing news apps for the best interoperability and discoverability.

More work in capturing, archiving and preserving news apps is needed, especially case studies that could test and update these recommendations. Documented case studies could then be used to establish best practices and workflows that could be implemented by other institutions at a broader scale. Such pathways will be the key in preventing these invaluable data journalism stories from being "irretrievably lost to future generations," as Rothenberg famously warned (1995, p. 42).

References

Anderson, D., Delve, J., & Pinchbeck, D. (2010). Toward A Workable Emulation-Based Preservation Strategy: Rationale and Technical Metadata. \iNew Review of Information Networking, \i15(2), 110–131. http://doi.org/10.1080/13614576.2010.530132

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

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

Castagné, M. (2013). Consider the Source: The Value of Source Code to Digital Preservation Strategies. \iSLIS Student Research Journal, \i2(2), 48–58.

Consultative Committee for Space Data Systems, B. (2012). *Reference Model for an Open Archival Information System (OAIS)*. Consultative Committee for Space Data Systems. Retrieved from http://public.ccsds.org/publications/archive/650x0b1s.pdf

Dappert, A., Peyrard, S., Chou, C. C. H., & Delve, J. (2013). Describing and Preserving Digital Object Environments. *New Review of Information Networking*, *18*(2), 106–173. http://doi.org/10.1080/13614576.2013.842494

Data Document Initiative Alliance. (2015). Welcome to the Data Documentation Initiative | Data Documentation Initiative. Retrieved April 5, 2016, from http://www.ddialliance.org/

European Commission : CORDIS : Projects & Results : Keeping Emulation Environments Portable. (2014, November). Retrieved March 13, 2016, from http://cordis.europa.eu/project/rcn/89496_en.html

Granger, S. (2000, October). Emulation as a Digital Preservation Strategy. *D-Lib Magazine*, *6*(10). Retrieved from http://www.dlib.org/dlib/october00/granger/10granger.html

Groeger, L., Ornstein, C., Tigas, M., & Jones, R. (2015). 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. http://doi.org/10.1177/0739532915600745

Han, T. (2014, February 19). Re: NICAR news apps archive designathon. *NICAR-L*.

Harris, J. (2013, November 14). And Remember, this Is for Posterity - Learning - Source: An OpenNews project. Retrieved March 27, 2014, from https://source.opennews.org/en-US/learning/and-remember-ones-posterity/

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

IMDb - Movies, TV and Celebrities. (2016). Retrieved April 5, 2016, from http://www.imdb.com/

International Federation of Library Associations and Institutions. (1997). *Functional Requirements for Bibliographic Records* (pp. 1–142). IFLA Study Group on the Functional Requirements fo r Bibliographic Records. Retrieved from http://www.ifla.org/files/assets/cataloguing/frbr/frbr_2008.pdf

Johnston, L. (2014, February 11). Considering Emulation for Digital Preservation | The Signal: Digital Preservation. Retrieved March 30, 2014, from http://blogs.loc.gov/digitalpreservation/2014/02/considering-emulation-for-digital-preservation/

Klein, S. (2012). News Apps at ProPublica. In J. Gray, L. Bounegru, & L. Chambers (Eds.), *The data journalism handbook* (1st ed). Sebastopol, CA: O'Reilly Media. Retrieved from http://datajournalismhandbook.org/1.0/en/delivering_data_2.html

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/

Library of Congress, & PREMIS Editorial Committee. (2016, February 11). PREMIS Data Dictionary for Preservation Metadata, Version 3.0 (Library of Congress) [webpage]. Retrieved March 20, 2016, from http://www.loc.gov/standards/premis/v3/index.html

Matthews, B., McIlwrath, B., Giaretta, D., & Conway, E. (2008). *The Significant Properties of Software: A Study.pdf* (pp. 1–97). United Kingdom: Science and Technology Facilities Council. Retrieved from http://www.webarchive.org.uk/wayback/archive/20100624233431/http://www.jisc.ac.uk/media/documents/programmes/preservation/spsoftware_report_redacted.pdf

Matthews, B., Shaon, A., Bicarregui, J., & Jones, C. (2010). A Framework for Software Preservation. *International Journal of Digital Curation*, *5*(1), 91–105. http://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. http://doi.org/10.1177/0739532915600747

McDonough, J., & Olendorf, R. (2010). *Preserving Virtual Worlds Final Report* (Library of Congress' National Digital Information Infrastructure for Preservation Program). Retrieved from http://hdl.handle.net/2142/17097

National Digital Newspaper Program | Library of Congress. (2015, November 5). Retrieved December 11, 2015, from http://www.loc.gov/ndnp/

National Information Standards Organization (U.S.). (2004). *Understanding metadata.* Bethesda, MD: NISO Press.

Pew Research Center Journalism Project. (2014, March 26). The Growth of Digital Reporting. Retrieved March 26, 2014, from http://www.journalism.org/2014/03/26/the-growth-in-digital-reporting/

Phillips, M., Bailey, J., Goethals, A., & Owens, T. (2013). The NDSA Levels of Digital Preservation: Explanation and Uses. In *Archiving Conference* (Vol. 2013, pp. 216–222). Society for Imaging Science and Technology. Retrieved from http://www.ingentaconnect.com/content/ist/ac/2013/00002013/00000001/art00047

ProPublica Inc. (2016). ProPublica. Retrieved April 5, 2016, from https://www.propublica.org/

Rothenberg, J. (1995). Ensuring the Longevity of Digital Documents. *Scientific American*, *272*(1), 42–47. http://doi.org/10.1038/scientificamerican0195-42

Silver, N. (2016). FiveThirtyEight. Retrieved 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

The National Archives. (2007). PRONOM | Welcome. Retrieved March 18, 2016, from http://www.nationalarchives.gov.uk/PRONOM/Default.aspx

The New York Times Company. (2016). The Upshot. Retrieved April 5, 2016, from http://www.nytimes.com/section/upshot

Von Suchodoletz, D., & Van der Hoeven, J. (2009). Emulation: From Digital Artefact to Remotely Rendered Environments. *International Journal of Digital Curation*, *4*(3), 146–155. http://doi.org/10.2218/ijdc.v4i3.118

Vox Media. (2016). Understand the News. Retrieved April 5, 2016, from http://www.vox.com/

Waite, M. (2013, September 12). Kill All Your Darlings - Learning - Source: An OpenNews project. Retrieved March 27, 2014, from https://source.opennews.org/en-US/learning/kill-all-your-darlings/