Challenges facing the preservation of born-digital news applications

Katherine Boss, New York University Libraries
Meredith Broussard, New York University, Arthur L. Carter Journalism Institute
IFLA News Media Section Conference. Hamburg, Germany, April 2016

News Applications

What is a news application, or "news app"?

- Single data journalism project that is custom-built by the news agency
- From an archiving perspective, boundaries are hard to define

Who is creating them and why should we care to archive them?

- Creation of news apps and interactive data journalism is exploding
- Represent some of the most innovative and complex journalism

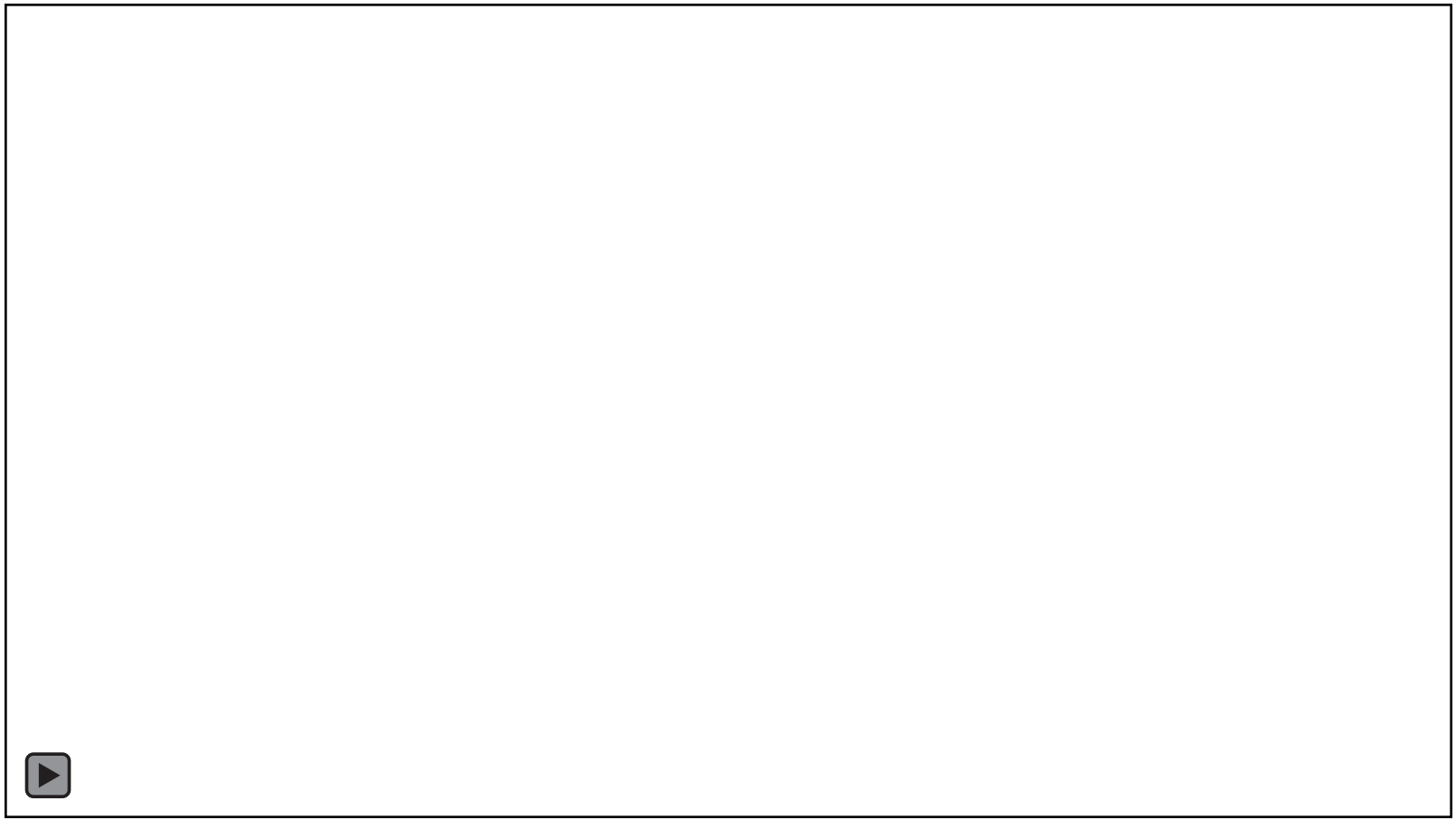
Aldhous, P., & Seife, C. (2016, April 6). Spies In The Sky: Here's Where FBI Planes Are Circling U.S. Cities. Retrieved April 11, 2016, from http://www.buzzfeed.com/peteraldhous/spies-in-the-skies

The Problem

Technical Issues: News apps are dynamic, not static, digital objects

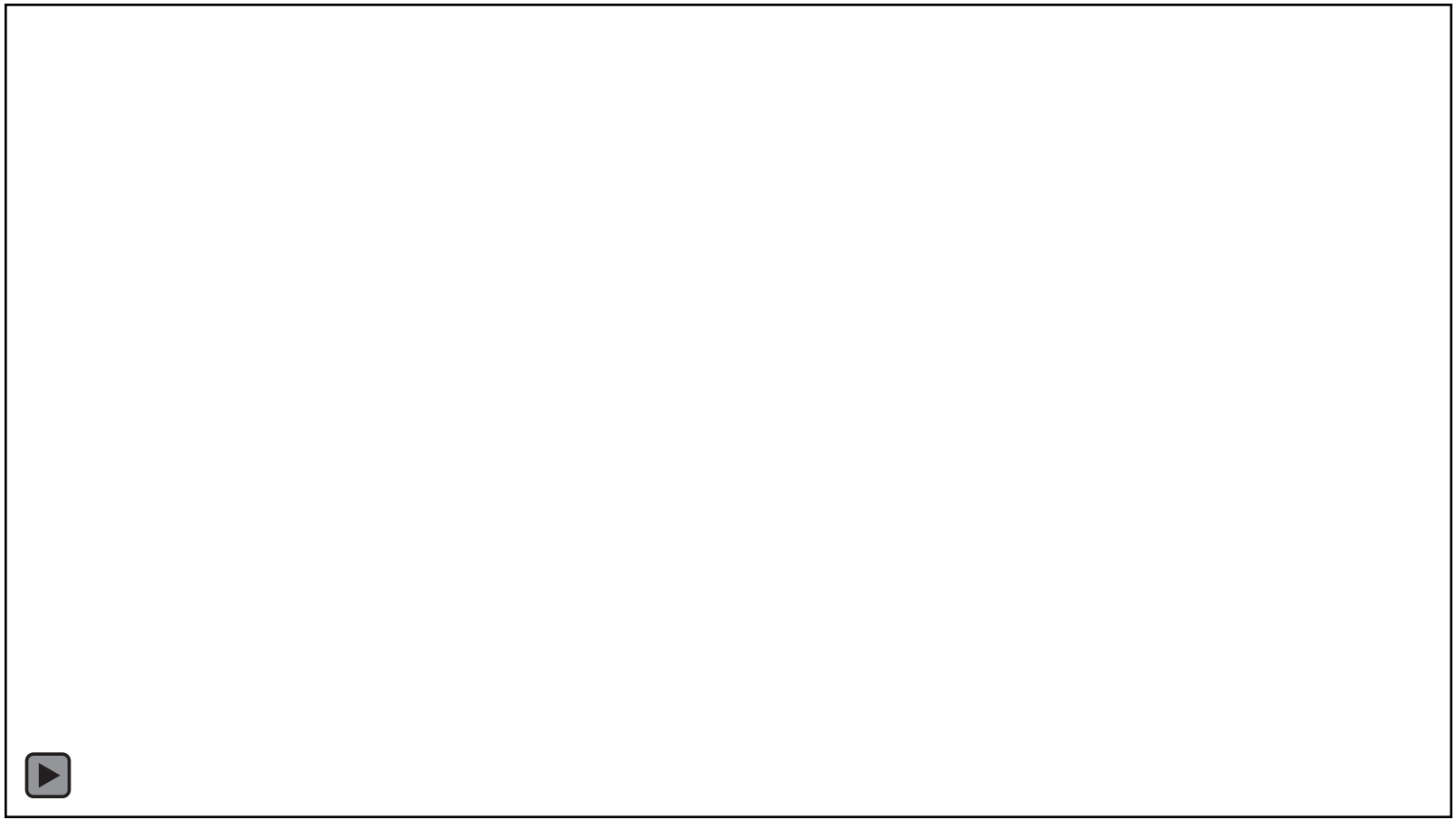
- Complex software dependencies, including the operating system, programming languages, compilers, software libraries, and so on (“dependency hell”)
- Cannot be fully captured through current web archiving
- External APIs (Google Maps etc) present extra challenges

ProPublica – screencast of live Dollars for Docs news app



Groeger, L., Ornstein, C., Tigas, M., Jones, R. G., 1, P. U. J., & 2015. (n.d.). Dollars for Docs. Retrieved December 12, 2015, from https://projects.propublica.org/docdollars/

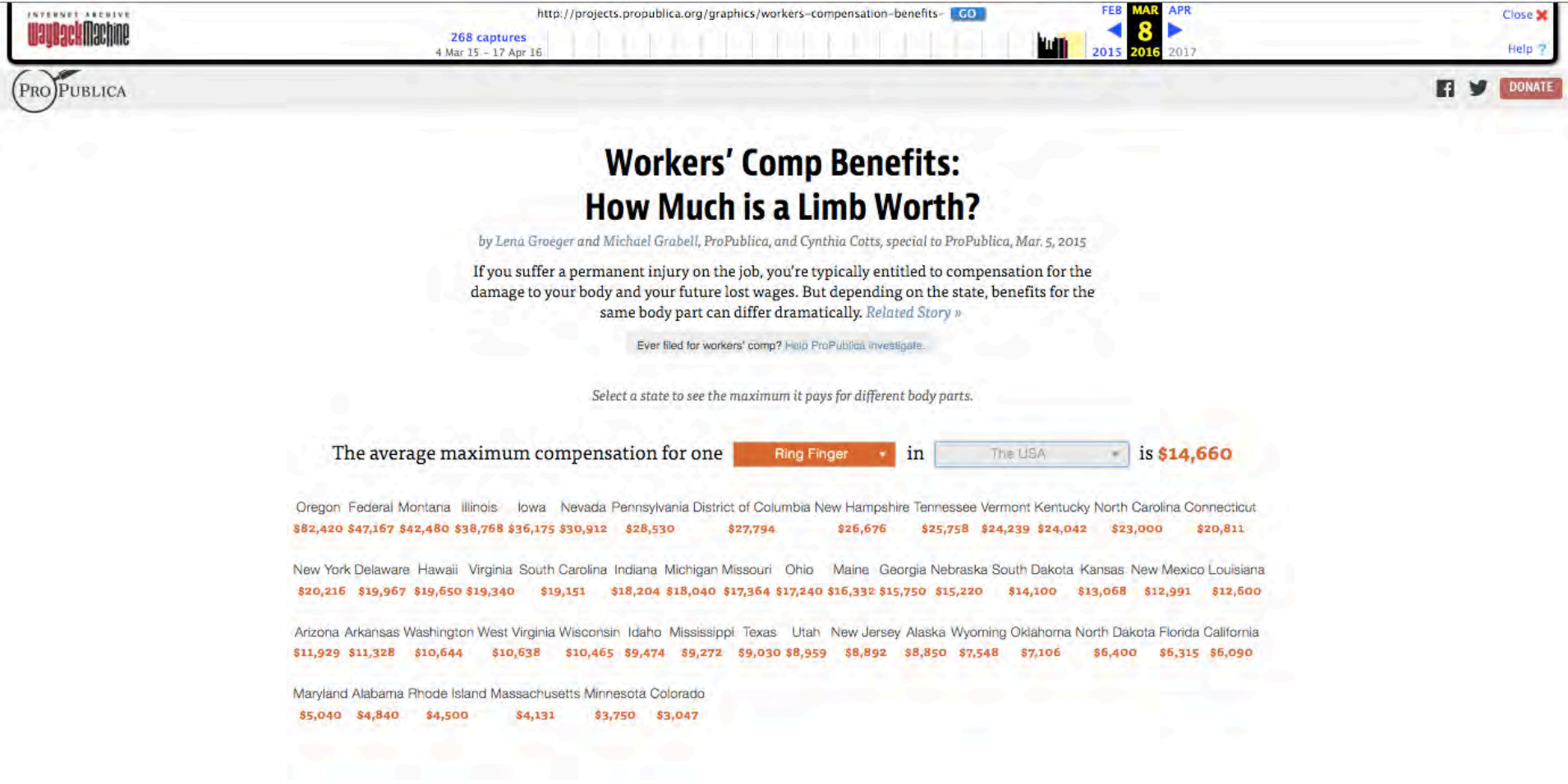
Internet Archive snapshot of Dollars for Docs, Mar. 31 2016



Screencast of Live Workers' Comp Benefits news app

Groeger, L., Grabell, M., ProPublica, Cotts, C., ProPublica, special to, 5, M., & 2015. (n.d.). Workers' Comp Benefits: How Much is a Limb Worth? Retrieved April 21, 2016, from http://projects.propublica.org/graphics/workers-compensation-benefits-by-limb

Internet Archive snapshot of Workers' Comp, Mar. 8 2016



Need a more thorough analysis of what isn’t being captured by current web archiving and why

- Insights into the technical challenges
- Test a spectrum of news apps, including projects using external APIs
- Consider if a new or updated web archiving tool could be developed
- Develop a registry of news apps or produce report on the scan

More Challenges

Conflicting incentives (legal, financial)

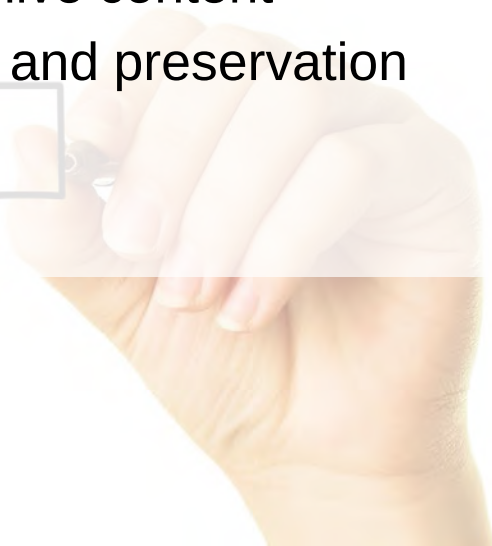
- Proprietary code and digital rights management issues
- News producers want to monetize their archives

Legal Strategies

- Form partnerships with non-profits and others willing to draft agreements
- Push for more open digital rights management

Landscape, Process and Resources

- Poor sense of the amount and type of news apps being created
- No automated workflows in place to capture or archive content
- Human and financial resources for digital archiving and preservation are scarce



Saving News Apps

Determine the “significant properties” of news apps

- Will affect what archiving and preservation solutions to pursue
- May vary from news app to news app
- Considerations for: content, context, rendering, structure, and behavior of the news app

- Must capture the database, the interface, data visualizations, outputs and analysis
- “Look and Feel” of the app as well as the content and functionality

Determine a technique for capture and archiving

- Will depend on the “significant properties,” and what level of human and financial resources can be devoted to the process
- Possible techniques include: technical preservation (computer museums, not very tenable), migration (better for static, not dynamic, digital objects), and emulation
- Digital preservation community currently favors emulation – need consensus on which emulator to use long-term, can migrate the emulator

Adopt a framework for archiving and preservation

- Migration or web archiving framework
- Emulation framework: Performance Model Framework for the Preservation of a Software System (Matthews et al, 2010)

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

Adopt a metadata description schema for best interoperability, discovery and access to news apps

- If web archiving can be a solution, we can use and update the current XML metadata schema
- If emulation must be used, there are several options, including PREMIS from the Library of Congress

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)	News Apps Registry Fields
Functionality	-Description of the typical characteristics of software. -Useful for efficient discovery and accessibility of the software in future	-Description of inputs and outputs -Description of operation and algorithms -Description of the domain addressed	-Descriptive Information	- Intellectual Entities or Object Entities (significantProperties, objectIdentifier, etc)	-Descriptive information on the application for discoverability: the name of the application and a description of its purpose -URL of the news app
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)	- Description of the components that constitute the software - Link to the open source software repository, if available (Github[1], SourceForge[2], Launchpad[3], 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.)	-Rights statements, software owner and license information, the news application development team member names and roles
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	-The inputs which a user enters through a keyboard, pointing device or other input devices, such as	-Not comprehensively addressed in the OAIS – may be categorized as the Significant	- Event Entities (eventDetail, eventOutcome, etc)	-User inputs and application outputs

[1] https://github.com/
[2] https://sourceforge.net/
[3] https://launchpad.net/

Mapping the framework to a metadata schema

- Framework for Preservation of a Software System => PREMIS v. 3
- Includes categories for: Functionality, Software composition, Provenance and ownership, User interaction, Software environment, Software architecture, and Operating performance

Needed Work

Saving News Apps is a group effort. We need:

- Conversations with media organizations about preserving their digital history and legacy
- More digital archivists and web developers
- More grant funding and institutional resources
- Case studies to test and update these recommendations
- Advocacy for more open digital rights management
- Establishment of best practices, workflows

Thank you! Questions?

Katherine.Boss@nyu.edu

MerBroussard@nyu.edu