

**Preserving News and  
Journalism:  
Fundamentals of Born Digital  
News Archiving**

**<http://bit.ly/1klZ4f2>**

**Digital Preservation 2014**

# Introductions

# Panel



Anne Wootton  
Pop Up Archive  
@annewootton



Leslie Johnston  
Director of Digital  
Preservation -  
NARA  
@ljohnston



Edward McCain  
Reynolds  
Journalism Institute,  
University of  
Missouri  
@e\_mccain



Aurelia Moser  
Knight Mozilla Open  
News Fellow  
@auremoser

# Intro to Newsroom Preservation

## Terms:

- What is 'born digital' in a newsroom?
- What are 'news apps'?

# Survey Results

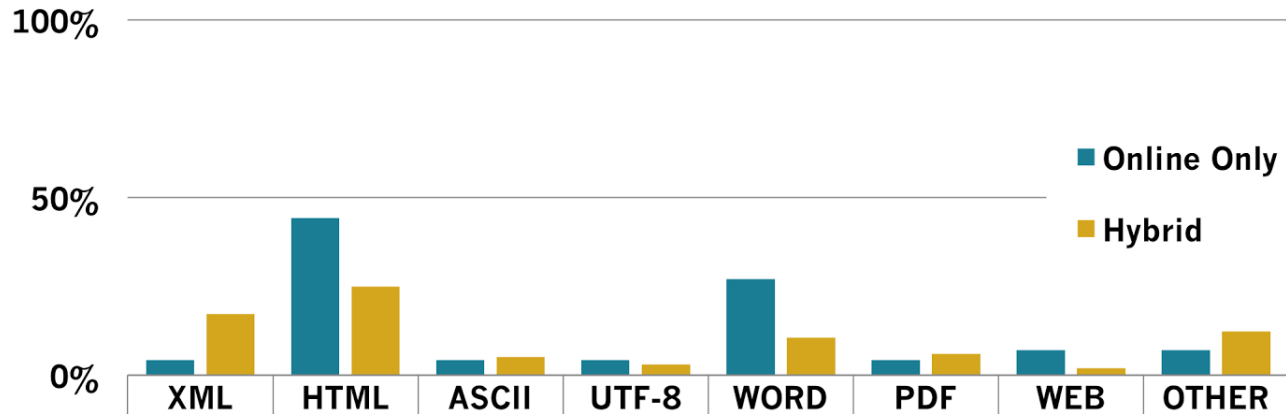


RJI phone survey of 476 news organizations:

- 406 were “Hybrid” enterprises
- 70 were “Online Only” publishers

# Survey Results

## Born-digital news text formats

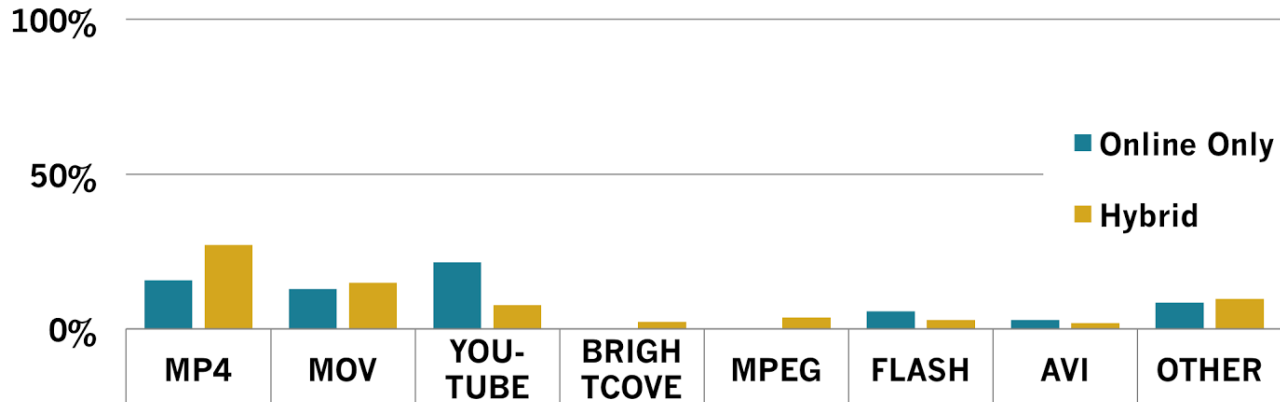


Q1aa: In what format is the BD text content produced?

	XML	HTML	ASCII	UTF-8	WORD	PDF	WEB	OTHER
■ Online Only	4%	44%	4%	4%	27%	4%	7%	7%
■ Hybrid	17%	25%	5%	3%	11%	6%	2%	12%

# Survey Results

## Born-digital news video formats



Q1cc: In what format is the BD video content produced?

■ Online Only	16%	13%	21%	0%	0%	6%	3%	9%
■ Hybrid	27%	15%	8%	2%	4%	3%	2%	10%

# News Apps

Interested in downloading the data? Go to the ProPublica Data Store.

## Has Your Health Professional Received Drug Company Money?

Name  State

Example searches: Klein, Duke University, Miami [More options +](#)

With 2.1 million records, this database represents:

- \$2.5 billion** in disclosed payments
- 15** companies
- ~ 43%** of total market share

### Payments in Your State

Click on a state to see payments made to practitioners and institutions there. See notes below.

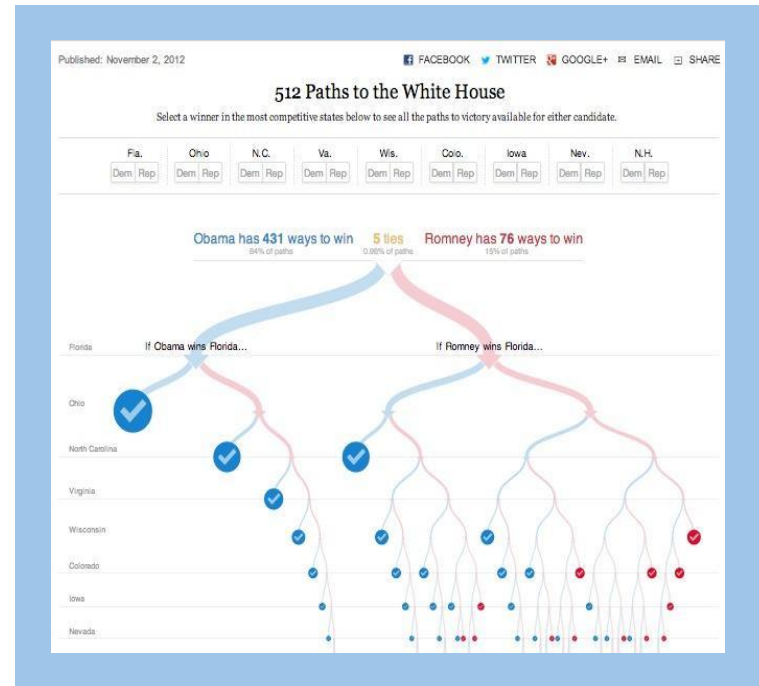
State	Total
Alabama	\$32,352,175
Alaska	\$516,535
Arizona	\$38,739,643
Arkansas	\$15,599,178
California	\$298,077,233
Colorado	\$37,667,949

### Company Disclosures

The totals listed here cover different time periods and spending categories, and aren't directly comparable. See notes below. [See what each company discloses >](#)

Company	Total Disclosed
AbbVie <small>Disclosed: July to Dec., 2012</small>	<b>\$18M</b>
Allergan <small>Disclosed: July 2011 to Dec. 2012</small>	<b>Ranges*</b>

ProPublica | Dollars for Docs | 3/3/14  
<http://bit.ly/1iJD2cq>

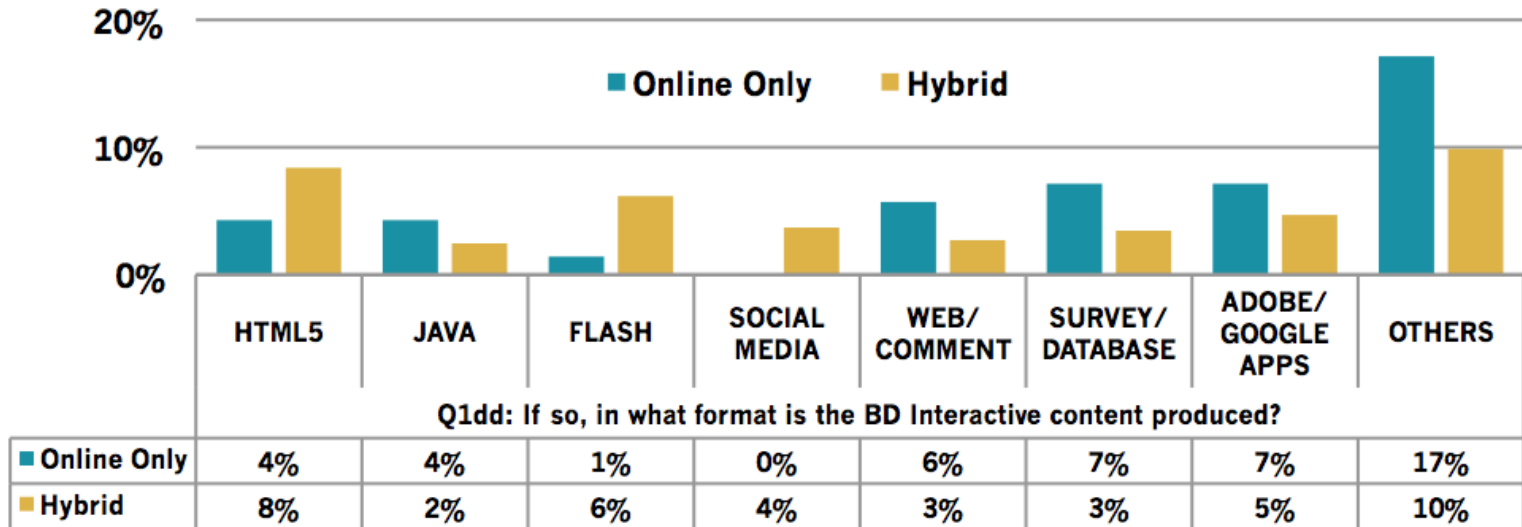


NYTimes | 512 Paths to White House | 11/2/12  
<http://nyti.ms/OIWyxL>



# Survey Results

## Born-digital news interactive formats



# Why are they worth preserving?

- What is the value of newsroom digital preservation?
  - Public Service
  - Posterity

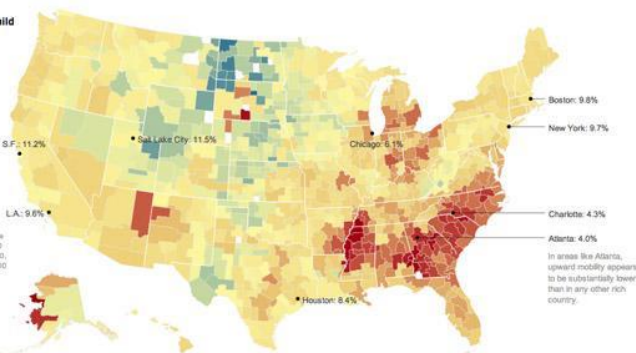
## In Climbing Income Ladder, Location Matters

*A study finds the odds of rising to another income level are notably low in certain cities, like Atlanta and Charlotte, and much higher in New York and Boston.*

The chance a child raised in the bottom fifth rose to the top fifth



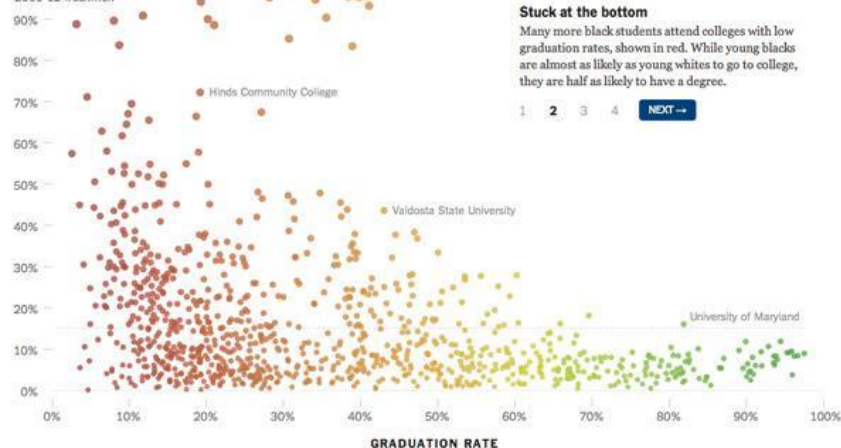
The top 10% is equal to family income of more than \$25,000 for the child by age 30, or more than \$100,000 by age 45.



## At Top Colleges, an Admissions Gap for Minorities

### BLACKS

as a percentage of 2011-12 freshmen



1 2 3 4 NEXT →

**Stuck at the bottom**  
Many more black students attend colleges with low graduation rates, shown in red. While young blacks are almost as likely as young whites to go to college, they are half as likely to have a degree.

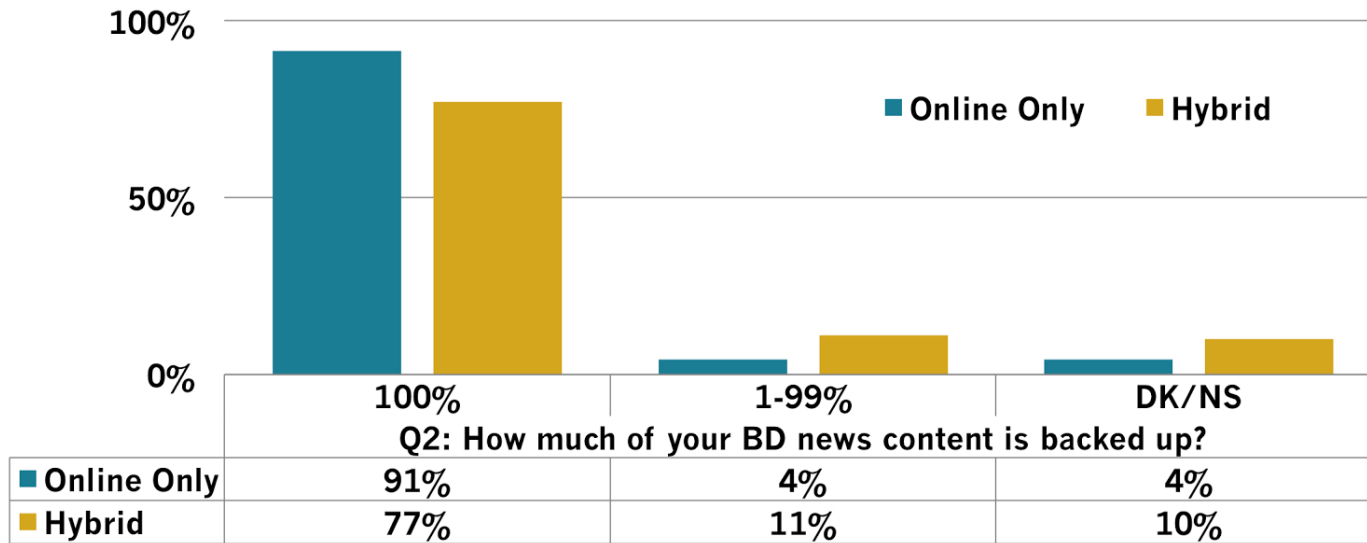
# Intro to Newsroom Preservation

## Concepts:

- What is being archived?
- What are the maintenance challenges?

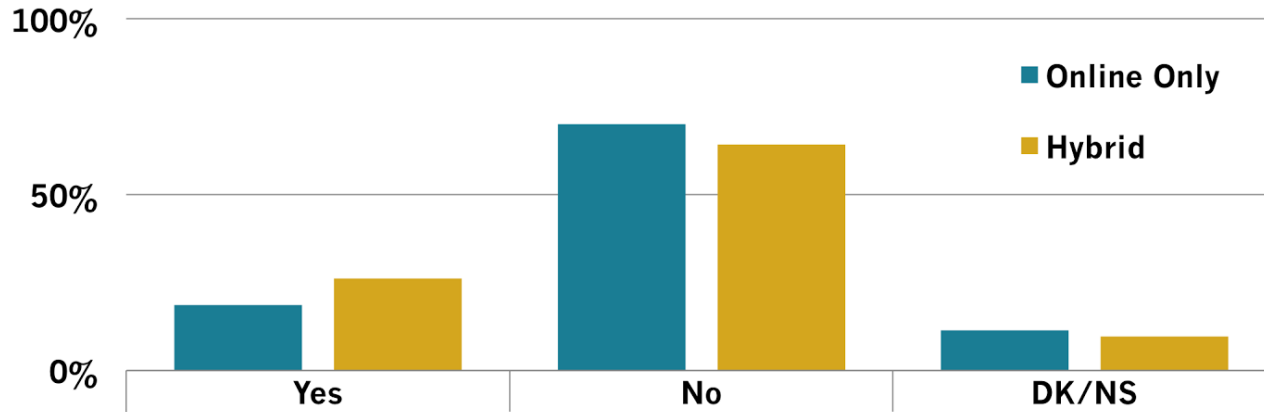
# Survey Results

## Amount of BDNC Backed Up



# Survey Results

## Written policies for BDNC

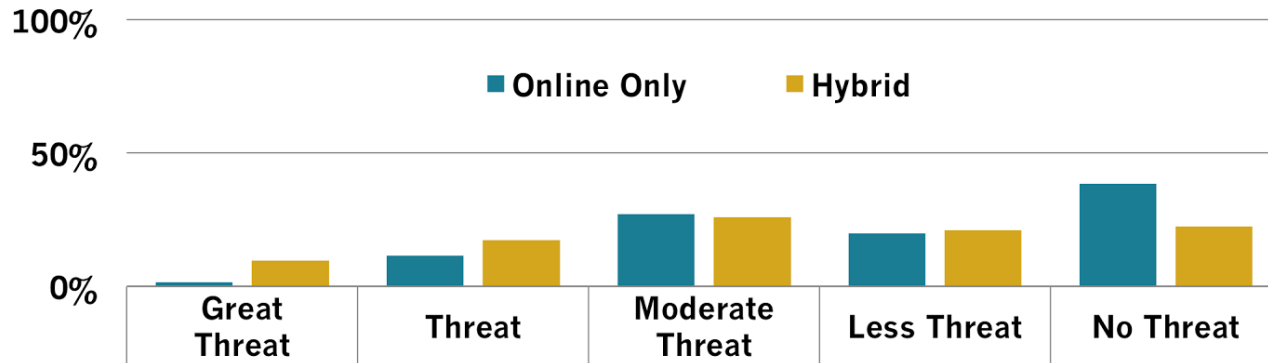


Q8: Does your news org. have written policies for managing BD materials?

■ Online Only	19%	70%	11%
■ Hybrid	26%	64%	10%

# Survey Results

## Insufficient policy as threat to BNDC



Q13c -  
Rank lack of or insufficient policy or plan for preservation as a threat

■ Online Only	1%	11%	27%	20%	39%
■ Hybrid	10%	17%	26%	21%	22%

# Current Status

# Preservation Practice

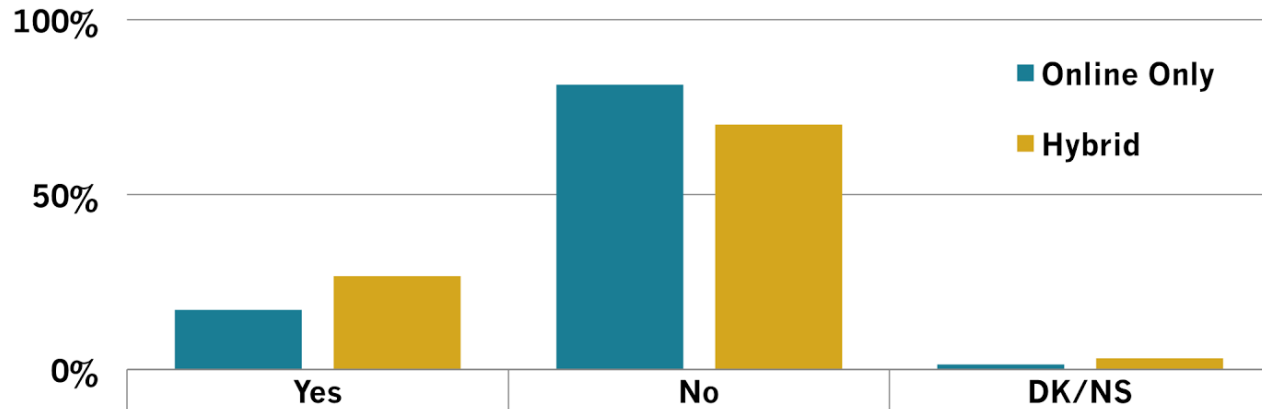
- What is the current approach to archiving in newsrooms?
  - Survey Results
  - Anecdotal





# Survey Results

## Significant Loss of News Content

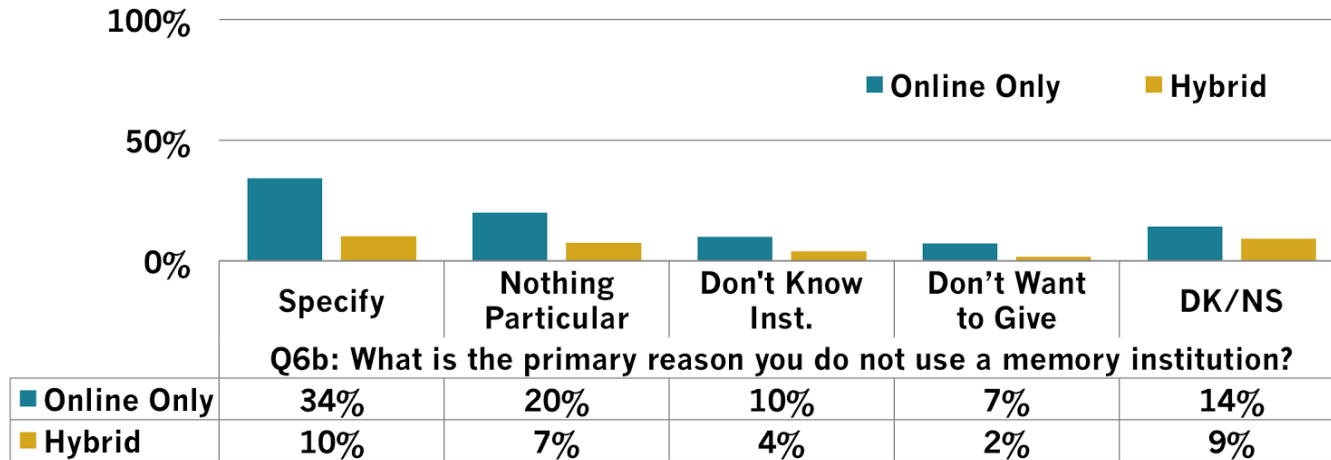


Q5c: Have you ever experienced a significant loss of news content?

■ Online Only	17%	81%	1%
■ Hybrid	27%	70%	3%

# Survey Results

## BDNC not going to memory institutions



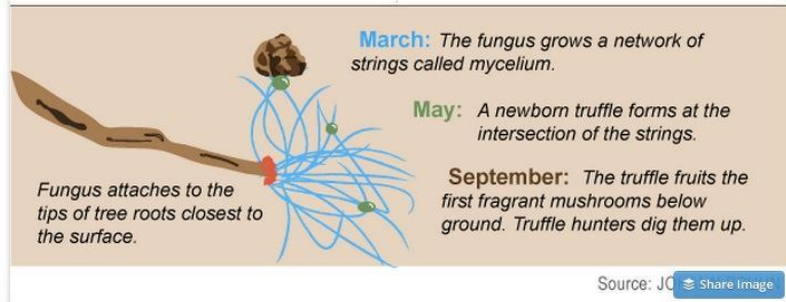
# What are large newsrooms doing?

## News Nerd First Projects

It's okay. We all sucked once. Curated by Tyler Fisher.

[Submit your first project!](#)

## The life of a truffle



This was my first attempt at an infographic, my first time using Adobe Illustrator, and my first time ever

Imgs: <http://bit.ly/1nAeCnL>  
<http://timesmachine.nytimes.com/>

The New York Times

TimesMachine

129 years of New York Times journalism, as it originally appeared.

WEDNESDAY, JULY 22, 1925

89 YEARS AGO

LONG STEP TO PEACE IS SEEN BY  
BRITAIN IN GERMANY'S REPLY

FINAL SCENES DRAMATIC

EVOLUTION BATTLE RAGES OUT  
OF COURT

Text of Germany's Reply to French  
Note

MORE IN THIS ISSUE

Hylan, Jno Francis  
United States Navy  
N Y C  
Sports  
Briand, Aristide

SEE FULL INDEX



SELECT AN ISSUE:

[VIEW IN TIMESMACHINE](#)

# What are small newsrooms doing?



ProQuest Historical  
Newspapers™



Brooklyn Public Library



BROOKLYN DAILY EAGLE  
1841-1902 Online™

Imgs: <http://bit.ly/1wZBZXo>



# Proposed Practice

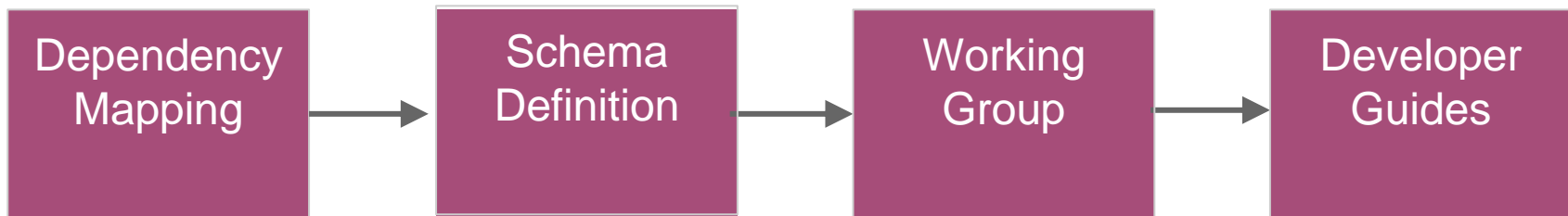
# Designathon Plan

Develop a preservation program for newsrooms



Img: Anne Wooton

# Some Outcomes



News Apps Model : Sheet1

Category	Artifacts	Attributes	Archival Requirement
Code	Front-end software, Back-end software, Data Code, Internal libraries	Actors, Documentation, Transformations, Versions, Decisions	Preserve
Data (Input)	Clean data, Raw data (copyright?), Metadata, Data Structure (field layout, data dictionary), UGC data, Live Data APIs, Reporting material,	Actors, Documentation, Transformations, Versions, Decisions	Preserve
Story (Output)	Narrative story, Links, Images/Audio/Multimedia, APIs we publish, UX, Viz Design, IA/Hierarchy/Taxonomy, Annotations, API Documentation	Actors, Documentation, Transformations, Versions, Decisions	Preserve
Infrastructure	The Internet, Browser, Server (OS, Language, Framework), Display APIs (external), Vendor libraries and dependencies, Platform (external hosting), Hardware, RDBMS/NoSQL, Bandwidth	Actors, Documentation, Decisions, Versions	Simulate, Emulate, Virtualize
Process	Code documented, Code "History" (git), Data Transformation, Data documentation, Cultural Context and Moment, Design Patterns, Edits of Story, Data sources (FOIA letters)	Actors, Documentation, Transformations, Versions, Decisions	Preserve
Response	Comments, Metrics, Awards, User Behavior, Logs, Inbound Links, Reaction/Media Coverage, Tweets, Impact, Intermediate Outcomes	Actors, Decisions	Preserve

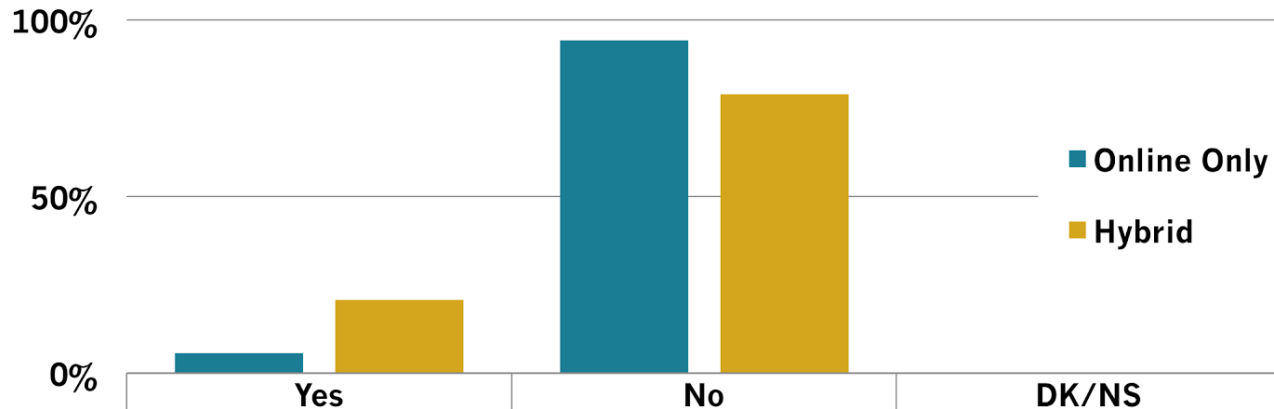
# Best/Better Practices

- Recommendations to news apps developers?
- Schema for cataloging and classifying news apps?
- Resource describing apps and their dependencies?



# Survey Results

## Does org. have a news librarian?



Q15c: Do you have a news librarian or equivalent position in your newsroom?

■ Online Only	6%	94%	0%
■ Hybrid	21%	79%	0%

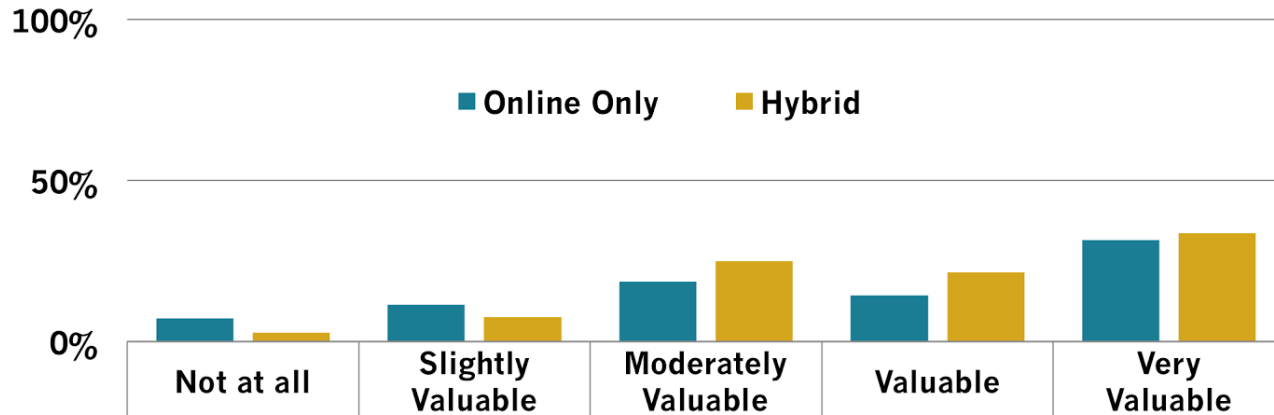
# Best/Better Practices

- Engaging partner communities in the process?
- Monetizing the archive?



# Survey Results

## Value of archives for ROI



Q12d: How valuable is having access to your archives for producing good RIO?

■ Online Only	7%	11%	19%	14%	31%
■ Hybrid	3%	8%	25%	21%	34%

# Flexible Practice

# Flexible Practices

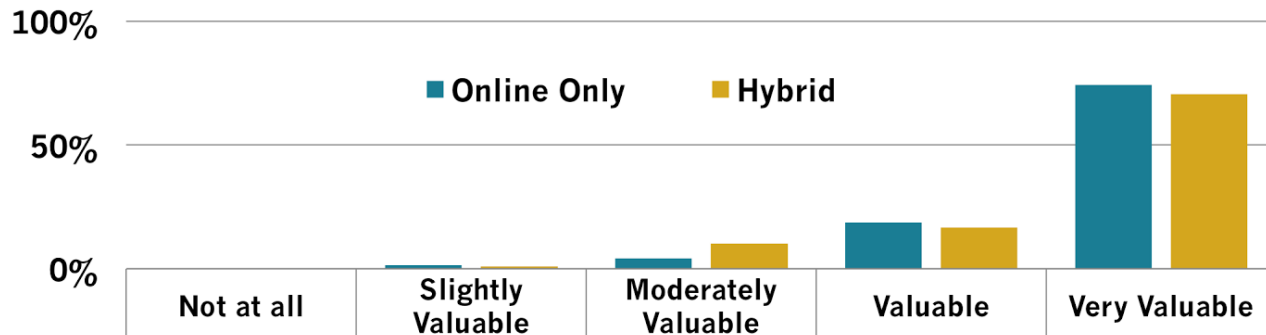
- What is an archive?
  - Values?
  - Threats?



Img: Result for “archive” keyword search in the Noun Project, correlates with “overworked”  
<http://bit.ly/1pc4T5j>

# Survey Results

## Value of archives for quality journalism



Q12c: How valuable is having access to your archives for producing quality journalism?

■ Online Only	0%	1%	4%	19%	74%
■ Hybrid	0%	1%	10%	17%	71%

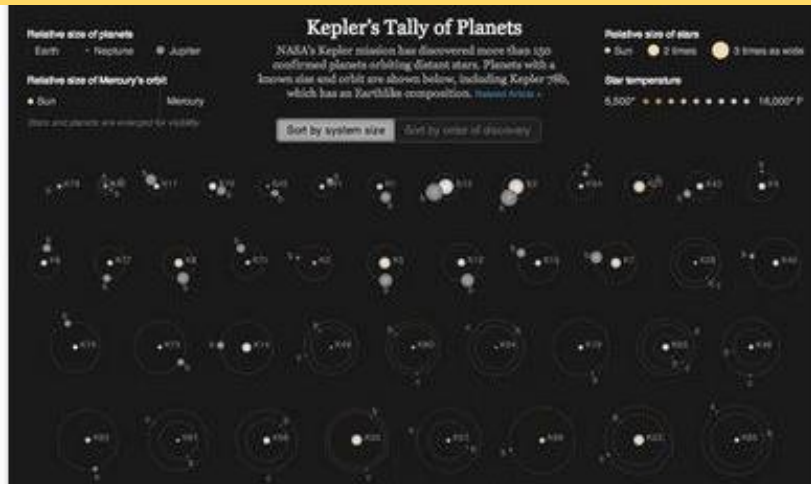
# Institution-Level Scaling

- Can we accommodate all papers and preservation needs?



Img: <http://bit.ly/1rAmBTB>

# Institutional Cases



**STATE'S SECRETS** *A cache of diplomatic cables provides a chronicle of the United States' relations with the world.*

LATEST IN THE SERIES: Jan. 3, 2011: Diplomats Help Push Sales of Jetliners on the Global Market

[More on WikiLeaks: The War Logs >](#)

**BROOKLYN Newsstand**

About | Brooklyn Collection | Support | Photo Search

Search Browse Clippings Sign-in

**THE BROOKLYN DAILY EAGLE.**

FOUR O'CLOCK. PUBLISHED BY NEW YORK, WEDNESDAY, DECEMBER 30, 1902.—VOL. 62, NO. 341.—22 PAGES. THREE CENTS.

WARSHIPS SEIZED BY ALLIED FLEET. FUSION UNDER GOOD: LOW TO GO TO NERVALLES. SAYS BRIBERY WAS TRIED TO BREAK COAL STRIKE. CHANGE OF POTASSIUM FOUND IN LEY'S BEER.

Search a keyword or name + Add more info

This screenshot shows the Brooklyn Newsstand website interface. At the top, there is a navigation bar with the site name and links for About, Brooklyn Collection, Support, and Photo Search. Below this is a search bar with options for Search, Browse, Clippings, and Sign-in. The main content area displays a newspaper clipping from The Brooklyn Daily Eagle, dated Wednesday, December 30, 1902. The clipping features several headlines, including 'WARSHIPS SEIZED BY ALLIED FLEET', 'FUSION UNDER GOOD: LOW TO GO TO NERVALLES', 'SAYS BRIBERY WAS TRIED TO BREAK COAL STRIKE', and 'CHANGE OF POTASSIUM FOUND IN LEY'S BEER'. At the bottom of the page, there is a search bar with the text 'Search a keyword or name' and a '+ Add more info' button.

large newsroom  
<http://nyti.ms/1h9NkB0>

small newsroom  
<http://bklyn.newspapers.com/>



# Tools of Entrapment

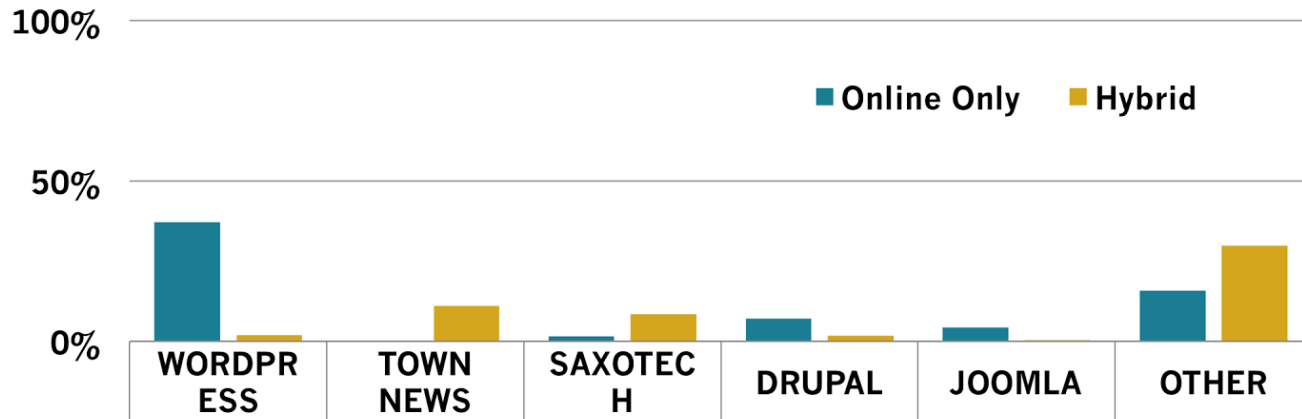
- News orgs rely heavily their CMSes
- Need public/private partnerships for buy in around better preservation practice
- NYTimes Scoop



Img: <http://bit.ly/1kbuxZ2> | <http://nyti.ms/1rAUvHF>

# Survey Results

## CMS types for BDNC



Q4bb0th: What specific CMS system do you use?

■ Online Only	37%	0%	1%	7%	4%	16%
■ Hybrid	2%	11%	8%	2%	0%	30%

# Value of the Archive

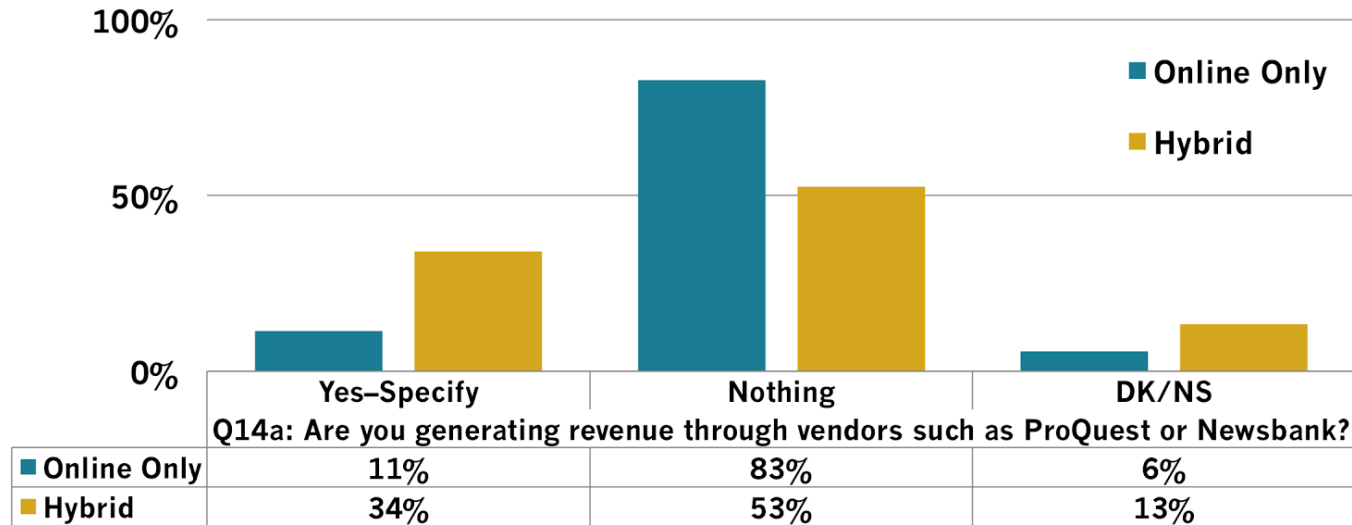
- How are newsroom archives quantifiably valuable?
- To what extent is monetization an option in the newsroom?



Img: <http://bit.ly/1nejEoE>

# Survey Results

## Revenue generation from vendors



# Future Practice

# Future Practice - Larger Papers

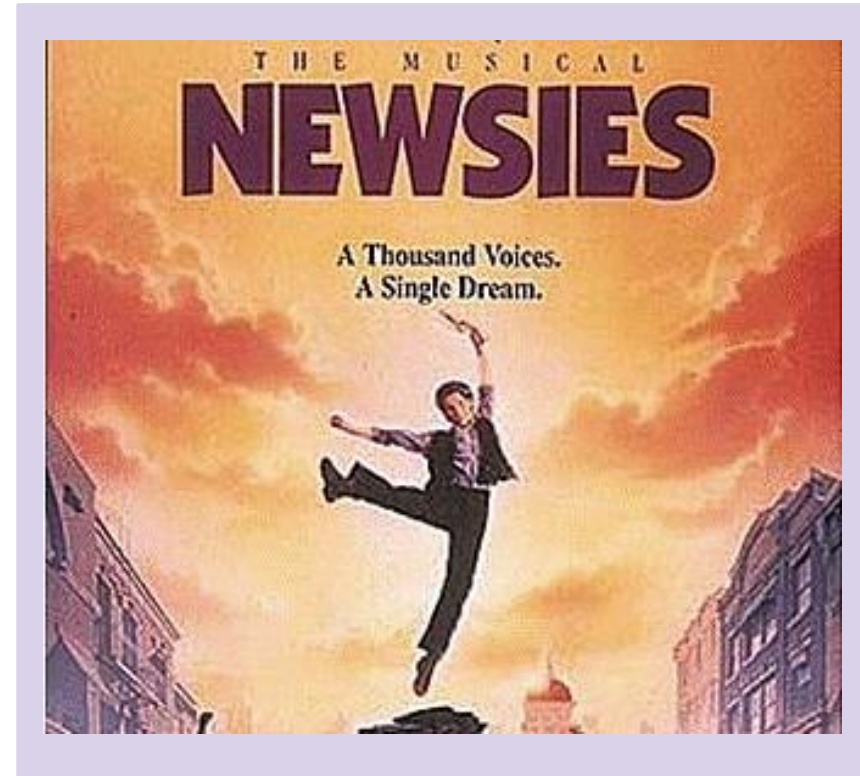
- Future media to preserve?
- Preserving leaks and breaches? Snowden files, Yanukovych leaked documents?

**FREEDOM  
= OF THE PRESS =  
FOUNDATION**

# Future Practice - Smaller Papers

- News co-ops for smaller papers?
- Enabling different media archives: audio, tv, broadcast
- Partnerships

Img: <http://bit.ly/1jUUw6Q>



# Questions + Takeaways

- Designathon Takeaways:
  - Wiki: <http://mzl.la/1pbwnYG>
  - Blogpost: <http://bit.ly/1nBMAbD>
- Pop Up Archive:
  - Blogpost: <http://bit.ly/1kbUisq>
- Survey Results:
  - Blogpost: <http://bit.ly/1wwlEt4>
- Next Steps ...



# Thank you!



Anne Wootton  
@annewooton



Leslie Johnston  
@ljohnston



Edward McCain  
@e\_mccain



Aurelia Moser  
@auremoser