Production and Distribution of Electronic News: Technology and Preservation Challenges

Megan Waters Bernal
mbernal2@depaul.edu
June 27, 2013

Global Resources Roundtable: Beyond the Fold: Access to News in the Digital Era
Newberry Library, Chicago, IL
Background

2005-2009: Director of Information Services at The Miami Herald Media Company

- News research
- Intranet, editorial, & archiving systems
- Print & digital archives
- Content licensing & syndication

Print Trends*

Source: Newspaper Association of America (NAA) Trends & Numbers
Composition of Newspaper Media Total Revenue Stream*

<table>
<thead>
<tr>
<th></th>
<th>2011</th>
<th>2012</th>
</tr>
</thead>
<tbody>
<tr>
<td>New Revenue</td>
<td>7%</td>
<td>8%</td>
</tr>
<tr>
<td>Circulation</td>
<td>26%</td>
<td>27%</td>
</tr>
<tr>
<td>Digital Ads</td>
<td>10%</td>
<td>11%</td>
</tr>
<tr>
<td>Print Newspaper</td>
<td>49%</td>
<td>46%</td>
</tr>
<tr>
<td>Nich Mktg.</td>
<td>8%</td>
<td>8%</td>
</tr>
</tbody>
</table>

Source: NAA American Newspaper Media Industry Revenue Profile 2012
Recent NAA Data

Recent Audience Trends
(% of Total Adults)

- Past 7 Days: Print/E-edition/Website
- Typically Watch Local news/evening
- Pure Print, Past Week
- Typically Watch Local news/-late
- Typically Watch National/network news
- Local Early Evening News (M-F) 6 - 6:30pm
- Newspaper Mobile - Past Month
- Newspaper Mobile Exclusive - Past Month
- All News Radio (M - S Cume)
- News/Talk/Information Radio (M-S Cume)
“[Library access and] preservation activities are built around the print newspaper, with its tangible outputs and regular publishing cycles. Electronic distribution requires other, more complex preservation models.”

CRL Mapping the Newspaper Industry, April 27, 2011, p. 42.
CRL Mapping the News Industry Report
Purpose & Methodology

• Outcome of 2009 National Digital Information Infrastructure & Preservation Program (NDIIPP) workshop at LoC.

• Explore strategies for collecting & preserving digital news on a national basis, including:
  – digital newspaper web sites
  – tv and radio broadcasts distributed via the Internet
  – blogs
  – pod casts
  – digital photos and videos

• Examine, analyze, and document workflows at 4 newspapers representing a broad segment of the U.S. newspaper industry (e.g. local, regional, national).

Key Questions from NDIIPP

• What is the nature of the *electronic facsimile*?

• How profound is the relationship between Web and print news “coverage”?

• What is the technical formatting and delivery of electronic news outputs?
Findings

- There are many **actors & systems** used to produce news content, licensed content is brought into news pages or sites and locally-produced content is licensed “out” to a plethora of syndication services and vendors.
- **Standards** are employed & **rich metadata** of interest to libraries exists for some types or levels of content *in certain enterprise news systems*.
- Print editions function as **fixed, static records** of interest to local constituencies whereas news web sites function as **regional portals** to a much larger and more dynamic realm of changeable content.
- The **proliferation of platforms, devices, and distribution networks** used to **read or access the news** further complicates our ability to fully archive the user experience.
- Common methods for **Web archiving** like those employed by the Internet Archive are **insufficient** due to the **rate of change and site architecture** of common news web sites, among other factors.
Actors & Content

Content sources
- Wire services
- Staff writers
- Photo agency
- Freelancers

Decision-making
- Editors
- Parent Company
- Publisher

File output
- Ads
- Text
- Art/Graphics
- Photos
- Video
- Audio
- Tabular data
- Databases
Typical News Organization
Systems Overview

Input

- Editorial System
- Archive

Output:
- Pagination System to printers, facsimile
- Web Production System to Web hosts, servers
- Third Party System to Web hosts, servers
Print/Web Content Overlap

- Online content, including entire stories, is often absent from the print edition
- Print content is often absent online
- The timing of publication varies between print and Web
- Multiple versions of articles appear online, but rarely in print
Brief News Standards Timeline

1965
• IPTC Founded

1970-80s
• IPTC 7901/ANPA-1312 is a 7-bit news agency text markup specification
• Still in use!

1990s
• Binary formats
• Information Interchange Model (IIM), first multimedia news exchange format aka “IPTC Fields”
• News Industry Text Format (NITF) XML specification

2000s
• NewsML
• IPTC Core Extension for use with Adobe XMP for photo metadata
• News Codes controlled vocabularies/taxonomies

2010s
• NewsML-G2
• Newscodes.org for linking
• rNews emerging standard for embedding structured metadata into HTML documents
Submission & Ingest of Content
Photo & Multimedia Flow
User Web Submission
Pagination to eFacsimile
Text-Specific Flow Out to Aggregators
News Web Site Architecture*

- Complex content management systems drive site architecture & content
- Site architecture is dense with 105 “gateway” pages
- Much subnavigation out of the gate: 5% of pages are linked to/from +1 gateway page
- 366,290 links monitored; average 735 per day added
- Most were advertising-driven
- Highly dynamic, unpredictable over time: No master list of articles tied to original gateway hierarchy
- Roughly 39% are linked for 1 day, 5% for 30 minutes or less

Characteristics by System/Process

**Editorial**
- Tied to newsroom processes & practices (assignments, story folders)
- Highly standards-based depending on system, staffing
- Often includes high resolution or raw original files
- Includes unpublished material

**Archive**
- Can be tied to newsroom processes or considered “outside”/unimportant to producing the news
- Highly standards-based
- May include “enhancing” of data
- Important as the record of everything published, rights management, licensing, but not “how we got the story”
Characteristics by System/Process

Pagination

- Tied to production print and ereader processes
- Highly standardized but oriented towards print publication needs
- Component parts are fused together, treated as a whole (e.g. the page)
- Richer file types and metadata are stripped out

Web

- Content comes from many sources and systems
- Less standardized
- Uses a diversity of file types which may or not be merged
- Platform & content licensing may occur at corporate level
- Richer file types and metadata are stripped out
- Metadata added is SEO-oriented
Reading Experience

• Users may access content under an umbrella brand like *The Chicago Tribune* via an Internet Browser, third party provider/site such as LexisNexis or Twitter, and e-reader devices
• The content they see may be an efacsimile of the print edition or not, it may also be certain, licensed material like staff writers only or only certain content types like text or video
• Web-based content is lower resolution and has less metadata attached
• With print advertising dropping and readership trends changing, more resources will flow to online operations and revenue streams
Where is the “high impact point of entry” for libraries in this flow?
Thank You!

Megan Waters Bernal
mbernal2@depaul.edu
http://www.linkedin.com/in/meaux/