PRESERVING NEWS IN THE DIGITAL ENVIRONMENT: MAPPING THE NEWSPAPER INDUSTRY IN TRANSITION

A Report from the Center for Research Libraries

April 27 , 2011

This report is an initial attempt to map the “lifecycle” for news content and information published in newspapers and online; and to clarify the relationship between the news content produced for those two major distribution channels.

The report includes two sections:

1. *Background and methodology*

2. *An overview of news workflow and systems*, organized around the three major stages in the lifecycle of news information

CRL will continue to refine and update the maps and lifecycle information contained herein, adding detail, particularly on technical processes that, because of the proprietary nature of much of the technology, we were not able to obtain within the timeframe of the study.

Contributors to this report were:

Jessica Alverson
Kalev Leetaru
Victoria McCargar
Kayla Ondracek
James Simon
Bernard Reilly

Illustrations were produced by Eileen Wagner, Eileen Wagner Design, Chicago IL.

TABLE OF CONTENTS

SECTION 1: BACKGROUND AND METHODOLOGY

The decline of the newspaper industry, combined with the ascent of digital media for news reporting and distribution, means that merely preserving newspapers and traditional broadcast media will no longer ensure future access to a comprehensive journalistic record. News production is no longer the periodic, linear process with a single, fixed output that it was in the era of the printed newspaper. It is now a continuous loop of news gathering, processing, versioning, output, response, and update. Therefore devising effective strategies for preserving news in the electronic environment requires an understanding of the “lifecycle” of news content.

Fundamental changes in this lifecycle are now occurring as major newspapers re-orient their operations and distribution channels from paper and printing to digital environments and platforms. How the news that appears in newspapers and on the Web is sourced and reported, how it is edited and processed, and how and in what forms it is distributed must all be taken into account when framing library action to prevent loss of this important class of historical evidence.

The Library of Congress has a vested interest in addressing this challenge. LC has been a key link in the newspaper supply chain for the major US academic and independent research libraries. This role is tied to LC’s copyright registration, acquisition and microfilming operations, and to its overseas operations. Today, understanding what processes and formats play a role in typical digital production workflows might provide the basis for a new approach to preserving electronic news at a national level

The present report grew out of a workshop convened by the Library of Congress National Digital Information Infrastructure and Preservation Program (NDIIPP) in September 2009 to explore possible strategies for collecting and preserving digital news on a national basis. For purposes of discussion at that forum LC defined digital news to include, at minimum, “digital newspaper Web sites, television and radio broadcasts distributed via the Internet, blogs, pod casts, digital photographs, and videos that document current events and cultural trends.”

Prompted by those discussions CRL proposed to examine, analyze, and document the flow of news information, content, and data for four major newspapers from production and sourcing, through editing and processing, to distribution to end users. Four newspapers provided a test bed for the project: *The Arizona Republic, Seattle Post-Intelligencer* (since 2008, *seattlepi.com*), *Wisconsin State Journal*, and *The Chicago Tribune*. The test bed newspapers were chosen to provide a variety of types of newspapers that represent a broad segment of the U.S. newspaper industry and its publishers. The report also draws from CRL analysis of three other news organizations: *The New York Times*, *Investor’s Business Daily*, and the Associated Press.

Our study addressed the following points of interest:

1. *The nature of the electronic facsimile:* LC was particularly interested in determining how one particular news format, the "electronic facsimile" of the printed newspaper, might be exploited for copyright registration and archiving purposes.

2. *The relationship between Web and print news "coverage":* LC wanted to understand how content posted to each newspaper's Web site compares with the contents of the paper edition.

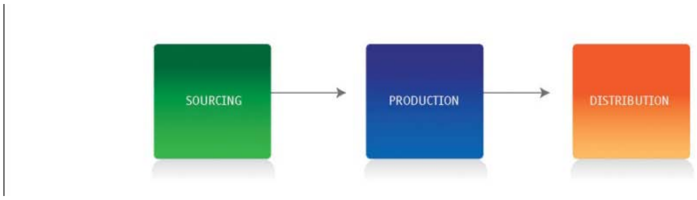
3. *Technical formatting and delivery of electronic news output:* To what extent are there extant or emerging industry technical standards for formatting, managing and/or disseminating news content? And what is the range of current practices for formatting and encoding news for distribution in print and on the Web?

Section 2 of this report provides an overview of news sourcing, production and distribution based on the practices of the test bed newspapers. Note is made where an individual newspaper deviates significantly from this generalized narrative.

The authors hope that this report can help pinpoint the potential "high-impact point of entry" in this workflow where libraries and other memory organizations can capture critical news content and metadata and ensure the long-term survival and accessibility of the American journalistic record.

SECTION 2: AN OVERVIEW OF NEWS WORKFLOW AND SYSTEMS

The production processes and workflows set up to produce newspapers in print are still very much at the center of the operations of news organization like *the Arizona Republic* and *Chicago Tribune*. These activities have always been built around daily and weekly news delivery cycles. Broadcast, and more recently the Web, however, has made possible a continual, accelerated flow of information and news, which is now approaching real time reporting. As a result the lifecycle of news is changing profoundly.



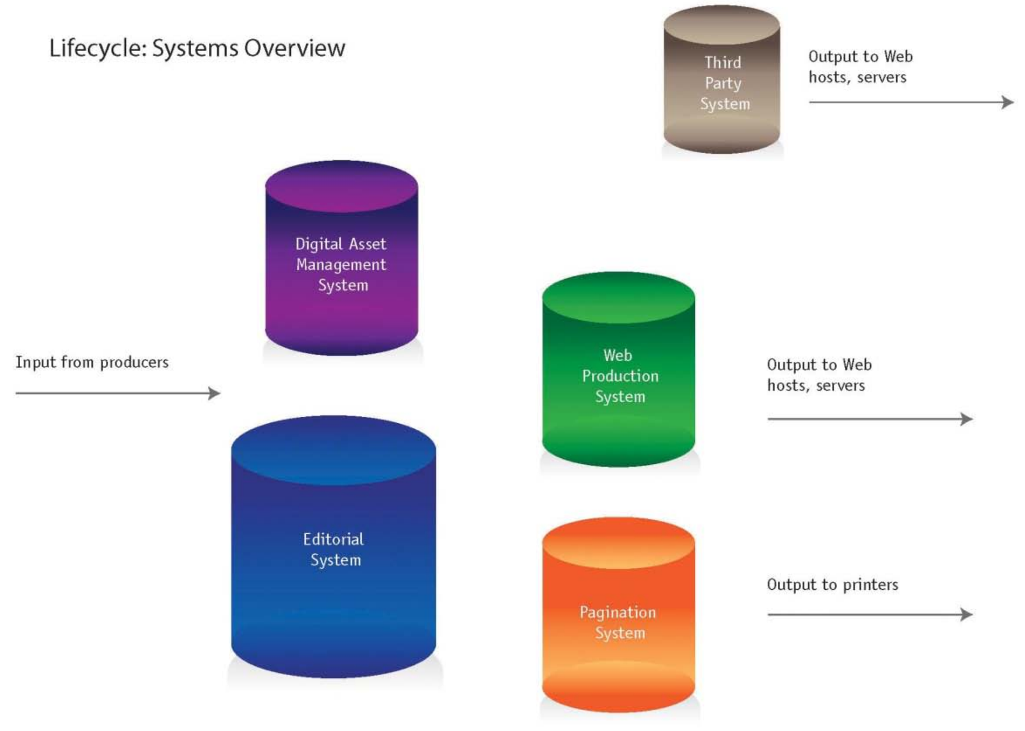
For purposes of the report we characterize the stages of this lifecycle as:

A. *Sourcing:* the gathering of news information and content by the news organization from those who create, report, and/or own that information and content;

B. *Editing and production:* the editing, processing, and enhancement of news content and information, and preparation of outputs for distribution through various media.

C. *Distribution:* dissemination and exposure of news content, and products and derivatives thereof, through print and online media.

Essential to each stage of this lifecycle are automated systems deployed to produce, modify, and annotate content and prepare the products of the newsgathering process. The major systems involved are editorial, digital asset management or archives, pagination, and Web production systems. In addition third-party providers, working in tandem with news publishers, employ their own systems to produce and deliver news content through the publisher's Web and print channels.


Lifecycle: Systems Overview
Third Party System
Output to Web hosts, servers
Digital Asset Management System
Input from producers
Web Production System
Output to Web hosts, servers
Editorial System
Pagination System
Output to printers

A. SOURCING NEWS CONTENT

Content/data:

- *News wire reports*
- *Article, feature texts*
- *Graphics*
- *Photographs*
- *Audio recordings*
- *Video recordings*
- *Tables*
- *Databases*
- *News object metadata*
- *User data*
- *Algorithms and programming code*

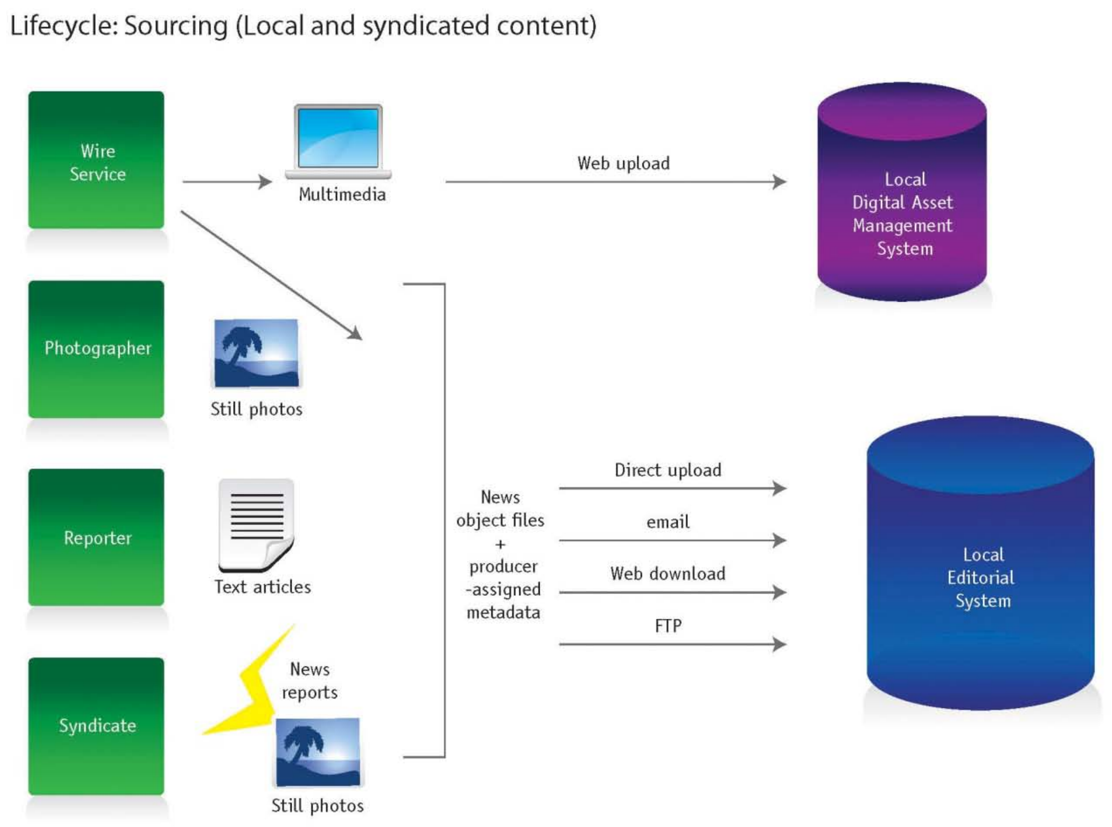
Actors:

- Standards organizations (such as IPTC, NAA)
- Newspaper publishers (*Arizona Republic*, *Seattle Post Intelligencer*, *the Tribune Company*, *Wisconsin State Journal*)
- Parent organizations (Gannett, Hearst Newspapers, Tribune Media)
- Affiliated newspapers and news organizations (print, broadcast)
- Editors
- Staff reporters, photographers, columnists, illustrators, cartoonists
- Independent, freelance reporters, photographers, columnists, cartoonists/artists
- Wire services (AP, Agence France-Presse)
- Syndicates and other content providers (King Features, Washington Post Writers Group, TV Guide)
- Photo agencies (Corbis, Agence France-Presse)
- Data and polling services (Dow Jones, National Weather Service, Nielsen Company, Major League Baseball)
- Advertising agencies
- Ad and data servers

The contents of a given newspaper issue or news Web site are produced by a host of individuals and organizations. Some content is produced by reporters, photographers, columnists, illustrators, and others under the auspices of the newspaper publisher (local content). Other content is supplied by wire services, photo agencies, data services, syndicates, affiliated newspapers, independent or freelance producers, and by parent organizations of the newspaper publisher (syndicated content). The main difference between sourcing for the print newspaper and for the Web news site is that for the latter considerably more content, and more processing of that content, is supplied by third party commercial service providers, such as advertising agencies or "ad servers," and Google Earth (third-party content).

In addition news websites, more than newspapers, make use of information and other content provided by readers (user-created content) either directly through web platforms like blogs, polls, and other forums controlled by the news organization, or indirectly through third-party social media sites like YouTube, Twitter, and FaceBook.   This proliferation and diversification of sources and ingest channels presents complicated workflow and rights management challenges to news organizations, and to those who seek to preserve news.

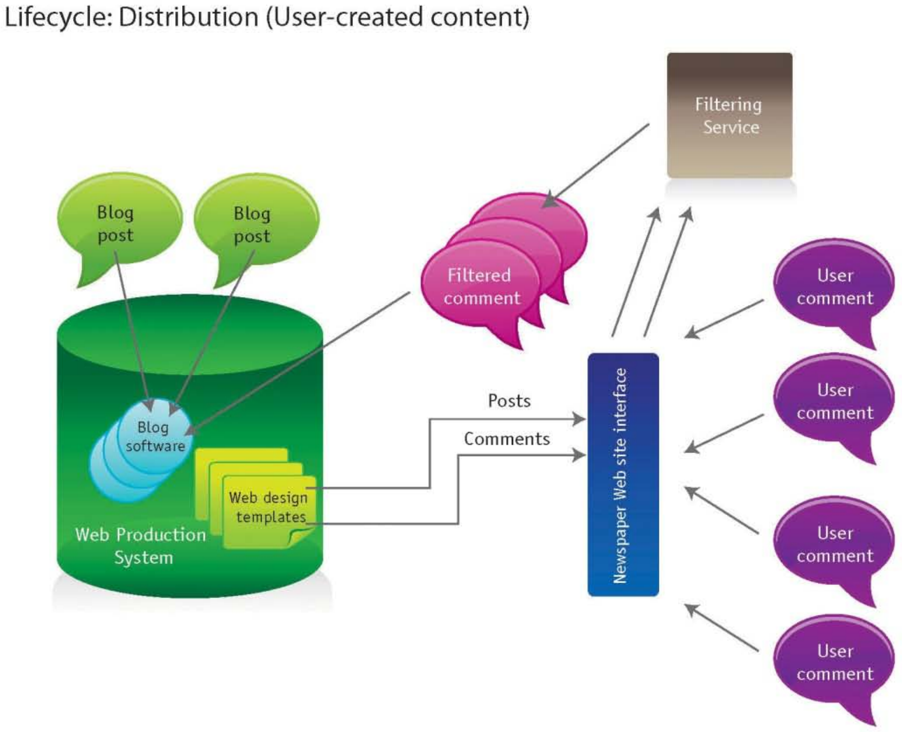
A.1 SUBMISSION AND INGEST OF CONTENT



Content is submitted to newspaper publishers by a multitude of individuals and organizations, under varying sorts of arrangements with the publisher. These include reporters and photographers, wire services, advertisers, and syndicates. Content is submitted by these parties directly to the newspaper publisher in a number of ways:

- Mail, courier and in-person delivery of hard copy and digital copy on portable media from reporters, photographers, columnists, artists, readers, and other contributors
- Email text and attachments
- Input of locally produced content to publisher’s production system directly by reporters, photographers, columnists, artists, and other authorized contributors
- FTP from syndicates and other proprietary suppliers
- Web download from syndicates and other proprietary suppliers
- Web upload through blogs and comment postings on the publisher’s web site.

Some publishers permit citizen journalists and readers to submit news stories and photographs by email and direct Web upload. The direct upload systems normally have a submission form and an SQL back-end for data storage. User-supplied content is often mediated by social media developers like Pluck, which screens and filters reader comments submitted to *The Arizona Republic* and other Gannett newspaper Web sites. (Pluck checks for abusive comments made in discussion forums, story chats and blogs on a 24/7 basis.)



Commercial advertising services and direct mail houses deliver page image files either to the publisher's production system or directly to the publisher's printer. Some third-party data and content providers, such as advertising, weather, and financial reporting services simply deliver URLs, Java Scripts and metadata to the newspaper publisher's production system that then "point to" images, videos, and other multimedia content residing on the provider's own specialized media servers.

A.1.1 SUBMISSION OF TEXT CONTENT

News text submissions to newspapers come in two general forms:

1) *News and syndicated text feeds* – Brief text reports of breaking news produced by syndicates and wire services like the Associated Press, Reuters; and discrete items like articles, columns, editorials and other writings produced by syndicates, other news organizations, and providers.

2) *Locally produced text* – Discrete articles, columns and other written texts, in a form intended for publication as stories, columns, letters, blog posts, or other element s of the newspaper or website.

Most editorial systems accept these texts in a variety of proprietary and open file formats, including plain text (.txt), Unicode UTF-8 (.utf8), rich text format (.rtf), Word (.doc and .docx), ASCII (.ascii), and email (Outlook .msg is the most common). Older news production systems like *NewsDesk* accept texts only in rich text (.rtf) or plain text (.txt) formats.

A.1.1.1 NEWS AND SYNDICATED TEXT FEEDS

Wire services, syndicates, and other providers of texts to newspaper publishers transmit bundles or packages of texts (feeds) that can consist of a single brief news report or a structured story or column. These texts are normally rigidly structured and transmitted to publishers in one of a number of standard markup formats designed specifically for news interchange. These markup formats include older standards like IPTC 7901[1] and ANPA 1312[2] and newer XML-based formats, including:

[1] IPTC 7901was one of the first news text exchange formats, introduced by the International Press Telecommunications Council (IPTC) in the early 1980s but is still in use by some wire services and news organizations. It was intended to facilitate international exchange of news reports, and takes into account technical and linguistic differences between countries and accommodates numerous alphabets.The format structures text transmissions into four sections: pre-header information, a message header, message text, and post-text information. These format elements are separated by control characters: SOH (start of header), STX or ETX (start of text), and terminated by EOT (end of text). More on IPTC 7901 at: http://www.iptc.org/site/News_Exchange_Formats/IPTC_7901/

[2] ANPA-1312 is a 7-bit news agency text markup protocol designed to standardize the content and structure of text news articles. It was created in the late 1970s by the American Newspaper Publishing Association (now the Newspaper Association of America),. Last modified in 1989 it is still a common method of transmitting news to

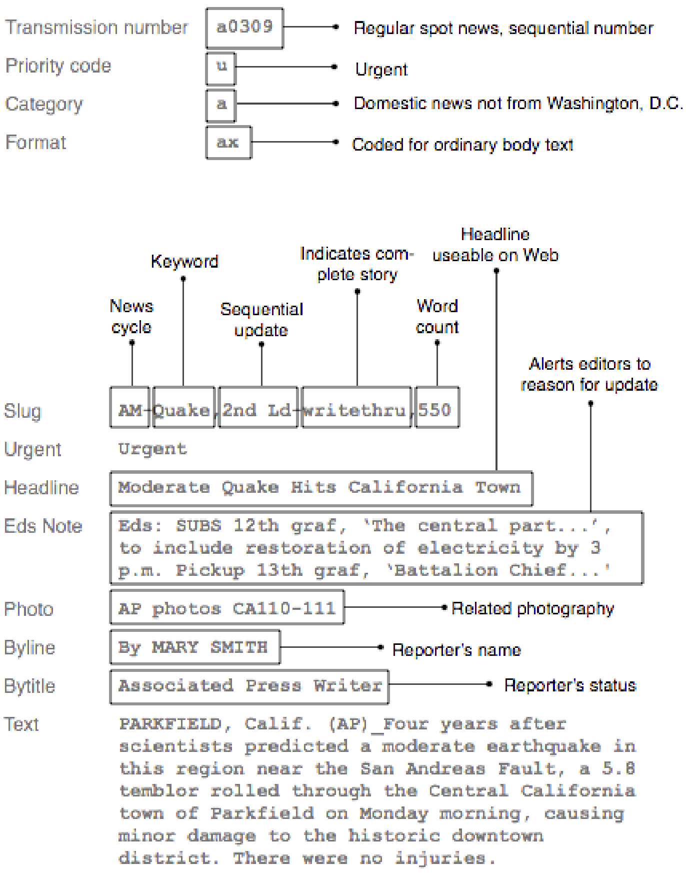
- News Industry Text Format (NITF), for text news items and articles [3]
- NewsML, for text, photo and multimedia news content, i.e., packages of text with accompanying photographs, databases, video, audio and other media[4]
- SportsML, for transmitting sports scores and data, and
- EventsML, a standard for event-related news information.

IPTC 7901, ANPA and NITF are exchange formats originally developed for structuring newspaper text content for transmission, and are used in marking up simple news item and article texts only. They are descendant from the codes developed by the Associated Press for delivery of text via dedicated newswire networks. The Associated Press coding is shown in the annotated text illustrated below.

newspapers, web sites and broadcasters from news agencies in North and South America. Using fixed metadata fields and a series of control and other special characters, ANPA 1312 was originally designed to feed text stories to both teleprinters and computer-based news editing systems.

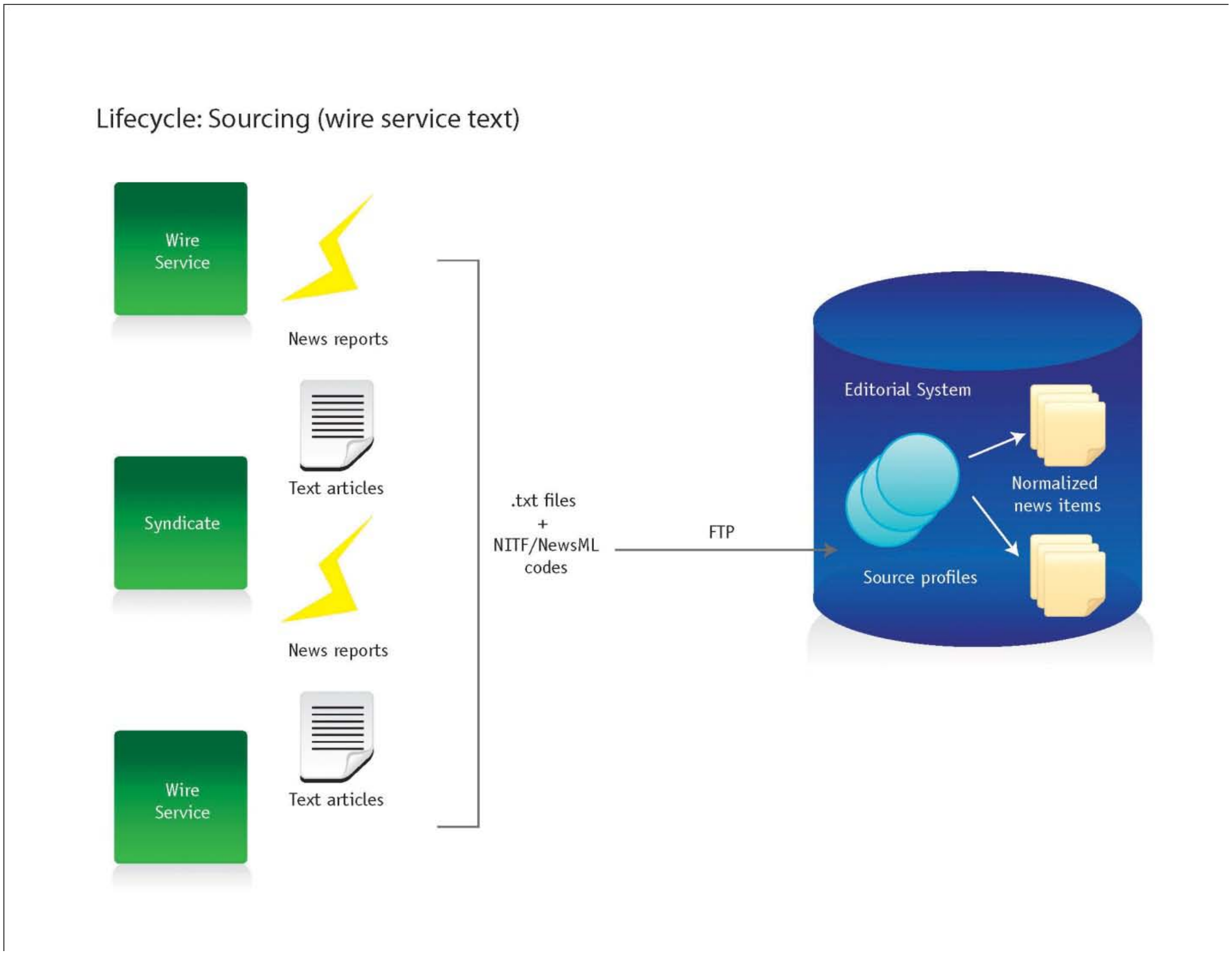
[3] News Industry Text Format (NITF), developed by the International Press Telecommunications Council (IPTC) in the 1990s, uses XML to define the content and structure of news items and articles. Because NITF metadata is applied throughout the news content, NITF documents are far more searchable and useful than HTML pages. NITF enables publisher systems to automate formatting of their documents to suit the bandwidth, devices, and personalized needs of subscribers. NITF-encoded documents can be translated into HTML, WML (for wireless devices), RTF (for printing), and other common formats used by news publishers. For further details on NITF see: http://www.iptc.org/site/News_Exchange_Formats/NITF/Introduction/

[4] NewsML was first developed in 1999 and updated in 2008 to "represent and manage news throughout its lifecycle." It was designed to provide a media-type-independent, structural framework for multi-media news. Beyond exchanging single items NewsML can also convey packages of multiple items in a structured file. It supports the representation of electronic news entities such as news-items, parts of news-items, collections of news-items, relationships between news-items and metadata associated with news items. For further details on the NewsML schema see: http://www.iptc.org/std/NewsML/1.2/specification/schema-doc/NewsML_1.2.html#Link06456060


Transmission number
a0309
Regular spot news, sequential number
Priority code
u
Urgent
Category
a
Domestic news not from Washington, D.C.
Format
ax
Coded for ordinary body text
Headline useable on Web
Keyword
Indicates complete story
News cycle
Sequential update
Word count
Alerts editors to reason for update
Slug
AM-Quake,2nd Ld-writethru,550
Urgent
Urgent
Headline
Moderate Quake Hits California Town
Eds Note
Eds: SUBS 12th graf, 'The central part...', to include restoration of electricity by 3 p.m. Pickup 13th graf, 'Battalion Chief...'
Photo
AP photos CA110-111
Related photography
Byline
By MARY SMITH
Reporter's name
Bytitle
Associated Press Writer
Reporter's status
Text
PARKFIELD, Calif. (AP)_Four years after scientists predicted a moderate earthquake in this region near the San Andreas Fault, a 5.8 temblor rolled through the Central California town of Parkfield on Monday morning, causing minor damage to the historic downtown district. There were no injuries.

NewsML™ and APPL (Associated Press Publishing Language) are newer, media-type agnostic news exchange standards that enable news organizations to provide a wealth of descriptive, structural and administrative information on transmitted news objects (e.g., reports, articles, photographs) that can be read by automated production systems at recipient newspapers. They enable production systems, for example, to manage articles, photographs and other content over time by providing information on rights status (publishable, embargoed, etc.) and information on authorship and copyrights (source, credit line, terms of use, etc.). They also enable news agencies and publishers to generate the same text in multiple languages; video clips in different formats; and the same photograph in different resolutions. (AP delivers news items in ANPA, NITF and APPA formats.) [5]

[5] The IPTC web site contains a wealth of information and vocabularies on the various text markup formats, at http://www.iptc.org/cms/site/index.html?channel=CH0112



The XML-structured data file for an NITF or NewsML news item includes tags that identify the various components of an item text, such as headline, byline, dateline, lead, body text, column headings, and so forth. The file can also incorporate additional tagged administrative information, such as level of urgency, embargo period, release date, version date, usage rights, and references to accompanying images. For example, a version of an article originally reported on November 25, 2008 but updated in 2010 would have the following NITF date tags:

```
<date.issue norm="20081125T143000-0500" />
<date.release norm="20100226T093000-0500" />
```

And the following NewsML date tags:

```
<contentCreated>2008-11-25</contentCreated>
<versionCreated>2010-10-19T16:25:32-05:00</versionCreated>
```

Normally the XML tagging will also include general news categories ("business," "weather," "sports," etc.), relevant geographical location(s), and subject keywords.

XML-structured content enables the automation of some processes within a newspaper's production system. Most production systems are designed to read the XML tags and to translate the structured data into functionality on various platforms. For example, event-related data structured in EventsML interacts with style sheets and templates in Web-oriented editorial systems like *NewsGate* and web production systems like Nstein's WCM to generate -- and automatically update -- event listings on the news Web site, and even to "purge" those listings when the embedded "end" or expiration dates arrive. Although the major news content sources employ their own proprietary variations on the standard IPTC tags and codes (Reuters, for instance, has its own "flavor" of NewsML) , the major editorial systems in use by the newspapers are able to translate most of these variations to their local codes automatically, through the use of profiles developed for each major news source. Other newspapers use third party data formatting services like *Fingerpost* to convert wire service article texts to their own preferred format before ingest. [6] (The *Chicago Tribune* uses *Fingerpost* to convert wire service feeds to ANPA format.)

Therefore, at minimum, texts transmitted by news agencies and wire services using IPTC/ANPA codes include information regarding the transmission standard used, date and time of transmission, author (if a feature or article), source of transmission, urgency of information, word count, and general news category. Texts marked up in newer, XML formats are more richly tagged and often include document type, rights and release-related information, version information, subject key words, date and place of origin, background on the story or image,and references to accompanying materials.

A.1.1.2 LOCALLY PRODUCED TEXT

Newspaper production systems accommodate unstructured texts as well. Locally produced text, i.e., text produced by staff or assignment writers, reporters, columnists, editors, and others is often created in common word processing applications and then uploaded by the authors or editorial staff directly to the newspaper's production system. In some cases article and feature texts are authored within the production system environment itself, using Microsoft Word or comparable, system-proprietary word processing software.

[6] Fingerpost, a UK-based software producer produces Fingerpost Information Processor (FIP), a service for http://www.fingerpost.co.uk/

For these texts the production system generates a certain minimal amount of metadata on ingest, such as assignment ID, author name, usage rights, date and time, general news category, and urgency, as well as subject information and references to accompanying materials such as photographs and charts.

> Therefore, at minimum, locally-produced texts submitted to, or authored within, a newspaper’ s editorial system will include information regarding the date and time of ingest, assignment or article ID, author, urgency of information, word count, and general news. Some texts are more richly tagged, with rights and release-related information, version information, subject key words, date and place of origin, and references to accompanying materials.

A.1.2 PHOTOGRAPHS AND OTHER NON-TEXT MEDIA

Individual producers and contributors of image and multimedia content to newspapers employ a variety of digital capture and recording devices. These include:

- Digital still photo cameras
- Video cameras
- Digital audio recorders
- Hybrid still/video cameras
- Cameras built into cell phones and PDAs

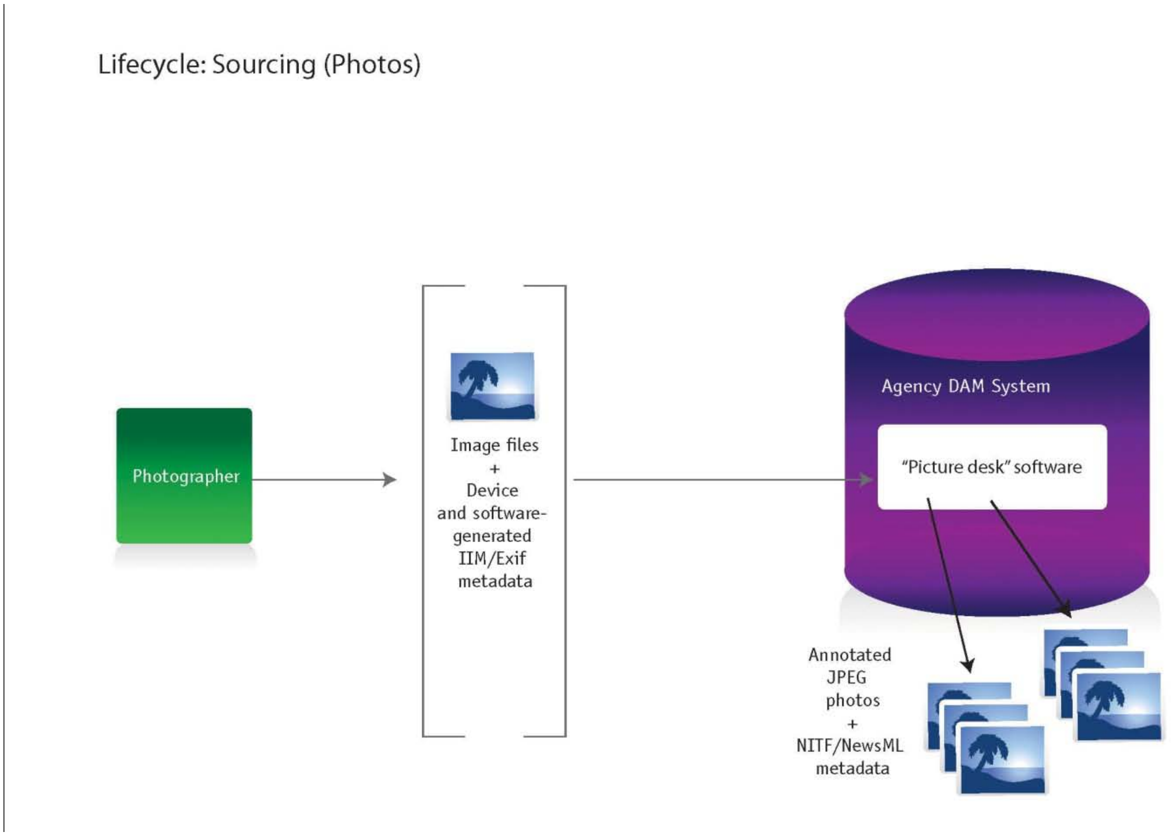
These devices employ a variety of proprietary and open formats and applications. Still photo and graphic image file formats accepted by the major news production systems include: JPEG, JPEG2000, TIFF, RAW, EPS, PNG, GIF, and PDF as well as numerous proprietary formats.[7] Newer newspaper

[7] According to Jon Swerens, *The News Sentinel’s CCI Cheat Sheets*, the *NewsDesk* editorial system accepts images in the following formats: Adobe Photoshop, Apple PICT, AWD Fax Format, BMP, Canon Raw Format (CRW), Casio Digital Camera, Commodore-Amiga IFF, Computer-aided Acquisition and Logistics Support (CALS), DCX, FAX (Brooktrout, LaserData and WinFax), FlashPix, Foveon X3F, FujiFilm FinePix S2, Pro RAF, Graphics Interchange Format (GIF), JPEG, JPEG 2000, Kodak Photo-CD, Minolta MRW, Kodak Professional Digital Camera, Leica Digilux 2 RAW, Nikon NEF, Olympus ORF, Paint Shop Pro Image, Panasonic DMC-LC1 RAW, Pentax *ist D PEF, Portable Bitmap, Graymap, and Pixelmap (PBM, PGM, PPM), PCX, Portable Network Graphics (PNG), Psion MultiBitMap, Raw Format, Seattle Film Works, SGI Image File, Sony DSC-F828 SRF, Sony Digital Camera, Sun Raster, TARGA, TIFF, WAP WBMP, Windows Icon, Windows Metafile, Windows XP Thumbs.db, X Bitmap, X PixMap, X Windows Dump.

content management systems, designed for Web production, also accommodate multimedia file formats such as WAV, MP3, Flash, MPEG3, others. In many instances (the *Chicago Tribune*, for example*)* content in video, audio and other multimedia formats bypass the newspaper's editorial system and is ingested directly into the Web production system.

A.1.2.1 PHOTO METADATA AND MARKUP

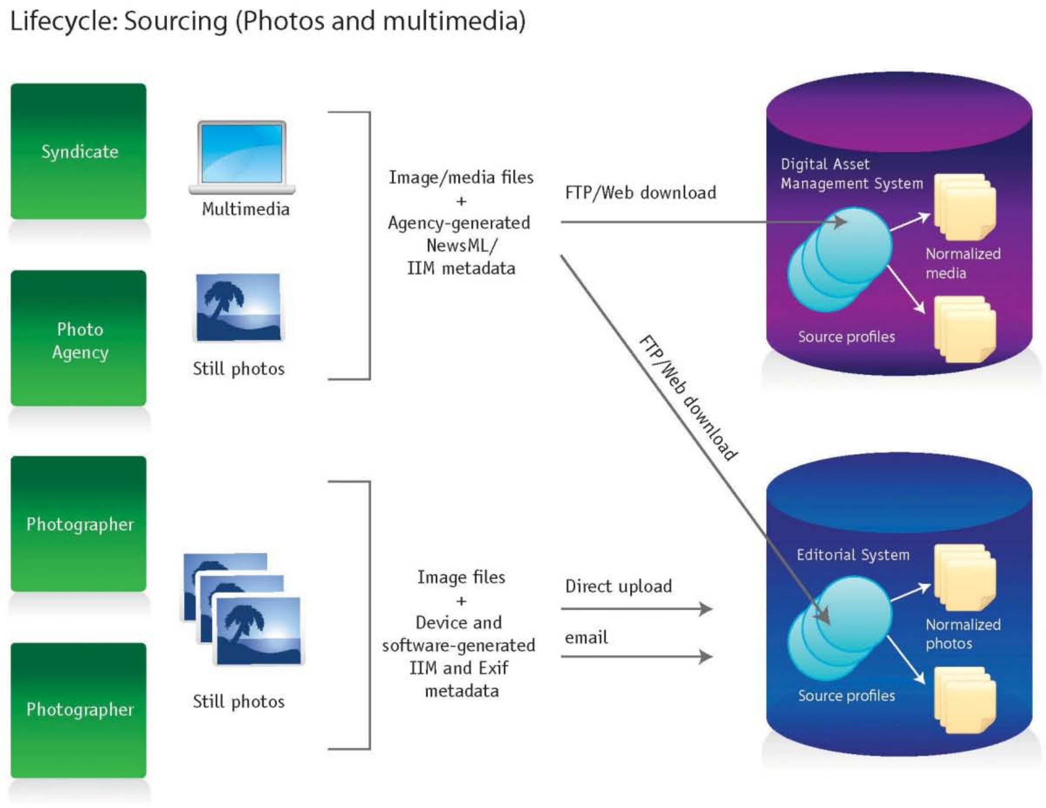
Today most photos are submitted to newspapers electronically, or are scanned and ingested electronically from hard copy. Staff and freelance photographers, wire services, and photo agencies include IPTC-fielded metadata with their electronic submissions. Some of this information is generated automatically by the capture device (cell phone, camera). More is added during the "post-camera" processes, such as editing in PhotoShop and still more in the "picture desk" and digital asset management systems used by the syndicates and photo agencies.


Lifecycle: Sourcing (Photos)
Photographer
Image files + Device and software-generated IIM/Exif metadata
Agency DAM System
"Picture desk" software
Annotated JPEG photos + NITF/NewsML metadata

There are several generations of standards for representing data about photographs. The principal standard for data about photographs, graphics and other images in wide use in the news industry since the 1990s was the Information Interchange Model (IIM). IIM is still widely used for photographs today.

IIM was created by the IPTC in 1991 as a set of metadata for digital text, photos, graphics, audio, and even digital video streams. IIM enables the transfer of these data objects between computer systems, providing an envelope around the object containing information on the type of content or document, file format, transmission date and time, etc. IIM metadata elements are known as "IPTC fields" in the "IPTC header" of digital image files. Some IIM metadata is automatically applied to photographs by digital cameras.

Exif is a multimedia metadata format standard adopted by most manufacturers of digital cameras to embed JPEG, TIFF and RAW format images with technical information. Exif metadata can include such information about the image as the make and model of the camera used, settings, timestamp, GPS location, photographer ID, and even face recognition details, as well as “color space” information such as RGB, Adobe RGBplus, orientation (“landscape” or “portrait”), and maximum available size of the image in pixels. Additional IIM metadata can be applied by photographers to their images “post-camera,” in PhotoShop, FotoStation, Extensis Portfolio, and other widely used image management software.



The various IPTC image metadata standards specify the types of information that can be included in specific fields in the IIM envelope for a digital object and identify the numerical and controlled vocabulary tags for those fields. IIM numerical tags identify such categories of information as the file format of the photograph (1:20), priority (2:10), key words (2:25), datelines (2:55), and author (2:80); and IIM symbols for general types of content, such as arts, culture and entertainment (ACE), disasters (DIS), health (HEA), etc., aid in directing the image to the appropriate newspaper "desk." Within these broad categories number codes designate more specific subjects, as fashion (01007000), earthquake (03002000), labor disputes (09004000), etc.[8]

Within the IIM fields there is considerable variation in the vocabularies used. The IPTC gives general directions that allow providers certain amount of latitude, such as "Not repeatable, Maximum 32 octets, consisting of graphic characters plus spaces" for the geographic source of a transmitted photograph or audio-recording. So codes representing the source of the transmission could be either "PAR-5" or "PAR-12-11-01," the latter incorporating the November 12, 2001 date of a photograph transmitted from the AP's Paris bureau.

IIM metadata is most often stored in the image file itself. To the metadata generated automatically by digital cameras, photographers and editors can add more metadata in the "post-camera" stage, i.e., in the course of image processing with software such as Adobe's Photoshop. In 1995 Adobe adopted IIM as a format for metadata in its *Photoshop* product. Hence many of the IIM fields map directly to metadata fields in Adobe, Apple and other image management software.[9]

[8] The 2007 IPTC white paper provides a great deal of information on the IIM metadata schema and its use with press and agency photographs, International Press Telecommunications Council *Photo Metadata White Paper 2007*, at: http://www.iptc.org/std/photometadata/0.0/documentation/IPTC-PhotoMetadataWhitePaper2007_11.pdf

.

[9] IIM tags include descriptive metadata (e.g., genre, headline, and subject information); administrative information (unique file ID number, title, date and place created, job ID, instructions like embargo, caption author); rights (photographer, and job title credit line, copyright line, rights contact, releases, usage rights, provider); and information on metadata schemas (such as IIM or the updated IPTC core scheme for XMP PLUS, regarding picture licensing rights; and more robust DIM2) For further detail on IIM, see the IPTC Web site at http://iptc.cms.apa.at/site/News_Exchange_Formats/IIM/

Today many syndicates and photo agencies provide more richly tagged photographs in an XML wrapper, usually using NewsML. NewsML provides much richer encoding possibilities for digital photographs as well as news texts and multimedia. NewsML formatted information for a photo transmitted by an agency could include metadata marked up as follows:

```
    <usageTerms>NO ARCHIVAL OR WIRELESS USE</usageTerms>
    <versionCreated>2010-10-19T02:20:00Z</versionCreated>
    <pubStatus qcode="stat:usable" />
  <urgency>4</urgency> ["1" indicates the most urgent]
  <fileName>USA-CRASH-MILITARY_001_HI.jpg</fileName>
  <headline xml:lang="en-US">A firefighter walks past the remains of a military jet that
      crashed into homes in the University City neighborhood of San Diego</headline>
   <description xml:lang="en-US" role="drol:caption">A firefighter in a flameproof suit
      walks past the remains of a military jet that crashed into homes in the University
      City neighborhood of San Diego, California December 8, 2008. The military F-18
      jet crashed on Monday into the California neighborhood near San Diego after the
      pilot ejected, igniting at least one home, officials said.</description>
   <creditline>Acme/John Smith</creditline>
```

Therefore, a photograph transmitted from a photo agency or a photographer to a given newspaper will, at minimum, normally include information about its authorship, source, technical characteristics, subject matter, urgency, and general terms of use. It might also include, however, an abstract or descriptive information, version history, and related media objects.

A.2 NORMALIZATION AND EDITORIAL MARKUP OF TEXT AND IMAGES

Editorial systems and digital asset management systems normally convert photographs on ingest to JPEGs. JPEG is the most widely used format for newspaper photos and all major editorial, digital asset management, pagination, and Web production systems accommodate them. Graphic images are normalized to EPS (Encapsulated PostScript) format, which is also a standard format used in print production.

Once within the editorial system texts and images are normally assigned additional metadata, using proprietary file formats and tags that generally correspond to IPTC subject codes and terms, augmenting the information originally attached to the object by wire services, reporters and other providers. At this point in their lifecycle news texts and photographs are highly structured in ANPA or XML and are heavily annotated with fielded information. For example, text received by *The Chicago Tribune* from the Associated Press may arrive with broad, standard IPTC subject category designations in the metadata (“news,” “sports,” “business,” etc.) and general keywords like "Illinois," "politics", "governor." Tribune editors then enrich this metadata by adding more specific, locally relevant keywords, like “Cook County,” “corruption,” and system-generated elements such as article ID, assignment number, references to related news objects, etc., and even descriptive abstracts.

However, as the story or photograph moves toward formatting for print, web and other platforms and operating systems and integration with other elements of the issue or Web feed, it is gradually stripped of much of that information.

B. EDITING AND PRODUCING THE NEWS

Content/data:

- Article, feature texts
- News wire reports
- Graphics
- Photographs
- Audio recordings
- Video recordings
- XML text and metadata
- Page images files
- Databases

Actors:

- Newspaper publishers (Arizona Republic, Seattle Post Intelligencer, the Chicago Tribune)
- Parent organizations (e.g., Gannett, Hearst Corporation, Tribune Media)
- Affiliated newspapers and news organizations
- Editors and newsroom managers
- Photo editors
- Media organizations, syndicates (e.g., News Corporation, Gannett, Hearst, Tribune)
- Associations and standards bodies (i.e., IPTC, NAA, ISO)
- Designers, layout artists
- News archivists
- Programmers and system designers
- Software producers (Adobe, Microsoft)
- System vendors (CCI, DTI, Nstein)
- System managers and hosts

Editorial or production activities, as we discuss them here, are the processes and tasks that take place under the auspices of a newspaper publisher or its parent organization, to produce a printed edition, a Web site, or other electronic derivatives of same, and to manage the content needed to support those activities. These activities involve the review, revision, editing, formatting, markup, and tagging of the sourced content and the assembly of a newspaper's print edition from that content, formatting of content feeds to text and electronic facsimile aggregators, and the development and formatting of content for the newspaper's Web site. These activities are augmented by services provided by other organizations, including printers, advertisers, and data providers that also produce content and data for a publisher's print and Web output.

In the newspaper era every newsroom functioned in essentially the same way. More than two centuries of practice in the US and the industry's dependence upon a small number of major syndicates (AP, UPI, the New York Times, Magnum, Dow Jones) for news gathering created a certain uniformity of practice. This uniformity applied to the news cycle (daily, weekly), writing style and formats (inverted pyramid, editorials, and advertisements), and conventions in placement of masthead, headlines, bylines, article text, and other page elements (multiple columns, sections, above and below the fold. Many of these processes went the way of the afternoon newspaper as the Web emerged as a primary platform for news access. But the effect of the print-era conventions is still apparent in the organization of the electronic newsroom. As Victoria McCargar writes:

> Newsroom technology . . . uses metadata schemes that reflect these structures, and something approaching a standard is central to virtually all news production. In part, this reflects the dominance of wire-service protocols that date back decades and have evolved along with technology but still maintain these basic structures.[10]

In the digital environment current practice for formatting and encoding news for print and Web distribution is shaped by a handful of industry standards and by a legacy of proprietary local systems and processes. Media convergence and consolidation in the news industry, however, is creating even more uniformity of practice today. As local newspapers are acquired by larger media groups, many production practices are being dictated by the home office. And back-office operations are being combined through implementation of enterprise-wide systems that provide common platforms for managing text, still image, video, audio, and database content.

This drive to uniformity is enabled by the networked nature of digital technology, which is changing the balance between internal and outsourced processes, and local and centralized control of those processes.

At the same time developments in media technology are giving rise to wider variety of derivative news products, which in turn makes the production process newly complex. In fact, not all content that ends up in the newspaper or on the newspaper's Web site even travels through the publisher's editorial system. Certain types of content supplied by third parties, such as print advertising inserts and supplements, are delivered directly to the printer by direct mail and advertising services providers. And

[10] Victoria McCargar, Associated Press Repository Profile, Center for Research Libraries, revised January 2011. http://www.crl.edu/sites/default/files/attachments/pages/AP%20Profile%20final%20doc_3_2011.pdf Accessed 11/12/10.

certain multimedia content and data feeds for Web display reside only on the services or data provider's servers.

Most editorial systems are built around the traditional concept of a news article or story. Following longstanding newsroom practice most text is maintained in the editorial system as part of a standard unit of content: it is a news item (for wire service reports) or a story, article, or feature. Photographs are normally managed either as individual items (from wire services and syndicates) destined to be linked with a related story or feature, or are bundled as an assignment, for later winnowing by editors. The richest and most extensive tagging is applied to content during the editorial process. Story text, photographs and other content are tagged with administrative information on authorship, date, status and ownership of rights, version, language, level of urgency, place of origin, and so forth. Specialized and highly localized subject tagging is applied in the editorial system, and abstracts are sometimes generated as well. Much of this information is later stripped out or lost, as the final published version of the news product takes shape.

B.1 PRODUCTION SYSTEMS

Most components of the print, Web and electronic facsimile editions of newspapers are produced using one or more of three integrated systems: an editorial system, an archive or digital asset management system, and a pagination system.

Today, as news organizations increasingly regard the Web as their primary distribution channel, the number of systems that can accommodate the complex content management needs of multi-platform news production is becoming smaller. The industry leaders in producing these systems are CCI Europe, Nstein, SCC, and DTI. The Tribune Company has implemented CCI's *NewsGate* as its editorial system. *The Seattle Post-Intelligencer*, along with the other Hearst newspapers, is in the process of implementing Nstein's *DAM*, *TCE* and *Web Content Management* systems. These systems integrate the functions of the legacy print production systems with the capability to format and deliver content to the Web and a range of electronic devices, such as mobile phones, PDAs, and e-readers.

A notable common feature of the newer systems is that they are often deployed and maintained centrally by a newspaper's parent organization, rather in the local newsroom itself. The ability of such systems to be accessed by reporters, photographers and other content producers via the Web and to ingest content from multiple remote locations means that they need not be instantiated at the local level. They can also be scaled to handle editorial and production work for several different newsrooms at once. As some of the systems incorporate administrative functions such as assignments, circulation management, and billing, they offer advantages to parent organizations that endeavor to increase their control over local newspaper operations and budgets.

These editorial, archives, and pagination systems tend to use proprietary software, create their own versions of standard content formats, and employ their own variants of industry standards for content markup and coding. IPTC codes are used alongside locally developed codes that identify placement in particular newspaper-specific sections, for instance, or indicate special rights arrangements regarding news content. Therefore, markup and formatting tend to be uniform only among publishers using the same production systems.

The individual editorial systems often consist of discrete modules which can be adopted individually and operated in tandem with existing legacy systems or with systems produced by other system vendors. The *Arizona Republic*, for example, employs the *NewsDesk* editorial system and *Layout Champ* pagination system, both produced by CCI Europe, alongside the *DC5* archive or content management system, which was developed by Gannett Technologies. (*NewsDesk* is one of the older editorial systems produced by CCI Europe and is still in use at many newspapers.) The Chicago Tribune uses CCI's next generation *NewsGate* integrated editorial and pagination system in tandem with *Assembler*, its own Web authoring system. The *Wisconsin State Journal* uses Advanced Publishing Technology's *Falcon Editorial 4.1.9* combined with Adobe's *InDesign* (design and pagination) system.

B.1.1 THE EDITORIAL SYSTEM

Systems: CCI *NewsDesk,* CCI *NewsGate,* Unisys *Hermes,* Nstein *WCM,* APT *Falcon Editorial*

The editorial system provides writing and editing functions, storage of text news objects prior to publication, encoding tools, and export of content to pagination and digital asset management systems, archives and other channels. Content enters the editorial system with what minimal metadata is supplied by the producer or provider, using standardized International Press Telecommunications Council (IPTC) codes. (See Section A.1, *Sourcing*, above.)

The text editing component of most editorial systems converts imported text (e.g. structured text from wire feeds, or word-processed documents submitted by reporters) to Microsoft Word or their own proprietary version of same. Stories are then composed and edited directly within the editorial system environment. News texts, which at this stage can be either draft articles or fast-breaking news items or updates from the wire services, are then copy-edited, marked up and tagged in proprietary SGML/XML formats, and stored. The tagging codes conform at least loosely to the IPTC standards, in order to optimize earlier tagging by the wire services and capture or authoring devices.

Once material is ingested, converted, and assigned IPTC metadata, it is sent to appropriate "pools" within the system. Wire service stories are automatically parsed according to attached metadata, and filtered into several categories in a "wire basket." Text is also tagged with additional subject keywords and names, and sent into a "text pool," which is filterable and searchable.   The story can grow and evolve through several iterations within the system. Here take place the beginnings of layout as the edited text is assigned a preliminary headline, lead, and an article ID, and linked with photographs, graphs, or other graphic material maintained separately in the digital asset management system.
As in the traditional newsroom, the "story" is a basic unit of content and organizing principle for editorial activities.   From the CCI NewsGate product brochure:

> The Story Folder -- the "homepage" of a story in NewsGate -- is the central collaboration area for everybody in the newsroom contributing to the creation and development of the story. In the Story Folder editorial teams are formed; assignments are distributed; background information is gathered; relevant contact information is stored; and content is collected, created, edited and published. . . . The Story Folder constitutes the central hub for research and story development.

After copy-editing, the texts are tagged with publication-related indicators for layout elements, such as publication date, section, byline, headline, etc. , as well as administrative information such as assignment, rights status, level of urgency, etc.

Ads for the print edition are either provided by advertisers in PNG or JPEG format, or are produced in-house from camera-ready material provided by the advertiser. (Ads for the Web edition follow an entirely different production path.) Multimedia formats often do not enter the editorial system, but rather bypass the system entirely and are ingested into the Web production system.

Once the copy editors complete their work, the XML article package for the print edition is released to the pagination system for final layout. The package is also forwarded to the Web production system.

Therefore, at this stage in the production process the content elements (photographs, article texts, etc.) will have minimal metadata, most of which is information relating to layout of the print edition. Such metadata will normally include issue date, section, page number, and references to related content, as well as text structure indicators such as headline, byline, and dateline. Photographs and illustrations will also include cutline or caption text.

B.1. 2 "PAGINATION" OR NEWSPAPER DESIGN AND LAYOUT

Systems: Unisys *Hermes*, CCI *Layout Champ,* DTI *Newsspeed*

After the articles and other components of a newspaper print edition are assembled and tagged in the editorial system they are usually exported to a pagination system where the page layouts for the print edition and e-facsimile edition are created. It is in the pagination system that most of the content for these editions of a newspaper is brought together for the first time.

At this stage the separate article text elements (headline, dateline, byline, lead, body) and related images, photographs, and charts are all separate files in the native formats of the generating editorial system or digital asset management (DAM) system. For text this is usually either a proprietary version of Word or an XML-formatted text document marked with minimal tags for paragraph and character idiosyncrasies like italics, diacritics, etc.

The accompanying photographs are low-resolution JPEG files, to be used for positioning only. The article components are linked through unique article or "news item" ID numbers generated by the editorial system, which enable them to be assembled by the layout artist. In the pagination system the article texts and images are then joined "physically" to the other elements and contents of the issue, such as charts, illustrations, masthead, editorials, columns, advertisements, etc.

Advertisements to appear in the body of print newspapers are often produced in a separate CCI module called *AdDesk*, and uploaded to the pagination system. (Printed advertising inserts are produced by third-party direct mail houses and enter the printing workflow separately. Banner and video advertisements are also produced separately and served directly to the Web by third party providers like AdTech, Google Ads, Pulse 360, and Brightcove.)

Pagination system software enables designers to lay out individual articles in the form in which they are to appear on the printed page, determining the article size, shape, type fonts, spacing, arrangement, and other design elements. These elements display visually within the pagination system environment. Programmed into the pagination system software are the particular rules and templates that give a particular newspaper its distinctive look, such as the type fonts and sizes used, ink color palette, number of columns, etc.

The design tools and templates often segment article or news item components, for example causing the headline, lead and beginning of an article text to appear on page one and the text to continue unbroken on an interior page. As one designer explains the process, "You design the article shape then fill them in with the headline, lead, byline and other article elements."

The most widely used pagination system is *LayoutChamp*, which is produced by CCI Europe. (Other widely used pagination systems are Unisys' *Hermes* and DTI's *NewsSpeed*.) *Layout Champ* is a proprietary software suite that employs its own native formats. These systems are often used in conjunction with "typesetting engines" like H&J, TeX and column justification algorithms, and often integrate with image editing software like *PhotoShop* and *Illustrator*, allowing editors and layout staff to make edits and alterations within the pagination environment. Other page design tools, including *Quark XPress*, and Adobe *InDesign*, are often incorporated with, or used in lieu of, pagination systems.

Once composed, the page files are tagged with codes in the pagination process that indicate the particular section and page number, with additional information on "zones." This section and edition coding is not uniform from one paper to the next because many of the section and edition names differ. (For example, the *Tribune* produces different sections for different regions of distribution, such as North Shore, South Side, national, etc., whereas the *Arizona Republic's* sections are Scottsdale, Tempe, North Valley, etc.). The section codes are important to the plate-setting and print production process.

Meanwhile the photographs and graphics selected for use in the edition are processed separately, normally using *PhotoShop* or proprietary software of the Digital Asset Management system. Here cropping and color adjustment is done, and the brief cut line is expanded to a full caption. In some instances the photographs are further enhanced for uniformity. (Gannett operates regional "toning centers" for this purpose.)

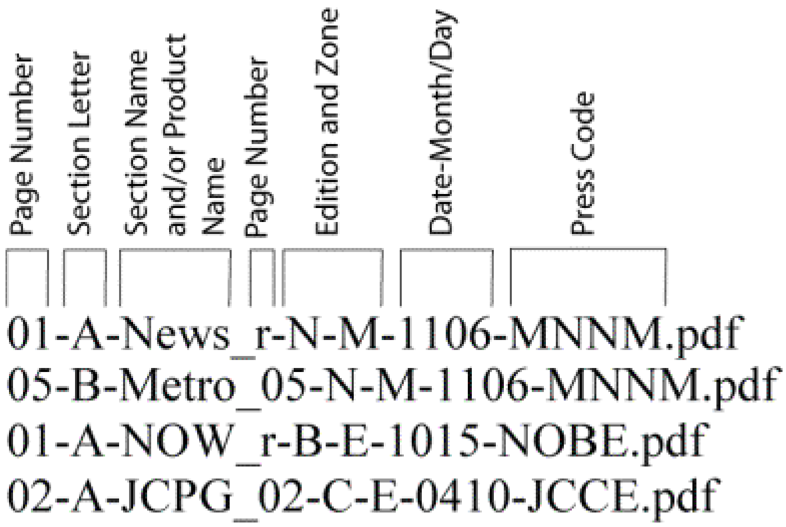
The set of composed page images for each edition of the print issue (e.g., "early," "final," metropolitan and regional) are then saved using Adobe Acrobat, in PDF or *Postscript®* format. The *Tribune* and *Arizona Republic* prefer *PostScript* for printing. (*PostScript* is a predecessor digital image format to PDF, developed as a language to drive high-quality image and text printing).

Once assembled, the page image files are then often run through a correction and optimization processor, such as *OneVision*, which uses one of a number of proprietary software types to enhance the color balance and consistency and overall image clarity. Image and graphics content, because they come from a number of different sources, may vary in quality, resolution, and color balance.
There is little uniformity in the metadata embedded in the output page image files from one publisher to the next and even among multiple titles produced by the same publisher. While Adobe imaging software does accommodate storage of a considerable amount of metadata, the page image files we examined had very little information. At best, the metadata actually embedded in these files included:

- File name (e.g., "ARIZONA REPUBLIC front page")
- File title ("06-20-10A_01-Z0.pdf" for the June 20, 2010 Arizona Republic front page)
- Type of software or service used to produce the PDF ("OneVision PDFengine","Acrobat Distiller Services")
- Date and time of creation ("6/20/10 3:01:41 AM","5/28/2010 4:20:03 PM")
- Date and time of revision
- Color correction and optimization application used ("Asura version 8.2")
- File size
- Page size (in inches)

This information roughly maps to the IIM fields developed by IPTC, but the field content is normally not consistent in structure or in controlled vocabulary.

Often, minimal information is also embedded in the titles of the individual page image files. But this too is not uniform. One source told us, "When we receive PDFs from publishers they come with file names set up by the publisher. We get all kinds of different naming conventions for publishers and there could be multiple naming conventions within a chain of papers which suggests they do not have any standard process.” An illustration of sample file names from different sources, reflecting slightly different naming schemes, appears below.



01-A-News_r-N-M-1106-MNNM.pdf
05-B-Metro_05-N-M-1106-MNNM.pdf
01-A-NOW_r-B-E-1015-NOBE.pdf
02-A-JCPG_02-C-E-0410-JCCE.pdf

In some cases there is also an index file that goes along with the set of page images, to guide the plate-setters and aggregators in properly sequencing the pages and sections of an issue. We were not able to determine what these index files look like. But inclusion of these files is not standard practice, and most sequencing is guided by the “profile” for the particular publication programmed into the plate-setter system.

The set of page image files for a daily newspaper edition are transmitted nightly to the designated printing plants. (This transmission normally occurs between midnight and 4:00 AM.) Some publishers send these files by an FTP feed. Others place them in a “drop folder” in a Web-accessible database, from which they are then retrieved by the printer’s plate-setter system.

B.1.2.1THE PLATE-SETTING PROCESS

Systems: Agfa's *Advantage*, Krause's *LS Performance*, ECRM's *Mako* and *Newsmatic*

"Plate-setter" or "imagesetter" systems are used to generate sets of CMYK (cyan, magenta, yellow and key) printing films or polyester plates for each newspaper page. [11] To administer the printing jobs printers use a "TIFF-catcher" software, such as *CtServer,* normally mounted on a PC workstation, which maintains a list of jobs submitted to the printer's RIP system[12] in "drop" folders. Each job is associated in the drop folder with a property set, which is a set of output characteristics, such as margins, cuts size, and output media type and width, associated with the particular job or newspaper. Placing a TIFF file in the drop folder submits that TIFF for tagging, scheduling and printing. The tags now include Photometric Interpretation information, such as positive or negative image, margin (white/black), resolution, image size, and colors (e.g., CMYK). The *Ctserver* operator then generates separate TIFF files for each color (CMYB plus) for each image. (In many cases the ECRM RIP system has already generated separations.) These computer-to-printer processes are increasingly automated.

Additional, updated page-image files are often transmitted to the printers before and even during the press run. These files incorporate corrections, updates, or even entire new elements, resulting in variations in content within the same day's edition. After printing, pages are combined with printed advertising, comics or magazine supplements which are produced from a separate workflow managed by specialty publishers and direct mail houses.

The same set of edition files used for printing is also transmitted to aggregators that publish the facsimile editions, such as *NewspaperDirect* (PressDisplay), *NewsStand* (ProQuest), *Active Paper Daily*

[11] *CtServer* software is used to send jobs to the plate-setter. It is also used to define the plate and pin bar requirements for jobs and communicate that information to the plate-setter, which displays the information on the control panel LCD. The use and operation of CtServer is documented in the *CtServer User's Guide*. CtServer "provides for setting up, monitoring and outputting jobs to ECRM imaging devices." Users' Guide is at http://www.ecrm.com/uploads/support_documents/ag4508914.pdf

[12] http://www.ecrm.com/newspaper/subcategory/RIP%20Upgrades/24/

(Olive), and *Image E-editions* (NewsBank). Many newspapers then archive copies of the PDF files for each section front (e.g., pages A1, B1, C1) in their digital asset management systems. Since these PDF files are simple image files, however, they are absent much of the rich administrative, structural and descriptive metadata associated with the various component parts of the newspaper issue applied by providers and in the editorial process.

Therefore, at this stage in the production process the content elements (PDF page images) will have minimal metadata, most of which is information relating to the particular issue, section and page of the print edition and the processing of the PDF file.

B.1.2.2 FORMATTING FOR TABLETS AND E-READERS

Many publishers also make versions of their news products available on e-readers and tablet computers like Amazon's Kindle and Apple's iPad. Although the e-reader formats have an even greater degree of functionality than the conventional electronic facsimiles, they normally begin with versions of the page images much like those generated for print and e-facsimile publication. Editorial and pagination systems, like *NewsGate* and *NewsDesk,* can create content for these special applications. Paginations systems such as Adobe *InDesign* and *CCI LayoutChamp* use special templates for the particular device formats to generate PDF page images and accompanying metadata. These page-image files with XML metadata are then processed by various vendors for distribution on various e-reader and tablet platforms.

The files are processed through special applications like *Skiff* and *WoodWing* in order to function on the proprietary e-reader platforms used by Amazon's *Kindle*, Barnes & Noble's *Nook*, and Apple's *iPad*, some of which involve touch-screen functionality. While there is considerable variation in the respective outputs of these platforms, the article files are richly embedded with codes and annotations that enable a high level of functionality when viewed on the target devices.

B.3. ARCHIVING AND DIGITAL ASSET MANAGEMENT

Some content acquired or generated by local news organizations has value for potential reuse. In particular, still images and video produced by staff photographers and by freelancers on assignment, articles and columns written by local reporters and editors, and illustrations by staff artists constitute "assets" or intellectual property considered by the news organization to be worth preserving over time. These assets include both published and unpublished materials.

Originally such materials were kept in newspaper "morgues" or clippings files. Some newspapers even maintained extensive back files of their published editions. Those paper files often contained materials to which a newspaper owned publication rights as well as materials obtained for one-time or limited use from reporters, photographers, syndicates, and other providers. Much of this infrastructure disappeared as the newspaper industry downsized during the 1980s and 1990s, and as the locus of many parts of the news operations shifted to the parent company.

In the 1970s and 80s, as newspapers introduced digital production methods, the amount of content a newspaper accumulated became formidable, and the proliferation of potentially usable content in multiple versions and multiple electronic formats began to present an enormous control challenge. Digital asset management or "archives" systems were then developed for large-scale commercial use, to provide the robust file and rights management capabilities necessary in the technology-driven media environment. Initially these systems were created primarily to manage photographs. Today, with the variety of media generated and managed by news organizations, and the "long tail" of value for digital content, the management and repurposing of all types of digital assets for the multiple delivery platforms (newspaper, magazine, radio, television, Web) under the control of the parent companies has again become a central activity of news organizations.

B.3.1 THE MAJOR DIGITAL ASSET MANAGEMENT SYSTEMS

Systems: Hermes *Doc Center*, Gannett Media Technologies' *DC5 Digital Collections*, *SCC Media Server*, SCC's *Merlyn,* NewsBank's *SAVE,* Nstein's *DAM.*

The *Arizona Republic* uses *DC5 Digital Collections,* an archives system developed by Gannett Media Technologies International, an affiliate of the *Republic's* parent company. Typical of the enterprise-level digital asset management systems, *DC5* is a powerful system with the capability to store and index multimedia objects and file types. This system manages the following *Arizona Republic* content:

- Published photographs with cut line
- PDF images of published front pages
- Associated Press wire photographs for which the Arizona Republic has publication rights
- XML marked up texts of published articles
- Raw texts of articles not yet marked up

Photos of all formats are converted to a baseline JPEG format on import to DC5. *Arizona Republic* staff photographers normally make initial selections of photos from their assignments and upload them directly to the system, where editors perform further selection, cropping and captioning. The Republic's *NewsDesk* editorial system integrates with *DC5* to append publication metadata to the photo's record, which then is stored for the longer term in *DC5.*

SAVE is a widely used archiving and digital asset management system operated and hosted by NewsBank. The system was originally developed to accommodate the texts of articles published by newspapers, to serve as a type of virtual morgue, but now accommodates page image files, texts, photographs, and graphic content.

Seattlepi.com and its parent company Hearst Corporation use Nstein's *DAM* system, a "media hub" for storing, repurposing and syndication of its text and rich media digital assets. *DAM* is an XML content repository that automatically interacts with Nstein's TME text-mining engine system to semantically tag newspaper articles with subject categories, and to thus enhance discoverability of content across all Hearst newspaper sites.

B.3.2 DIGITAL ASSET MANAGEMENT PROCESSES

Today many of these systems are used to store and manage all types of digital content, including stories and other raw and published text, locally produced and wire photos, information graphics (ingested as EPS files from vector software), third party-produced PDF documents, and PDF-format published newspaper page images. Some of the formats accommodated by the major systems in use today include:

- JPEG, TIFF, PSD, RAW and other photo formats
- EPS and native graphic formats like Illustrator™ or Freehand™
- PDF files with full-text indexing and searchable content
- Multimedia files such as QuickTime™, MP3, WAV, AVI
- Text in ascii, XML, SGML, and HTML
- Native file types from Quark, Adobe, Microsoft, and Macromedia software.

Local tagging of news content in the digital asset management system is often enhanced through third-party semantic analysis, annotation and tagging by syndicates and third parties. Wire services like Reuters and Associated Press provide this kind of data tagging for client news organizations. Many AP members submit their news text, audio and image content for tagging through AP's Content Enrichment Initiative. And Reuters' *Open Calais* service also provides rich subject and name analysis and tagging and annotation of content files in all media.

Photographs chosen for publication are usually edited and formatted in the DAM system. The DC5 system, used by Gannett newspapers, generates multiple versions of the same image:

1. A low-resolution JPEG used only for page layout (known as "FPO" or "for position only.") This is what a layout editor sees in the *Layout Champ* pagination program.

2. A high-resolution JPEG that is sent to a color-correction or "optimization" lab. This image is converted from RGB to CYMK, and settings are applied that ensure optimum print quality on the local presses, and then is later inserted in the page image files sent to the printer.

3. The original high-resolution RGB image (JPEG) that is the archival version.

Some archives systems, such as *SAVE* and *SCC Media Server* also feature some minimal text editing capabilities as well, and can generate output files (originally prepared in pagination systems) directly to printers and other content vendors including Factiva, NewsBank, and Lexis-Nexis.

In some instances two digital asset management systems are used in tandem. The Wisconsin State Journal uses *Merlin* as its local archiving system for high-resolution published images after editing and enhancing, and the *SAVE* system for text and as a generator of content for aggregators.

Unlike the morgues of the past, however, digital content management systems normally only hold content for which a publisher either owns rights or has licensed rights for extended periods of time. Blog posts and other user-generated content, for example, are not archived by the newspapers in the digital asset management systems, because they are the intellectual property of the post writer. Uploaded through a Web interface, blog posts are archived in the native blog platform, like WordPress, but are rarely captured for long-term retention.

Archive systems usually provide greater storage capacity and functionality for managing image and multimedia files than editorial systems. This is the point in the news lifecycle where the most heavily annotated and tightly controlled content files reside. Media files and article texts stored in these systems are richly tagged with granular metadata relating to production date, authorship, provenance, version, usage and resale rights, publication history, expiration dates, related news objects, and so forth. The formats for such metadata generally follow the IPTC standards and vocabularies, or employ system-specific or locally relevant tags mapped to same. Published stories and accompanying images, although filed separately, can be linked to each other through this metadata and to corresponding page images. And published images can be linked to outtakes from the same assignment.

B.3.3 CENTRALIZATION AND OUTSOURCING OF ASSET MANAGEMENT AND PROCESSING

Developments in the news industry are fueling a growing tendency for large media organizations to manage their digital assets on an enterprise-wide basis. With the consolidation in the news industry, content produced by local newspapers is shared with other news organizations under the same parent company. Photographs by *Arizona Republic* staff photographers, for example, are maintained under the control of Gannett, the *Republic's* parent company, in its own “enterprise-wide” digital asset management system. With media convergence news stories, photographs, video, audio, and other content can be versioned for several media platforms under the local newspaper or parent’s control.

Because of the heavy storage demands imposed on these systems by multimedia files, even much enterprise news content is now not actually maintained locally. The New York Times, for example, uses Amazon's S3 service, a "cloud" data center, to store its photographs. In some instances even the actual management of a newspaper's content is outsourced. Many of the DAM systems can be either licensed and integrated locally with a newspaper's IT environment or deployed on a "software-as-a-service" basis, hosted by the system's producer. And content archived in the *SAVE* system for many newspapers is hosted by NewsBank.

Some systems rely on third-party service providers for subject tagging and indexing of content held in their own digital content management systems. SCC *MediaServer* accommodates the use of Thomson Reuters' *Open Calais* service for semantic analysis and tagging of Hearst newspapers and magazine's archived content. (*Open Calais* can be used either as an on-site application or as a hosted solution managed by Reuters.)

C. DISTRIBUTING NEWS CONTENT

Content/data:

- Articles, feature texts
- Graphics
- Photographs
- Audio recordings
- Video recordings
- Tables
- Databases
- Data visualizations
- News object metadata
- Web design templates
- Style sheets
- User data
- Algorithms and programming code

Actors*:*

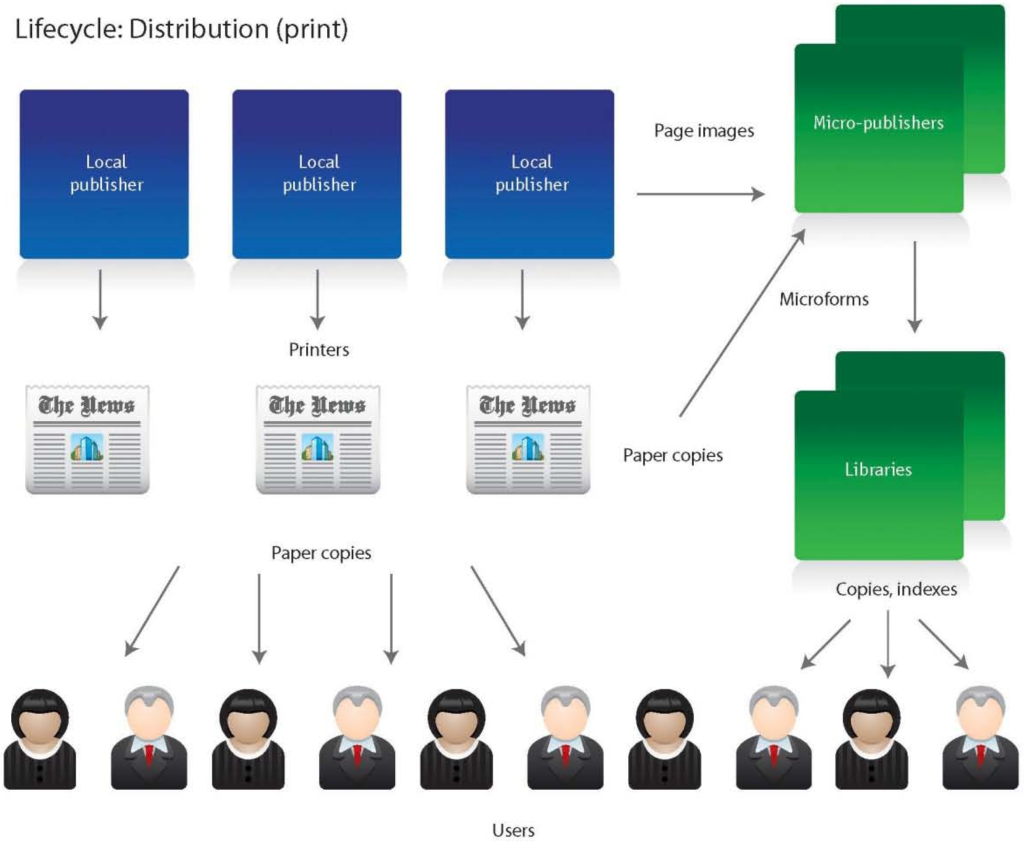
- Newspaper publishers (Arizona Republic, Seattle Post Intelligencer, the Tribune Company, Wisconsin State Journal)
- Parent organizations (Gannett, Hearst Newspapers, Tribune Media)
- Affiliated newspapers and news organizations (print, broadcast)
- Text aggregators (LexisNexis, NewsBank, Bloomberg)
- E-facsimile aggregators (Olive, NewspaperDirect, Tecnavia)
- Micropublishers and e-publishers (ProQuest, Gale, NewsBank)
- Search engines and news reader services (Google, FastTrack News, News Rover)
- Wire services, syndicates and photo agencies
- Image processing services
- Printers
- Bloggers
- Social media platforms (FaceBook, Twitter)
- Libraries
- Indexing services (Reuters, Nstein)
- API programmers
- Ad-servers (AdTech, Yahoo!)

Most newspaper publishers now distribute their news content both in print and electronically. They output electronic page image files for print editions and for republication as electronic facsimiles. The facsimiles are sometimes available on the publisher's own Web platform, but more often are republished by aggregators like ProQuest, Tecnavia and NewspaperDirect. The publishers also generate electronic text feeds of their locally produced news content for re-aggregation in commercial news databases like LexisNexis, and for syndication through affiliated organizations like the Associated Press. Finally, newspapers generate formatted text and multimedia content that is disseminated through their own Web platforms and through a variety of mobile devices such as cell phones, PDAs and tablet computers.

Printing newspapers, for all but a few publishers, is still largely a local operation. But most distribution activities are now managed at the parent company level. Or they rely on specialized service providers such as YouTube, WordPress and others operating at the national or global level. The early steps in the preparation of text, data and still image content for these distribution channels are taken in the editorial system. But content destined for Web and mobile platforms soon parts company with versions of the same content for print and aggregator distribution. Publishers use a common, in-house work stream to produce inputs for “typesetting” and production of print and electronic facsimile editions. The production streams for Web site, mobile device, and other electronic content diverge from the print and electronic facsimile workflows in the editorial system.

C.1 PRINT DISTRIBUTION

Print newspapers follow a long-established distribution model that has not changed significantly in over a century. Print newspapers are sold by the issue to consumers by retail vendors; on a subscription basis directly to households and businesses; and to institutions (including libraries) indirectly through subscription services. Delivery to subscribers is through carriers under the auspices of the newspaper publisher or a sub-contractor, and by mail. Some newspapers are given away free and are largely supported by advertising.



In the past libraries have represented a secondary "market" for printed newspapers, but have provided a useful service for news publishers. Libraries have served as repositories for the long-term safekeeping of newspaper back files. Local libraries and historical societies have systematically acquired local newspapers for their current awareness and historical value. And the Library of Congress (like other national libraries) has been able to amass substantial newspaper holdings from publisher deposits of newspapers with the U.S. Copyright Office.

In the past libraries have not only stored and provided access to newspaper back files, but have invested a great deal in indexing and bibliographic analysis of the newspapers as well. Hundreds of local and state historical societies and state and academic libraries have created detailed indexes and "clippings files" that identified locally important or special interest content in those publications. These stewardship activities formed the foundation upon which major preservation and access programs like the NEH's long-running United States Newspaper Project was built.

Given the bulk and fragility of aging newspapers and the resource-intensive nature of the indexing and cataloging work, these services had a significant cost. For a long time that cost was underwritten by local, federal and philanthropic grant making, as a public good. Commercial microform publishers have played an important role in serving this secondary, research market as well, generating revenue from the sale of microfilm and microfiche copies of the newspapers under license from the publishers. Some micropublishers have employed the same model to distribute page images of the same newspapers in digital form.

The US Copyright Law affords libraries certain rights to make copies of the newspapers available for study, research and other non-commercial purposes. Publishers have also granted micropublishers the right to provide access to materials for a secondary audience (scholars and family historians) in return for royalty payments. This symbiotic set of relationship has enabled materials of marginal commercial value to newspaper publishers to be maintained for the long term.

These preservation activities, however, are built around the print newspaper, with its tangible outputs and regular publishing cycles. Electronic distribution requires other, more complex preservation models. With the advent of the Web, the newspaper publishers themselves are implementing or finding new platforms through which to maintain their back files and generate revenue in the process.

C.2 DISTRIBUTION THROUGH ELECTRONIC FACSIMILE (E-FACSIMILE)

Three of the four test bed newspapers -- the *Arizona Republic*, *Chicago Tribune*, and *Wisconsin State Journal* – are produced also in electronic facsimile form.[13] The electronic facsimile is a digital reproduction, page by page, of the printed newspaper, produced by the publisher as an output of the editorial process and aggregated and distributed, usually on a subscription basis, by services such as *NewsStand* (LibreDigital), *NewsMemory e-Editions* (Technavia), *PressDisplay* (NewspaperDirect), *Active Paper Daily* (Olive Software), and *Image E-editions* (NewsBank).

Production of the electronic facsimile makes use of the page images delivered to printers to create the plates and films for printing the paper edition of the newspaper. The facsimiles correspond closely to the print editions, and are therefore to a large extent considered by the publishers to be a faithful representation of the edition of record.

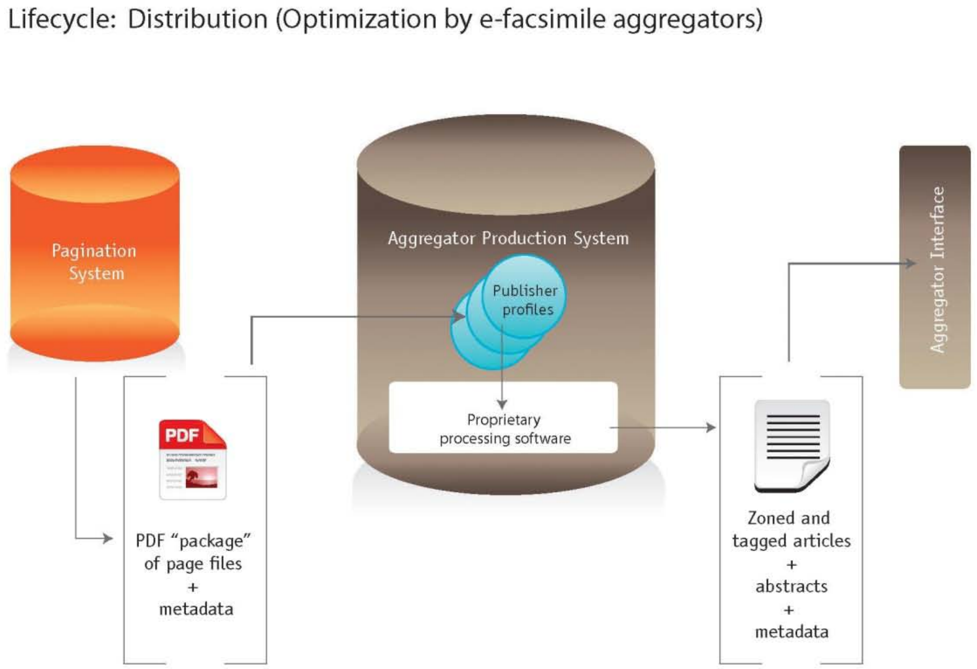
The lifespan and terms of access to the electronic facsimile varies. These are generally available through the facsimile publisher's system for a limited period of time. Newspaper issues that appear in the *PressDisplay* service, for example, are available for only three months after the issue's original publication date, and then are purged from the system. The *Tribune*, on the other hand, makes e-facsimiles of its back issues available "permanently" to those who subscribe to Tecnavia's e-reader subscription service. And the *Arizona Republic*'s own Web site provides free access to PDF images of the front pages of the newspaper's print edition for only the most recent seven days.

C.2.1 IMAGE FEEDS TO E-FACSIMILE PUBLISHERS

While page image files provided by the newspapers to printers and electronic facsimile publishers are essentially the same, there is a fundamental difference between the printed newspaper and the e-facsimiles made available through products like *PressDisplay* and *NewsStand*, resulting from heavy re-processing of those files by the e-facsimile publishers. The pages that appear in products like

[13] The *Seattle Post-Intelligence*r is a web-only newspaper and therefore does not produce an electronic facsimile.

NewsStand have been "optimized" for the aggregators' proprietary viewing platforms. That is, the aggregators enrich the image files with structural and subject metadata that facilitates searching and navigation within the aggregator interface, and interaction with the content from other newspapers in the same product. The aggregators' systems use their own proprietary software to "unwrap" the newspaper output files and to segment or "zone" the contents into sections, articles, images, and other units; to tag these segments in XML with subject and geographical information for searching; and to generate searchable text abstracts.


Lifecycle: Distribution (Optimization by e-facsimile aggregators)
Pagination System
Aggregator Production System
Publisher profiles
Proprietary processing software
Aggregator Interface
PDF
PDF "package" of page files + metadata
Zoned and tagged articles + abstracts + metadata

Zoning and indexing involves complicated algorithms and in many instances human checking and correction, since articles often run from one non-consecutive page to another in the printed newspaper and page images, following no predictable pattern. In some instances (*Olive Daily Paper*) each article is extracted from the PDF into a separate file, and can be displayed separately in either facsimile or plain text format. Some organizations use third-party software, such as Nstein's *TME (Text Mining Engine)*, to organize, categorize, and generate abstracts of its news content. This kind of processing gives the

electronic facsimile a degree of discoverability and navigability far superior to that of its original print version.
Because of the intense competition among service providers in this area, the aggregators' processes and software are tightly protected and are usually patented and copyrighted.

C.2.2 VARIATION IN CONTENT BETWEEN PRINT EDITION AND THE E-FACSIMILE

Our comparison of the electronic facsimiles and printed editions of the three test bed print newspapers determined that in every case there tended to be slight but important textual differences between the two versions. Most often, the electronic facsimile is the newspaper's metropolitan or national (rather than neighborhood) print edition, only one of several editions of a daily newspaper produced in printed form. In some instances, moreover, the text of the print editions that we examined had been slightly revised or updated in the course of the press run. Revisions could include, for example, addition of the final score of a sports event, or correction of a misspelled name. These minor revisions did not always show up in the electronic facsimile. The discrepancy occurs because in many cases the page image files are made available for pickup by the facsimile publisher in a single scheduled "drop." Thus updated PDF files sent to the printer subsequently in the course of the press run are not reflected in the electronic publication.

As production and user monitoring technologies become more sophisticated, the variation in content even within single print editions is becoming increasingly common. New geographic information systems, like ArcIMS offered by the GIS services firm ESRI, for example, permit targeted insertion of different advertising content into copies of the same issue of a newspaper printed for distribution in different neighborhoods or postal areas. While production systems have always permitted changes to be made at the time of printing, the new digital content management capabilities and the new norms of Web publishing are making it possible to vary advertising content significantly within a single press run.

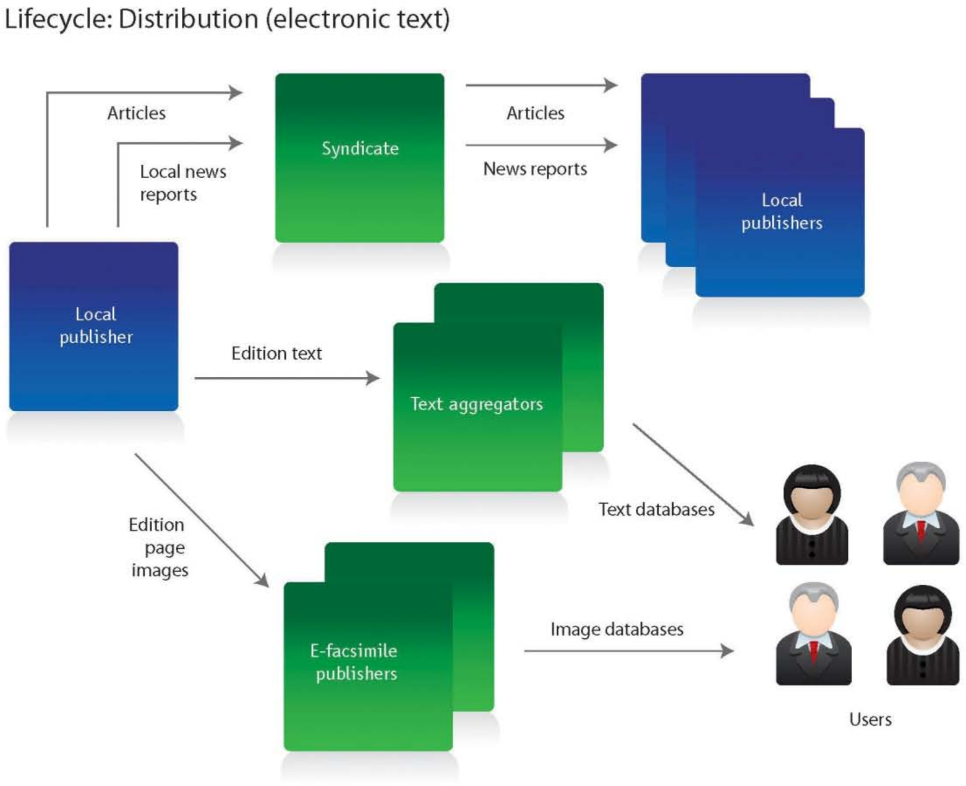
Finally, separate advertising inserts and printed features distributed with the print edition do not appear in the facsimile edition. The contents of these inserts is not produced using the newspaper's own editorial system, but are normally produced by third-party printing and direct mail services like Vertis Communications and QuadARM, and inserted during post-press operations.

C.3 DISTRIBUTION THROUGH TEXT AGGREGATORS

Text aggregators and syndication provide additional distribution channels, and revenue streams, for newspaper publishers' content. Aggregators collect news text from thousands of newspaper publishers, broadcasters, and other news outlets and combine that content in large subscription databases.[14] These services increase the value and usefulness of the content through aggregation, and are designed primarily to serve advanced academic, legal, public policy, and business researchers. LexisNexis® was the first in this arena, introducing its first news database, *Nexis*, in 1979. Later, NewsBank and Dow Jones followed with *Access World News* and *Factiva* respectively.

Syndicates and content partnerships like the Associated Press and Washington Post Writers Group also aggregate text content from their member newspapers. AP member newspapers submit fast-breaking local news reports and feature articles by staff writers to the syndicate, which then redistributes that text to other local publisher members and subscribers.

[14] "Factiva.com (Factiva) is a news and business information tool with more than 28,000 sources from 200 countries in 23 languages. . . . Sources include regional and industry publications, web and blog content, newspapers, journals, magazines, television and radio transcripts, photos, audio and video..." *VIP Report: Product Review of Factiva.com*, November 2010, http://www.dowjones.com/collateral/files/dj-factiva-vip-report-v2.pdf. Accessed 2/21/2011



The text transmitted to text aggregators is normally from the final metropolitan edition of a newspaper. Newspapers send only locally produced text to the aggregators: wire service and other syndicated content, photographs, and multimedia files are not included.

Feeds of the contents of a given newspaper issue are normally sent to text aggregators in ASCII, UTF-8, or another compatible text format. The feed is normally a single continuous stream containing all of the articles for a particular newspaper issue combined. The transmitted article texts are generated either from the newspaper's pagination system, once the page and section numbers have been established, or from its digital assets management system. In some instances, as with *The Chicago Tribune* and the *Arizona Republic*, the texts are submitted to the parent company's central digital asset management system, and then transmitted to the aggregator as part of a larger combined feed.

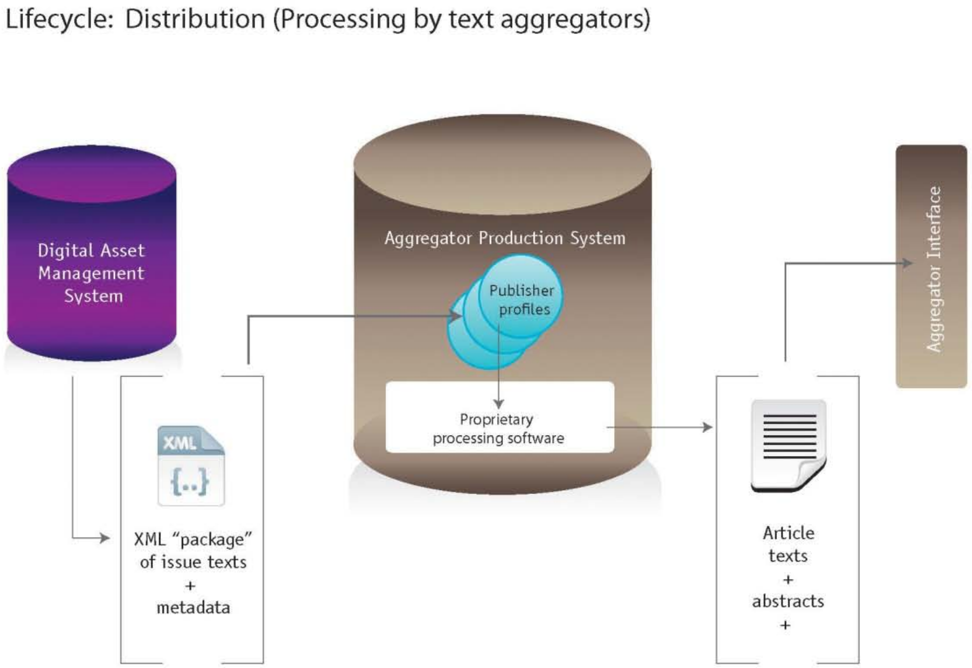
The individual elements of the texts sent to the aggregators – headline, byline, dateline, lead, body, etc. -- are relatively standardized, because of longstanding newsroom practice. There is some variation in the formatting and tagging of the texts, however, which requires a certain amount of parsing and normalization on the part of the aggregator. In most instances the text feeds are marked up in either NewsML, or, in the case of *The New York Times* NITF, a subset of SGML. Other publishers, like *Investor's Business Daily*, send minimally formatted text, with field names spelled out (e.g., "<FIELD NAME = BYLINE>").

Again, as with the page-image files, the transmitted texts carry only the minimum metadata needed to enable the aggregator's system to detect and process the essential units of the news issue. Absent is information about rights, provenance, subjects, versions, level of urgency, and other aspects of the content that were applied by producers and editors earlier in the lifecycle. Metadata embedded in a LexisNexis feed for an article in the June 1, 2010 issue of *Investor's Business Daily*, for example, included the following:

- File name for the day's feed ("nexis_20100601.txt")
- Publication day ("Tuesday")
- Edition ("National")
- Section and page ("A1MON_A1" for front page of the Monday edition)
- Article ID number ("30845")[15]

The transmitted files normally indicate the publication title and publication date for the news items included in the transmitted package. For articles and other component items the tags identify the headlines, bylines, leads, sections and pages, but otherwise give no information as to how the original printed article text was formatted, which version is being transmitted, or about rights status.

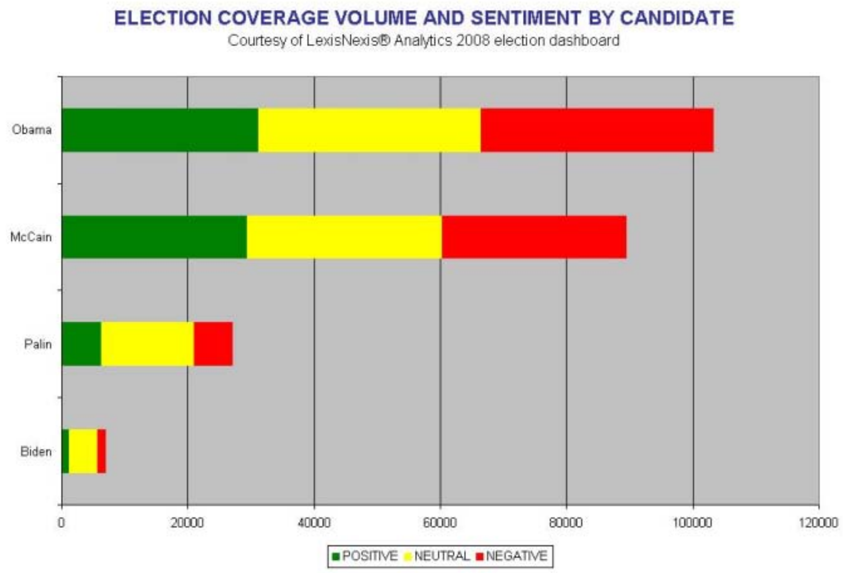
[15] Information provided to the authors by the *Investor's Business Daily*.


Lifecycle: Distribution (Processing by text aggregators)

Digital Asset Management System
Aggregator Production System
Publisher profiles
Proprietary processing software
Aggregator Interface
XML
{..}
XML "package" of issue texts + metadata
Article texts + abstracts +

The aggregators then "normalize" this information, converting the newspaper-applied tags to their own formats using a profile created for each feed source. The profile maps the original, newspaper-applied tags to the metadata formats and style sheets used in the aggregator's own content management system.

The aggregator databases are frequently used by the business, financial and policy research communities for longitudinal or time lapse analyses. Investment firms in particular have developed sophisticated algorithms to analyze large bodies of raw business news text, to track the performance over time of particular companies, funds, and industries. Similarly, researchers in the public policy field and in the social sciences have applied text-processing engines to large corpora of news text to detect political bias in media reporting, and to map trends in public opinion and thought. In recent years, with the growth of financial and policy research, aggregators like Reuters and LexisNexis® have begun to layer text-mining and sentiment-analysis tools and services over their databases, to increase earnings generated with the data.[16] The chart below was created using computer-assisted analysis of comparative media coverage of the major candidates in the 2008 US presidential election campaign. The analysis was performed using the LexisNexis® Analytics 2008 election dashboard.



The immanence of news text in the major aggregator databases is not assumed, and it is very difficult to tell with certainty what content is actually present. Text is only licensed to an aggregator on a limited term basis, and newspaper publishers can withdraw their content from a database for any of a number of reasons, such as unfavorable terms or insufficient earnings. In rare instances content is removed wholesale. The landmark 2001 U.S. Supreme Court decision in *Tasini v. the New York Times* forced the deletion from aggregated databases of all identifiable articles created by freelance writers. In that decision the Court found that newspaper publishers did not own the right to redistribute in electronic form work produced on a freelance basis for their publications. Thousands of articles produced over decades of publishing were removed from LexisNexis® and other databases.

[16] Bernard F. Reilly Jr., "When Machines Do Research: Automated Analysis of News and Other Primary Source Texts," *Journal of Library Administration*, **49**: 5, 2009, 507 – 517. On the analytical services and products, see *LexisNexis Analytics: an analytical vision on information to support your decision-making*, at http://www.lexisnexisanalytics.com/en/analytics/produits-index.html, accessed 12/11/2010; and Factiva / Dow Jones and Reuters: Media Intelligence from Factiva, at http://factiva.com/factivainsight/producttour/monitortour/insights01.html accessed 12/11/2010.

C.4. DISTRIBUTION VIA THE WEB

During the late 1990s, soon after the first Web browsers were introduced, newspapers were quick to adopt the Internet as a platform for distribution. Long before this, since the days of the telegraph, news reports and information had been distributed electronically over wires by the Associated Press. But the Internet made it possible for the first time to deliver structured newspaper content directly to individual consumers electronically. Obviously, substantial differences exist between the newspaper content a publisher issues in print and what appears under that same publisher's electronic masthead.

Back files: Once a locally produced news story is retired from the active or current pages of a newspaper's website, it is often posted as the "archived" version or "version of record" in a separate part of the Web. Some newspapers outsource maintenance of these archived stories and features to archiving services like NewsBank, NewspaperArchives.com, and ProQuest. These services add value by formatting and indexing the stories and presenting them in searchable databases, which are normally hosted by the archiving service, but made to appear seamlessly connected to the newspaper site.

The Wisconsin State Journal site provides access to two archives of back files: digitized back files of the WSJ print edition maintained by Newspaper Archives.com, accessible only to subscribers; and individual articles from the website presented as HTML text. A search of the HTML version yielded results back to 1969.

While workflow for the print newspaper is organized around a 24-hour news cycle, the Web is served a continuous stream of content. Locally produced news content is revised and uploaded by publishers to the Web not daily or weekly but throughout any given 24-hour period, and not in editions but headline by headline, article by article. The result is that much website content is dynamic rather than static.
It was evident even before our analysis that most news Web sites bear little resemblance to the corresponding newspapers and that what resemblance there is, is quickly fading. The most obvious difference is the incorporation of multimedia video, audio and other dynamic content in the Web edition. But even apart from these differences in format, the relationship between the contents of a newspaper's website and the print edition is rapidly evolving in terms of the nature of the content itself and its sourcing.

We conducted our study during a time of transition: when first-generation websites launched by newspapers like *The Arizona Republic* and the *Wisconsin State Journal*, and conceived as extensions of the print newspaper's legacy forms and formats, still survive alongside dynamic, Web-based models, such as *azcentral.com*, *madison.com, seattlepi.com* and *Chicagotribune.com*.

The relationship between the *Wisconsin State Journal* website and the larger and more dynamic *madison.com* site maintained by Capital Newspapers exemplifies this transitional status. *Madison.com* serves as a news gateway for Madison, Wisconsin, and incorporates content from other Capital Newspapers-owned publications, *The Capital Times* and *77 Square*.[17] The *Wisconsin State Journal's* Web presence is one of seven main "channels," on the *madison.com* site, each of which is represented by an icon on the site's landing page.

The seven channels include locally-produced content (the Wisconsin State Journal, the Capital Times, 77 Square, Sports, Obituaries, and Entertainment), Community-sourced information (Communities, Marketplace) and feeds from third-party providers. *Madison.com* displays content from each of the channels and each of the channels links to and displays content from others.

The orientation of the online WSJ is more predominantly local than the print edition, which carries a considerable amount of national news. Local news headlines display prominently in a sidebar near the top of the page, and all sections (News, Business, etc.) default to local news. In addition, the site's featured photo and headlines always focus on local stories. While the print edition of the *Wisconsin State Journal* may feature national stories from other newspapers and the Associated Press on the front page, local and state-related stories consistently dominate the landing page of the site.

Like many news sites *madison.com* employs social media technologies and Web analytics to mine user opinions, preferences, behaviors, and traits. The *madison.com* site devotes an entire channel, the Communities channel, to "crowd-sourced" content, where readers can post stories, videos, events and photographs.

The newer model of the news Web, however, is exemplified by *seattlepi.com*, the Hearst Seattle Media's "flagship site." Like the *Wisconsin State Journal* site, *seattlepi.com* also focuses heavily on information of local interest, such as crime, regional politics, and local sports teams. But *seattlepi.com* is even more fundamentally different from its now defunct predecessor, the *Seattle Post-Intelligencer* newspaper. It

[17] In April 2008, *The Capital Times* ceased print publication and became a web-only publication.

features not only original staff reporting and breaking news, but blogs by staff and readers, links to other journalism and news Web sites, community databases, and photo galleries. Through partnerships with other Seattle media (i.e., radio and television broadcasters), *seattlepi.com* also has access to video and audio produced by their local staff.

The site also serves as a feeder of local Seattle-area news to the Hearst News Service, which in turn functions as a kind of miniature Associated Press. The site, in turn, draws freely from a pool of news stories, features, business information, sports, and profiles from other Hearst newspapers. This consolidation of news media enabled by the Web is a phenomenon undermining the predominance of AP and has co-opted many of that venerable syndicate's sources of local news.

C.4.1 GENERAL DIFFERENCES BETWEEN PRINT AND ONLINE CONTENT

On the basis of our analysis of the test bed newspapers there are a few generalizations one can make about the difference between what still goes into the printed newspaper and what a news organization directs to the online reader. While *The Chicago Tribune* and *Arizona Republic* newspapers function as "fixed" or static records of civic events and containers of information of interest to their local constituencies, their websites function as regional portals to a much larger and more dynamic realm of text, image, and multimedia information, combining with the content produced in their local newsrooms content and applications produced by third-parties and even by consumers themselves. The presence of this content dramatically changes the nature and impact of the news reported and the experience of the user.

In general, we observed the following differences between the contents of print editions and the content of the same publishers' Web editions:

1. *Online content, including entire stories, is often absent from the print edition.* Because "space" is unlimited in the online environment, the newspaper's website content is often richer and more plentiful. Additional articles, multimedia presentations like audio recordings and videos, and "portfolios" or "slide shows" of photographs are the most common locally produced Web-only features. The enlarged capabilities afforded publishers by digital technologies enable newspapers to use assignment photographs that would, because of space limitations, have been relegated to "outtakes" of the print editions.

2. *Print content is often absent online.* While all *Chicago Tribune* print articles also appear online, certain stories appearing in the print editions of the *Arizona Republic* and *Wisconsin State Journal* (particularly news briefs) do not appear on their websites. Some articles appear both in print and online but are not included in the online "archives" of the paper's articles. Some photos and graphics included in the print editions of the papers are also absent in the online editions.

One driver of this variation is rights. Often absent from the online newspaper are wire service stories and syndicated material, political cartoons, photographs, financial data, and content provided by commercial sources or other news organizations like the *New York Times*. Newspapers often acquire only either the rights to publish such materials in print format, or one-time or limited use rights. On the other hand, the Web is not as facile at displaying longer explanatory pieces of journalism which are best viewed in the print (or tablet) environment.

3. *The timing of publication varies between print and Web.* The newspaper Web site is usually the first place breaking news stories appear. Many news stories appear online as soon as they are reported and are updated several times on line before appearing in print. Even a newspaper with a relatively modest Web presence like the *Wisconsin State Journal* posts new stories almost every hour on weekdays.

Through simple human analysis, we determined that different Web pages and sections have different updating schedules. For example, sports pages remain fairly static during the daytime; and business pages are rarely updated on Sundays. John Smalley, editor of the *Wisconsin State Journal*, reported that there is no formal updating schedule for WSJ web content, but that in general, the goal is to spread out news updates throughout the day in order to keep the site current. Stories on the *Wisconsin State Journal* website normally appear in the print edition the following day. (Similarly the *New York Times* print edition contains the final, edited and corrected versions of news story that appear earlier on the Times Web site, which in turn become the versions of articles posted in the Times online archive.)

4. *Multiple versions of articles appear online, but rarely in print.* Stories and information on news sites are updated frequently as events occur, while the print edition is obviously static and confined to the 24-hour cycle. Particular to "breaking news"—but not confined to that category—developing stories may be updated with additional content as stories unfold (the ongoing developments of the protests at the Wisconsin State Capitol is a prime example of constant news updating on the Wisconsin State Journal Web site). These updates represent multiple revisions of the same article, rather than distinct new articles.

Story updates are normally indicated by a timestamp and the word "updated." Our analysis determined that some updating timestamps, however, appear automatically without there actually being changes in content.[18]

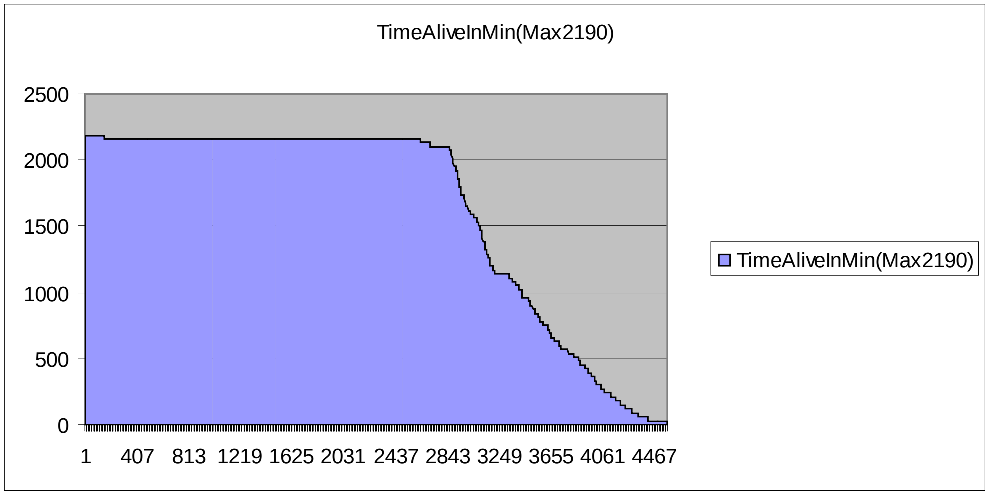
[18] For example, all of the stories displaying under the "State" tab in the side box of the Wisconsin State journal appeared with an "Updated 12:00pm" timestamp on October 25, 2010, although all the headline stories were from October 24, 2010, suggesting that this may be an automated feature.

The differences are due to a number of factors: financial, legal, and technological. While the ability of the Web to accommodate a wide variety of digital media enables newspapers to distribute more and richer content, the rights obtained or purchased by the newspaper for the use of particular content are often a drag on that new freedom.

C.4.2 THE VARIABLE LIFESPAN OF WEB CONTENT

In contrast to the relatively regular news cycles of the print era, there is little predictability or uniformity in the rate at which Web news content is changed or updated. Yet some generalizations can be made. A computer-assisted analysis of the *Chicago Tribune* Web site yielded a granular picture of the rate or "velocity" of updates on news Web sites, and how the rate or velocity varies for different types of content.

The analysis, undertaken by Kalev Leetaru at the University of Illinois, Urbana-Champaign, involved a crawl of all 105 "section root" pages on the chicagotribune.com site. (The "section root" pages list all of the stories on the rest of the site.) The chart below illustrates the number of page URLs against minutes of persistence for a two-day period.



The analysis showed that in general business, entertainment, and sports news tended to be updated most frequently (sometimes several times within the half hour), while features, opinion, travel, and blog content changed less frequently. Hence the difference between print and electronic versions of newspaper content will vary considerably by type of content. (A copy of Kalev Leetaru's report is attached here as Appendix A.)

C.4.3 SOURCING WEB CONTENT

Most newspaper Web sites draw upon three different types of content feeds:

1. *Native content,* i.e., locally produced or syndicated content aggregated in and output from a newspaper or parent company's own editorial and/or Web production systems. This content usually resides on various Web servers under the control of the newspaper or its parent company.
2. *Third party-produced content and applications* produced and hosted on Web servers maintained by specialized content providers such as AdTech, AccuWeather, Google Earth, Yahoo!, Monster.com, Brightcove, and Morningstar.[19]
3. *Reader or user-produced content*, i.e., comments, queries, photographs, videos, data, and other content, and links to other content supplied by readers through news web site blogs, polls, and other means.

In general Web content of the first type corresponds most closely to material included in the print newspaper than does third-party or user-produced content. This native Web content is normally generated in locally managed production system as part of the same work flow as print news. Third-party content, on the other hand, is provided largely by specialized organizations set up to serve dynamic content and real-time data specifically to the Web. AdTech, for example, a subsidiary of AOL Inc. based in Germany, provides an integrated technology platform for serving video ads to websites, mobile devices and other platforms. The ad videos that appear are customized by AdTech for a given end user, i.e., keyed to the reader's preferences or geographical location, which is determined through an interaction between the AdTech software and information (e.g., IP address, browsing history) stored on the individual user's device or browser.

It is true that traditional print newspapers have always incorporated third-party and user-produced content, such as wire service photographs and text, syndicated cartoons and editorial columns, and letters to the editor. But in the past that content has normally travelled through the newspaper's editorial process en route to publication. Today, much of the video advertising, financial data, and even feature article content “appearing” on the *Chicago Tribune, Seattle Post-Intelligencer,* and *Azcentral.com* sites is delivered directly to the user’s browser by the third-party service providers. Similarly, blog posts and posted videos are hosted in third-party blog applications like WordPress or reside on social media sites like YouTube.

Notwithstanding the preponderance of outside providers of content to newspaper Web sites, there is a notable emphasis on the *Wisconsin State Journal* and *Arizona Republic* sites on information of primarily local interest. This apparent paradox reflects the growing ability of major national and even international information aggregators to target and provide customized services to local audiences.

Locally Produced Content vs. Third Party Content

To estimate the pervasiveness of third-party news content on seattlepi.com, we examined content and links to articles (excluding advertisements) that appeared on the landing pages of the seven sections of seattlepi.com on November 1, 2010: Home, Local News, Nation/World, Business, Sports, Arts & Entertainment, and Life.

Based on this assessment, we estimate that approximately 80% of seattlepi.com content comes from either third party providers or readers.[1]

	Seattlepi.com	Third Party	Reader	% of all Reader + Third Party)
Home Page	46	78	39	72%
Local	57	93	6	63%
Nation/World	16	78	6	84%
Business	28	54	14	71%
Sports	15	66	38	87%
A&E	11	121	50	94%
Life	15	59	48	88%
TOTALS	**188**	**549**	**201**	**80%**

C.4.4 WEB PRODUCTION WORKFLOW AND SYSTEMS

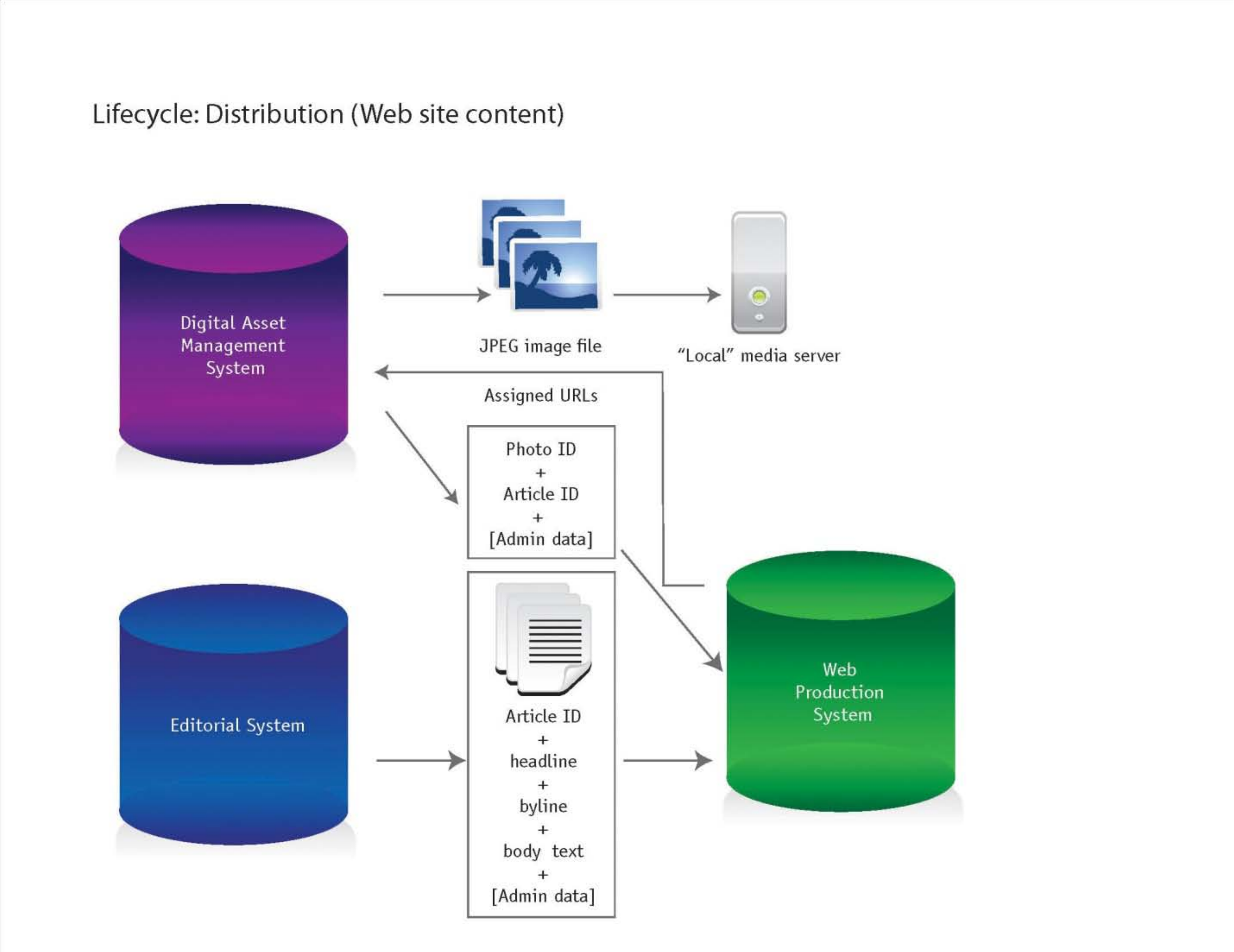
Systems: Tribune Company's Assembler, Nstein's Web Content Management, Town News's Blox Content Management System, Drupal.

The production of online news content represents perhaps the area of greatest change in traditional media organizations today. Production systems and workflows are being radically re-engineered to facilitate the transition from sourcing to distribution via a variety of media platforms (print, Web, mobile, and syndication). The new-generation sites like *seattlepi.com* are the product of newer integrated media management systems designed to speed news content to the Web and other digital applications first and create print editions as a secondary output. These systems deploy robust content management and social media technologies, and capabilities for the dynamic exchange of content and services with third-party providers of advertising, finance, sports, and entertainment information. The news industry is witnessing, in effect, a merging of the various production systems to accommodate the accelerating news cycle and growing demand for real-time information.

There are two basic types of systems for producing online news: 1) content management systems like Town News's *Blox Content Management System*, used by the *Wisconsin State Journal*, and the Tribune's *Assembler*, developed for use with conventional editorial systems like *Falcon Editorial* and CCI *NewsDesk*; and 2) "Web-first" authoring and content management systems such as CCI's *NewsGate* and Nstein's *Web Content Management* (WCM).

News organizations with the first type of system rely on a variety of discrete modules to manage the sourcing, production, layout, and online publishing. Most editorial systems (like *Falcon Editorial)* and some pagination systems (like *Layout Champ*) generate XML output of individual articles, tables and other feature content tagged with structural formatting codes for feeds to Web and wireless devices. "Versioning," or production of multiple versions of the same article for different media platforms, is done through the application of XML coding. This coding is generated and applied in a system's XML editor, and the application of HTML templates for dynamic design of Web pages enable layout staff to preview the content in newspaper, Web and Wireless Application Protocol (WAP) formats.

NewsGate exports an XML "story package" (text, photos, graphics, codes) to *Assembler*, the Tribune's own proprietary Web production system, where the package interacts with templates or style sheets containing the newspaper's pre-established design characteristics. Content to be published, i.e., HTML text, and graphic, video, and JPEG image files, is then posted on separate servers within the paper's domain.


Lifecycle: Distribution (Web site content)
Digital Asset Management System
JPEG image file
"Local" media server
Assigned URLs
Photo ID + Article ID + [Admin data]
Editorial System
Article ID + headline + byline + body text + [Admin data]
Web Production System

Drupal is a multi-purpose, open-source Web content management software used by many smaller newspapers like the *Arizona Republic* and *Wisconsin State Journal* to repurpose and customize print content for the Web. Drupal can exploit the NITF tagging of content applied in the editorial system, which identifies structural elements such as headers, teasers, and sections, and provides administrative information such as editor name, issue numbers, publishing dates, and links to related content. Using

XML coding editors can also set “purge” dates for articles, to “unpublish” them automatically at a given time or date.
Seattlepi.com, which produces content for the Web and electronic distribution only, uses the Nstein *Web Content Management* system (WCM) for creating and formatting its Web site content. This system is part of a suite of digital content production and management systems produced by Nstein for high-end publishing companies, and was implemented enterprise-wide by *seattlepi.com's* parent company Hearst Corporation. *WCM* “allows writers, editors and Web producers to create, edit and assemble content for multi-channel publishing.”

These Web production systems encode text, image and multimedia content in XML-based formats like NewsML or its predecessor NITF, industry standards developed to format and tag news content for exchange between systems and organizations. The XML tags identify and link the various elements of a story or article, such as headline, byline, teaser, and describe characteristics like version, filing date, rights status (often including a “purge” date). Tags applied in the Web production software also tag items to create linkages between pages with related content.

These tags interact with design templates created in proprietary Web production systems like the Tribune Company's *Assembler,* Nstein's *WCM* and with open source Web authoring systems like *Drupal*. Templates and style sheets (In formats like XSLT)[20] are developed by Web programmers for these applications, and determine the basic visual and functional layout of the Web site, like the number of columns on a landing page, and the relationship between the landing page and the other sections of the site like Sports, Home page, Business.

The Web production systems also name and post image and multimedia files associated with the article to various specialized Web servers, and insert the URL in the XML document for the article. News organizations use separate, specialized media servers for images, videos, graphics, and other multi-media and non-textual materials. Assignment of the URLs often follows a logical, hierarchal classification scheme that names and organizes files according to the major types of content, such as

[20] For an example of an NITF-encoded objects translated to a Web page, see http://www.iptc.org/site/News_Exchange_Formats/NITF/Examples/

"news," "articles," "columns," "images." Thus a photograph accompanying an op-ed column by E.J. Montini in the *Arizona Republic* bears the URL:

http://www.azcentral.com/columns/images/montini_t.jpg

The systems then output in various formats developed for the destination platform. For the Web the production system generates and posts on the local Web server HTML documents and style sheets (.html and .css files). The HTML documents include the texts of articles and other features to appear on the page, along with basic tags that indicate the kind of content included and provide basic formatting instructions, and the URLs for content (like photographs, pages, databases, and other media) residing on other servers. The style sheets cue the browser on visual layout of the page, determining where individual texts and photographs and features appear, to display the content in the distinctive visual formats established by the Web production system for the site.

Scripts developed in the systems cue the browsers to display text and other content, and to query local and third-party hosted databases and content servers to return content from those sources to the browser display.

Some content on the Web site is generated and updated by databases maintained by the newspaper. Such databases are often locally built, using open source or proprietary software, and are programmed to interact directly with the Web production system to post content to the Web site automatically in real-time article-by-article or batch updates. [21] Other database-generated content, like the lists of "Most Read" and "Most Emailed" features that are a commonplace on newspaper landing pages, are generated using databases incorporated in the Web production system. Thomson Reuters runs the database behind the New York Times "Most Emailed / Recommended for You" feature. Such "personal preference engines" automatically identify and point to articles, columns and other features of the news site related or analogous to items accessed previously by the viewer. These databases rely on profiles of a given viewer assembled by the database using upon information on the viewer's browsing history.[22]

[21] For the New York Times web submission process, see Jacqueline Maher, "Building a better submission form," New York Times Open blog, May 25, 2010, , http://open.blogs.nytimes.com/2010/05/25/building-a-better-submission-form/ accessed 1/25/2011.

[22] "NYTimes Recommendations creates a personalized list of recommended reading, pointing you to NYTimes.com content we think you'll find interesting. Recommendations are based on what you've read on NYTimes.com (both the standard site and the mobile site). Recommendations are generated approximately every 12 hours, to ensure that you see new recommendations each day. Each NYTimes.com article is assigned topic tags that reflect the

Some visual content appearing on a newspaper's website is generated from databases, using data visualization applications. The VIDI software developed for Drupal creates dynamic charts and graphs that enable viewers to interact directly with the database.[23]

External data feeds are also set up to send content and updates directly to the newspaper's editorial system. Providers like Dow Jones, AccuWeather, and Major League Baseball send information in NewsML, SportsML, and other XML formats continually to the newspaper's editorial system where it is automatically set into the local style sheet and uploaded to the Web. Updates of this type of information are sometimes only seconds apart.

News Web site content is enhanced for searchability in a number of ways. Many news publishers now use third-party data indexing services to perform subject and semantic analysis, tagging and indexing of their Web content. In some instances the tags and annotations are created upstream of Web production, in the digital asset management systems. Wire services like Reuters and Associated Press provide this kind of data tagging service for news organizations. Since 2008 AP members have been encouraged to submit their news text, audio and image content for tagging in AP's enhanced metadata suite under the AP Content Enrichment Initiative. The content is then enriched with the APPL tag set and returned to the news organization for further use.[24] Reuters *Open Calais* service also provides rich subject and name analysis and tagging and annotation of content files for client news organizations.
A certain amount of content enrichment can also occur in the newer Web production systems. Nstein's *Text Mining Engine* and *Web Content Management* systems, for example, ingest and automatically

content of the article. As you read articles, we use these tags to determine your most-read topics. NYTimes Recommendations uses browser cookies to build your reading history and determine your top sections and topics. NYTimes Recommendations uses your NYTimes.com reading history to identify your most-read topics and sections. Based on what you've read in the past 30 days, NYTimes Recommendations suggests additional content you might like. The more you read, the more accurate the recommendations become." http://www.nytimes.com/content/help/extras/recommendations/recommendations.html#recommendationsq01 accessed 4/21/11

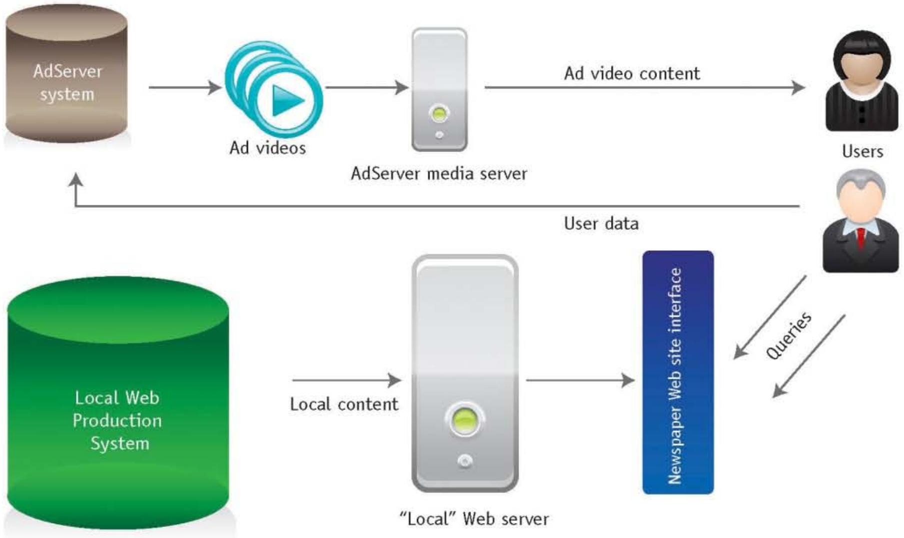
[23] For the widely used VIDI tools developed by the Jefferson Institute for Drupal, see the VIDI project site at http://www.dataviz.org/

[24] For more information on the AP's Content Enrichment program see http://www.ap.org/contentenrichment/index.html

analyze unstructured text, identifying context, meaning, personal and corporate names, categories, entities and even sentiment. The systems then return standardized, structured XML text that is optimized for discoverability within the publisher's site and on the open Web.

Many third party providers, such as advertising and analytics services, serve content directly to the Web or, more accurately, to the user's browser. These services capture information from the user's browser, identifying preferences, geographic location(s) and transaction histories associated with the user's IP address. This information in turn prompts an algorithm to determine what new ad content to display to the user.

Lifecycle: Distribution (Third-party content)



C.4.5 SECONDARY WEB DISTRIBUTION NETWORKS

Once exposed on a news organization's website, news content is discovered and viewed by millions of individual readers. This content often finds its way into secondary distribution networks. Many users in fact discover news information through intermediary agents and applications -- the Huffington Post, the Drudge Report, Google News, Twitter, and FaceBook -- rather than directly from the news sites themselves. These services harvest current news content exposed on the open Web to provide timely specialized information to subscribers. This activity is an outgrowth of the accelerated news cycle and the growing demand for real time information.

"News reader" services like *NewsRover, LibreAccess, FastTrack News*, and *Google Reader* employ their own proprietary search engines and Web crawlers to identify and retrieve relevant syndicated news from various news publishers. These services continually download and annotate that content and generate customized web feeds in their own XML-based RSS and Atom-formats. Subscribers to these services then view the content in customized displays on their computers, mobile phones, tablets, and other devices. *FastTrack News* is a portable usenet client and RSS feed reader, blog client and news aggregator. Once subscribed to particular news feed, the *FastTrack News* software checks for new content at user-determined intervals and retrieves updated data.

While many publishers have varying degrees of control over whether and how their content appears in the news reader networks, some publishers strive to better expose their news content to mobile and tablet device platforms like Apple's iPhone and iPad through customized applications. These applications, or "apps," are software programs that prompt interaction between digital data and content exposed by a publisher or media organization, and the operating system of a particular digital device or search engine. They enable, for example, the import, organization and display of content from a news database, and automatic updates of that content, on a smart phone. Other apps are produced independently by programmers and made available on the Web through a number of App sources, such as Apple's App Store and Google Apps.

The proliferation of platforms and distribution networks for news presents daunting challenges to the capture and reproduction for future researchers how the typical user experiences today's news. The number of versions of essentially the same news content that are generated will make it very difficult to characterize the “typical” user.

CHICAGO TRIBUNE

CONTENT VELOCITY ANALYSIS

KALEV LEETARU

OVERVIEW

This report presents the findings of a small pilot study examining content velocity on the Chicago Tribune's website, http://www.chicagotribune.com/. Conversations with the Tribune indicated they were unable to provide detailed metrics on content additions to their website, so a custom crawling script was deployed to download the HTML contents of the 105 "gateway" pages (the article listing pages of each section) every half hour for 34 days, from 9/15/2010 through 10/19/2010. The resulting 136,605 snapshots were used to characterize the paper's linking velocity over this period.

Primary findings are that 83% of the Tribune's links were to the DoubleClick.net advertising network, with just 11% of links pointing to Tribune pages with an average link lifespan of 56 hours and a range from 18 hours to 7 days. Only 5% of pages were linked to from more than one gateway page, suggesting the Tribune requires users to employ extensive subnavigation to locate all stories, and crawling-based archival systems must utilize more extensive crawling strategies to locate all content. No master "latest content" portal pages or RSS feeds are available, meaning that a page-by-page crawl of all top- and sub-level gateway pages is required to compile a list of new pages, greatly increasing the resource requirements for archiving the Chicago Tribune.

Roughly 39% of Tribune URLs are linked for a day or less and an average of 735 new links are posted each day to Tribune content. The highest number of new links are added on Thursdays, while Sundays have the fewest updates. Content sections exhibit strong stratification in number of links, percentage of links directed to other Tribune pages, and link lifespan, suggesting that content characterizations must be topically-oriented on large news sites, rather than site-wide.

Archiving large news websites like the Chicago Tribune require considerable resources to discover new content as it is published and queue it for archival. Unlike traditional websites which may be large, but have only a small portion of their content change daily, news websites are by definition highly dynamic. They can add hundreds or thousands of new pages each day, make changes to historical content on an ongoing basis, and operate on a 24/7/365 schedule, making them worse-case scenarios for archiving. Further complicating matters, many outlets like the Tribune do not maintain master chronological lists of their content, relying instead on a single table of contents page for each primary content section. As links are expired off this contents page, the articles themselves remain on the site, but a reader or crawler's ability to discover that link is substantially hindered, requiring knowledge of the article in question or keyword searching of the site. Indexing all new content published to the Tribune web site each day requires crawling its 105 content gateway pages at a fast enough speed that it is able to catch new URLs even during breaking news situations when multiple new articles an hour may be published.

To assist archival crawlers, news websites can provide chronologically-sorted RSS feeds that list the most recent 50-100 URLs updated. Some sites provide one RSS feed per content section, while others provide a "master" RSS feed that covers the entire site. An archival crawler simply has to download a copy of this single master RSS feed to access a machine-readable list of the last 100 URLs to have changed on the site, making quick work of downloading the latest updates. Such RSS feeds have the added benefit that human users of the site can use them to keep on top of breaking updates, improving the overall user experience of the site. A second option is the use of specialized sitemaps, such as the Google News Sitemaps service, [25] which uses a customized XML format [26] to help guide crawler traversal of a site.

ANALYSIS

During the 34 day monitoring period, there were a total of 366,290 links in the monitored gateway pages. Of these, 304,684 (83%) were links to the DoubleClick.net advertising network, while 41,220 (11%) were to Chicago Tribune pages, and the remaining 20,386 (6%) were to other external sites, including Tribune partners. That more than 83% of the links across the Chicago Tribune's website during this period pointed to paid advertising demonstrates the importance advertising plays even in mainstream online news websites, and the density of paid commercial messages against the journalistic news content developed by the papers themselves. Note that this study did not attempt to measure the screen real estate in pixels or percentages occupied by advertising versus journalistic content, but such a high link density would suggest a significant amount of link space is dedicated to advertising content.

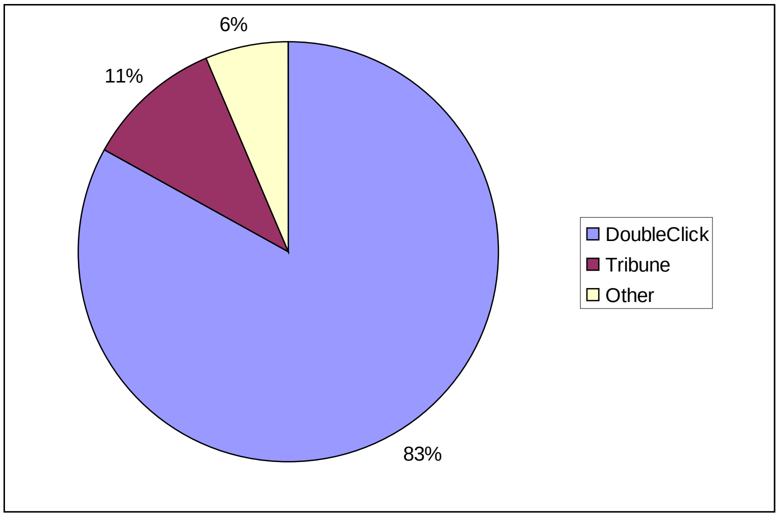


Figure 1 - Breakdown of Chicago Tribune links by target

Appendix A, Domains by Links, contains a complete list of all domains linked to by the Chicago Tribune and the number of distinct URLs linked to at that domain. Of particular note is that several of the Tribune's partner sites are among the most-linked-to domains, including chicagobreakingsports.com and chicagonow.com. Two other papers in the Tribune stable, the LA Times and the Orlando Sentinel, are frequently linked from the Tribune. In the world of print, a newspaper very rarely runs a story from a primary competitor, yet in a strange twist, in a one month period, the Tribune forwarded its visitors to more than 515 articles on the Sun Time's site, making it only the second most-linked to external domain, behind Twitter. In all, the Tribune linked to URLs hosted by 458 different domains and subdomains.

Just 1,950 of the 41,220 Tribune links (5%) were linked to from more than one gateway page, and those that were tended to be links to other main pages like links from sub-level pages such as /sports/cubs/ back to the main /sports/ page. This is a critical finding from the standpoint of archiving the Tribune, as it suggests that

[25] http://www.google.com/support/news_pub/bin/answer.py?hl=en&answer=75717

[26] http://www.google.com/support/news_pub/bin/answer.py?answer=74288

to fully collect the entirety of the Tribune's content, crawling only root-level pages like /sports/ will not be enough: it will be necessary to crawl every single sub-level page like /sports/cubs/, /sports/football/, etc. This significantly increases the resources necessary to archive the paper.

Appendix B, Tribune URLs by Lifespan, shows all 41,220 Tribune URLs linked from the monitored gateway pages during the 34 day monitoring period, ordered by the number of minutes the URL was linked to. Since snapshots were taken every 30 minutes, lifespan is measured in multiples of a half hour, with 30 minutes indicating the link was found in only a single snapshot (likely a high-velocity content section), while those with long spans measuring in days indicate URLs that were kept linked from the site for long periods of time. Note that the lifespan of a URL measures only the length of time a link was kept to it from one of the 105 monitored gateway pages, not the length of time the article was kept on the Tribune site. In most cases, pages remain on the site indefinitely, and access transitions from gateway links to requiring keyword search engine use for discovery.

There were 2,029 URLs (5%) linked for 30 minutes or less, while an additional 928 were linked for 60 minutes or less. Nearly 39% of URLs (16,249) were linked for a day or less. The figure below shows the total number of URLs at each lifespan duration. There appears to be fairly even distribution of lifespans beyond the one day mark, with the only outlier being a strong clustering at a duration of 6 days, suggesting a weekly rotation cycle for some Tribune content sections. That 61% of the Tribune's content is nearly evenly distributed among linking lifespans (as opposed to primarily being clustered in the 1-3 day range) suggests that the Tribune uses a management strategy emphasizing using a large number of gateway pages to keep links online, as opposed to a small number of high-velocity linking pages.

Eliminating URLs linked for two or more weeks (these are likely gateway page links or links to standing features) leaves 30,226 URLs (73%), with a mean link lifespan of 56 hours.

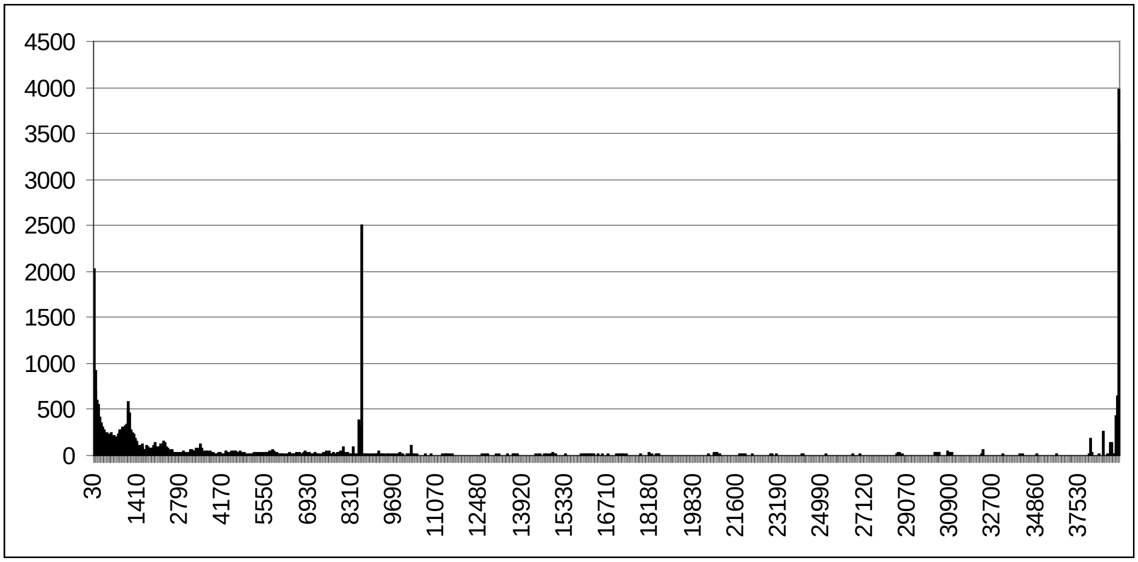


Figure 2 - Histogram of number URLs updated by time interval

Looking only at Tribune links, the figure below shows the total number of new links added to the site each day after the start day on 9/15/2010, exhibiting strong weekly stratification. There is an average of roughly 735 new links per day. While the Drudge Report had the most updates on Tuesdays and the fewest on

Saturdays, [27] the Tribune has the highest number of new pages added on Thursdays, with the fewest new pages on Sundays. A linear trendline has been plotted through the center of the timeline indicating that the Tribune appears to have reduced its update rate in nearly linear fashion throughout the duration of this analysis. It is unclear whether this is a longer-term reduction in the Tribune's update velocity, or whether it simply reflects a specific news cycle profiled during this period.

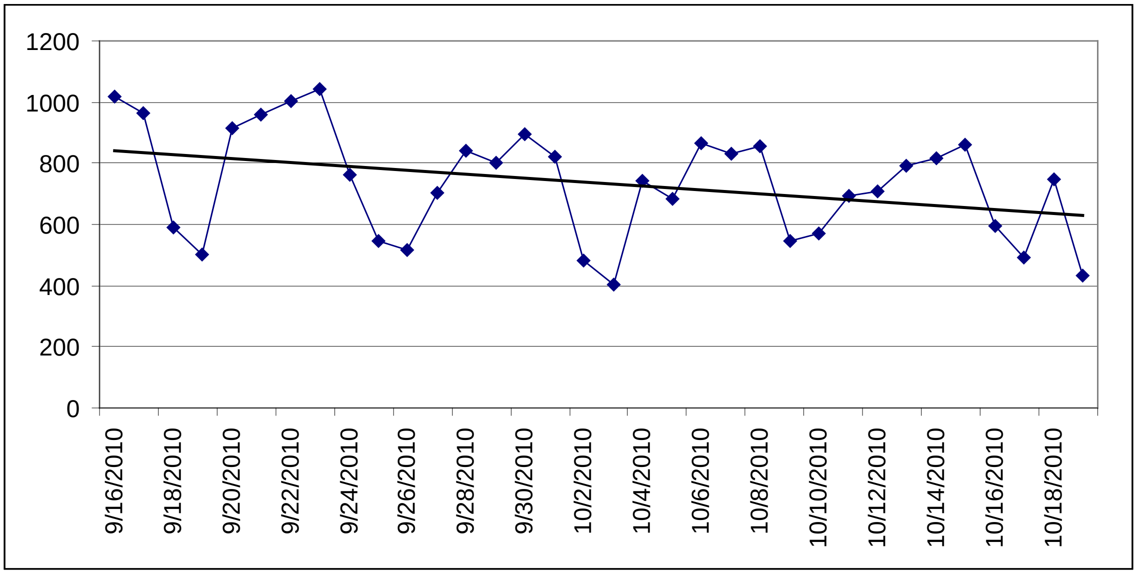


Figure 3 - Number new Tribune links seen by day

Appendix D, Stats by Section, contains a breakdown of each of the 105 gateway pages and the total number of links to all sites, number of links to other chicagotribune.com pages, number of Double Click advertising links, and average lifespan in hours of chicagotribune.com links, for each section. The blogs and special elections section of the Tribune have very little advertising, high link lifespans, and low content volumes. The only three sections to have sub-24-hour lifespans are the main Business, Sports, and Celebrity sections. Even the front page has a link lifespan of 25 hours, suggesting relatively slow turnover on featured stories, despite that section having the highest total number of links. Among individual content sections, the Sports section has the highest number of links at 9,940 (3,384 to Tribune pages), nearly one thousand more than the business section. Ranking by total number of links to Tribune pages, as opposed to all links, Nation & World news comes in at third, followed by Entertainment and then Local News.

Ranking sections by the percentage of their links that point to Tribune pages versus external properties, the special elections.chicagotribune.com section was highest at 95%, followed by the Tribune's major blogs. The highest non-blog section was its Sports page at 34%, its main page at 29%, and its Nation & World news at 28%. The pages with the lowest density of Tribune links are several of its travel sections with less than 4%. That no major content section has more than a third of its links pointing to Tribune pages shows the tremendous role external resources and advertising play on the Chicago Tribune's website.

Table 1 - Gateway pages ordered by average link lifespan

Section	Total	Tribune	%Tribune	DoubleClick	Lifespan
/business/	9003	2180	24.21	5134	18.97
/sports/	9940	3384	34.04	5192	20.21

[27] http://www.uic.edu/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2500/2235

/entertainment/celebrity/	5403	1090	20.17	3900	22.08
/features/horoscopes/	5503	244	4.43	5200	25.08
/	13030	3855	29.59	5200	25.48
/technology/deals/	3345	689	20.60	2602	25.61
/news/nationworld/	5704	1628	28.54	3900	29.38
/news/local/chicago/	4927	426	8.65	3900	29.77
/sports/football/bears/	5310	1083	20.40	3903	35.15
/sports/college/	5362	1118	20.85	3897	35.19
/health/	5394	789	14.63	4096	36.19
/news/education	4477	464	10.36	3894	36.62
/news/opinion/blogs/	8041	735	9.14	3892	37.48
/news/opinion/share/	5573	345	6.19	5168	38.01
/entertainment/	5762	1458	25.30	3848	38.10
/news/politics/	4710	619	13.14	3903	38.84
/news/local/	6309	1057	16.75	3879	43.35
/news/opinion/	5037	831	16.50	3900	45.79
/news/columnists/all/	4867	908	18.66	3898	47.37
/news/columnists/all/	4867	908	18.66	3898	47.37
/news/corrections/	4301	337	7.84	3903	50.16
/sports/baseball/whitesox/	5028	850	16.91	3900	50.73
/sports/highschool/	4847	838	17.29	3798	52.01
/sports/hockey/blackhawks/	4808	696	14.48	3903	53.07
/features/columnists/	5627	359	6.38	5204	53.35
/sports/baseball/cubs/	4966	808	16.27	3903	55.50
/travel/	5827	582	9.99	5175	59.53
/sports/basketball/bulls/	4636	501	10.81	3900	61.74
/features/food/	4492	414	9.22	3903	63.51
/news/opinion/commentary/	4676	501	10.71	3894	64.61
/news/opinion/editorials/	4397	434	9.87	3903	67.04
/features/gardening/	4290	318	7.41	3900	71.08
/business/yourmoney/	5880	255	4.34	5204	71.90
/entertainment/tv/	4963	407	8.20	3901	73.22
/business/columnists/	5566	307	5.52	5200	73.65
/features/style/	5617	347	6.18	5204	79.39
/features/tribu/	4380	432	9.86	3887	80.27
/business/investors/	2847	197	6.92	2600	80.98
http://newsblogs.chicagotribune.com/clout_st/	429	390	90.91	2	83.75
/travel/escapes/	4263	277	6.50	3900	84.22
/entertainment/dining/	4483	404	9.01	3903	84.73
/sports/golf/	4339	376	8.67	3903	88.00
/news/local/suburbs/	1539	291	18.91	858	90.48
/sports/tennis/	4167	213	5.11	3903	93.86
/sports/smack/	6045	365	6.04	3900	95.24
http://newsblogs.chicagotribune.com/tribnation/	486	400	82.30	2	95.80
/entertainment/events/	4407	376	8.53	3903	96.14
/news/opinion/letters/	4552	589	12.94	3903	97.72

/business/smallbusiness/	5493	234	4.26	5200	99.85
http://elections.chicagotribune.com/	148	142	95.95	2	100.53
/features/family/	5619	244	4.34	5200	101.67
/news/local/southsouthwest	4381	230	5.25	3897	102.07
http://leisureblogs.chicagotribune.com/the_theater_loop/	535	455	85.05	5	108.16
/news/watchdog/	4205	239	5.68	3903	110.04
/features/games/	5451	199	3.65	5196	110.67
/health/agentorange/	2825	177	6.27	2598	114.83
/news/strange/	4566	601	13.16	3903	115.10
/entertainment/movies	4458	354	7.94	3900	115.47
/news/local/northnorthwest	4380	235	5.37	3900	124.02
/business/problemsolver	5608	298	5.31	5200	125.72
/travel/midwest/	5416	166	3.06	5194	127.27
/features/lottery/	5416	155	2.86	5204	127.47
/travel/family/	4156	202	4.86	3897	128.26
/sports/soccer/	4214	253	6.00	3903	128.29
/news/local/suburbs/evanston/	1517	176	11.60	866	128.56
/news/local/suburbs/joliet/	1512	172	11.38	866	129.76
/news/local/suburbs/arlington_heights	1457	170	11.67	863	131.12
/news/local/suburbs/tinley_park/	1404	170	12.11	866	131.49
/news/local/suburbs/wheaton/	1532	170	11.10	866	131.51
/news/local/suburbs/orland_park	1396	171	12.25	866	131.62
/news/local/suburbs/wilmette-kenilworth/	1379	173	12.55	863	131.82
/news/local/suburbs/northbrook	1470	171	11.63	863	131.95
/news/local/suburbs/libertyville	1514	171	11.29	866	131.96
/news/local/suburbs/schaumburg/	1448	168	11.39	866	132.19
/news/local/suburbs/crystal_lake/	1475	168	11.31	866	132.19
/news/local/suburbs/des_plaines/	1485	168	11.19	866	132.19
/news/local/suburbs/glen_ellyn/	1501	168	11.60	866	132.19
/news/local/suburbs/plainfield/	1497	169	11.29	866	132.26
/news/local/suburbs/deerfield	1393	172	12.35	866	132.26
/news/local/suburbs/elgin/	1481	169	11.40	866	132.27
/news/local/suburbs/bolingbrook/	1483	169	11.41	866	132.27
/news/local/suburbs/grayslake/	1391	170	10.81	866	132.34
/news/local/suburbs/gurnee/	1436	170	11.00	866	132.34
/news/local/suburbs/elmhurst	1545	170	12.22	866	132.34
/news/local/suburbs/downers_grove/	1573	170	11.84	866	132.34
/news/local/suburbs/hinsdale	1449	170	11.73	866	132.34
/news/local/suburbs/naperville	1619	170	10.50	866	132.35
/travel/deals/	5414	164	3.03	5200	132.45
/news/local/suburbs/glenview	1453	172	11.84	866	132.48
/news/local/suburbs/oak_park-river_forest	1650	172	10.42	866	132.48
/news/local/suburbs/winnetka-northfield/	1382	172	12.45	866	132.48
/news/local/suburbs/highland_park-highwood/	1452	173	11.91	866	132.55
/news/watchdog/nursinghomes/	4148	192	4.63	3900	133.18
/entertainment/theater/	4306	258	5.99	3900	133.33

/travel/other/	5459	206	3.77	5196	135.91
/travel/unitedstates/	5465	208	3.81	5200	136.04
http://featuresblogs.chicagotribune.com/printers-row/	123	79	64.23	2	140.56
http://leisureblogs.chicagotribune.com/turn_it_up/	443	406	91.65	2	142.45
/travel/chicago/	5453	196	3.59	5197	146.86
/news/local/west	4429	217	4.90	3900	146.86
/news/local/suburbs/barrington/	1553	169	10.88	1030	154.80
http://featuresblogs.chicagotribune.com/features_julieshealthclub/	679	378	55.67	2	162.62
http://newsblogs.chicagotribune.com/towerticker/	378	270	71.43	2	163.84
/features/askamy/	4071	238	5.85	3768	168.75
http://leisureblogs.chicagotribune.com/taking-off/	165	90	54.55	2	172.70

INTEGRATED EXTERNAL LINKS AND JAVA SCRIPT

Some sections, such as Julie's Health Club [28] have a standard set of external links provided for each posting. For the Health Club features blog, each post has a list of keywords at the bottom with links to a Technorati search for each, along with a link to a Technorati search of all external blog posts linking to that blog post. This results in a high density of external linking in these content areas. Since these links are part of the overall "experience" of those content sections, and to help distinguish content sections that rely more or less heavily on these links, they are considered along with all other links for a given section. Further investigation of the site suggests that these "integrated" external links are present only on the Tribune's small number of hosted blogs and are not used in its primary content sections.

In addition, Health Club uses an embedded Java Script widget from Outbrain.com to display a list of external sponsored advertising links beneath each posting based on its content (known as "contextual advertising"). Downloading, executing, and evaluating the output of embedded Java Script blocks like this was beyond the scope of this study, but suggests that dynamic content generated on-the-fly when a user visits a page is on the rise among newspaper websites, especially for contextual sponsored advertising content.

PROCESS AND METHODOLOGY

At the start of the project, Bill Adee, former Digital Editor of the Chicago Tribune and current Vice President of Digital Development and Operations for the Chicago Tribune Media Group was contacted to see what details the Tribune itself could offer on velocity and churn rates. Email correspondence with Mr. Adee [29] suggested that 100% of the Chicago Tribune's print content is reproduced on its website, http://www.chicagotribune.com/, and that 15% of its web site's contents is only available online (such as blogs). When asked "How often are web-based articles updated, to add new information, correct errors, etc? What percentage of its online content is later edited to reflect updated information, as opposed to a new story being issued with the updated details", Mr. Adee responded that online stories were "updated often," especially on their partner Chicago Breaking News Center site (http://www.chicagobreakingnews.com/). In

[28] http://featuresblogs.chicagotribune.com/features_julieshealthclub/

[29] Personal correspondence with Bill Adee 9/7/2010.

further correspondence, [30] Mr. Adee noted that there was no central RSS feed for the entire Chicago Tribune website because multiple content management systems are used to populate the site. He was unable to provide more detailed indicators on specific rates of change within each content section.

Given that the Chicago Tribune was unable to furnish specific detail on the velocity of change on its site, it was necessary to set up a targeted web crawling system to externally monitor the rate of new content posted to the site. The size of the Chicago Tribune website and total volume of content made it infeasible to attempt to download the full article contents of the entire site to monitor for individual article change to examine how often articles are updated after posting. A random selection of roughly 50 URLs from across sections, however, showed no content updates to those articles during a 24 hour period, so it is unclear how widespread post-updates to article content may be, and it is likely that this may be concentrated more heavily on their partner "breaking news" properties as opposed to the flagship http://www.chicagotribune.com/ domain. More thoroughly measuring this phenomenon would require mass-downloading of large portions of the Chicago Tribune's site on an hourly basis, which would entail significant resource consumption and thus was outside of the scope of this study.

Initially, it was hoped to use the site's RSS feeds as a quick mechanism to identify all new content across the site. Many news websites offer site-wide master or section-wide master RSS feeds that contain a complete list of all new content of the current day arranged in chronological order. Unfortunately, the Tribune's RSS feeds [31] do not function in this manner and appear to be ordered based on popularity, rather than publication/posting date. For example, the "Latest News" feed, which would presumably contain the most recent stories posted to the site, included only 10 stories at 9:45AM on November 8, 2010, with four of the stories dated the previous afternoon and the other six ranging from 4AM to 9:40AM November 8th. Other feeds examined showed similar ordering unrelated to post date and did not reflect the majority of content contained on the table of contents pages of each section. Thus, it was determined that the RSS feeds would not suffice to measure content velocity and a full-fledged external web crawl was necessary.

Given the size of the Chicago Tribune's site and the need for tight-interval regular snapshots of the site's structure, it was decided to use a "gateway page scan," which downloads only the primary "table of contents" ("gateway") pages of each content section (such as the front page of the sports section and its links to all of the sports stories of the moment) and measures update velocity based on the average lifespan of a secondary page linked from that gateway page. In other words, the rate of change of the sports section is determined by downloading the main sports front page (/sports/) every 30 minutes and examining the list of links on that page. In conversations with CRL, it was decided to monitor only the Chicago Tribune's flagship property http://www.chicagotribune.com/ as opposed to its myriad partner properties in order to make results attributable solely to the Chicago Tribune's editorial controls, eliminating the role that partner organizations would play in the update cycles of affiliated websites.

Given the complex site layout of a typical Chicago Tribune web page, and the need for high accuracy on measuring rate of change, it was decided that a manual review of the site would be used to identify the gateway pages, rather than an automated content characterizer. Automated content characterizer crawlers use algorithms that balance the size of the page's textual content against the density of internal links and the overall site's link structure to determine if a page is likely a news page or a table of contents page. Such algorithms must crawl the entire site in order to generate baseline data and to find all possible gateway pages, which was not feasible for this study given the size of the Tribune's site.

[30] Personal correspondence with Bill Adee 9/7/2010.

[31] http://www.chicagotribune.com/services/rss/

Instead, a manual review of the Tribune site was used to identify all 105 gateway pages, such as http://www.chicagotribune.com/news/local/. A complete list of all pages is included in Appendix C. In many cases, both a root-level page such as http://www.chicagotribune.com/sports/ and a sub-level page like http://www.chicagotribune.com/sports/baseball/cubs/ were monitored independently. A manual examination of root and sub-level pages showed that the Tribune displays only a small number of the top stories on its top-level pages, leaving the majority of its content in each section to be linked only from the sub-level pages. This required archiving top-level and sub-level gateway pages individually.

A customized crawling script written in PERL and using the LWP networking library [32] was written to download the HTML content of all 105 gateway pages every 30 minutes and this was run for approximately one month (34 days), from 9/15/2010 through 10/19/2010, with a 3 second delay between each page fetch to reduce load on the Tribune's servers. A TCP timeout of 20 seconds was set for the crawling agent, but never triggered, indicating the Tribune's servers were responsive during all crawling runs. To ensure returned results matched those seen by a human user, robots.txt adherence was disabled in the crawlers and an HTTP User Agent field was transmitted mimicking a Macintosh user with the Google Chrome browser. Aggregate network latency resulted in an average of 43 snapshots per day, as opposed to the theoretical 48, for a total of 1,301 snapshots of each of the 105 gateway pages, or 136,605 total snapshots covering the entirety of the http://www.chicagotribune.com/ site over a 34 day period.

CONCLUSIONS AND FUTURE WORK

This study has examined a 34 day period of the Chicago Tribune, examining the rate of change and linking patterns of its flagship online property http://www.chicagotribune.com/. The lack of a master chronological list of new and updated content via RSS, portal page, or other mechanism is a significant hindrance not only to scholarly archival and analysis of the paper, but, more critically, for the Tribune itself. If Tribune readers must wade through multiple sub navigation pages to view all articles under a given content section, rather than subscribing to an RSS feed or other push mechanisms to receive a list of all new content each day, that makes it difficult for its readership to fully consume and utilize its content. This is by no means limited to the Tribune and reflects a larger problem with many news media websites.

Resource limitations prevented this study from examining the possibility of individual articles changing over time through post-updates as opposed to new articles being released, and no attempt was made to characterize how article content or positioning influenced linking lifespan, though these could be examined in future studies.

Overall, the picture of a modern mainstream news website reflects one heavily dependent on advertising revenue, integrating paid sponsored links throughout its web properties in excess of 83% of all links, with the average content link remaining on the site from 18 hours to 7 days. Content sections exhibit significant stratification in linking behavior, suggesting site-wide characterizations are less useful for understanding large news media websites and a more intricate content-specific characterization approach must be utilized. In conjunction with a previous longitudinal study of the Drudge Report, [33] this study demonstrates the considerable insights that can be derived from a media property through external monitoring and the utility of customized crawling activities in informing both media research and archival.

[32] http://search.cpan.org/~gaas/libwww-perl-6.02/lib/LWP.pm

[33] http://www.uic.edu/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2500/2235