



## **Repository Profile**

### **The Associated Press**

By Victoria McCargar<sup>1</sup> and staff at CRL  
Version 1/21/11

### **About the Long-Lived Digital Collections Case Studies**

With funding from the National Science Foundation, CRL undertook a two-year analysis of eight established, “long-lived” collections of digital data and content. These case studies built upon the TRAC criteria for trustworthy digital repositories and the audits of the Portico, LOCKSS and ICPSR repositories conducted by CRL in 2006-2010 to test and refine those criteria.

The CRL case studies serve a different purpose than the aforementioned audits. While the audits probed the soundness of repository organizational and technical infrastructure, the case studies will identify practices, strategies and mechanisms that have enabled repositories to sustain massive digital collections over substantial periods of time.

The Associated Press e-AP repository is the subject of one of the studies. AP is the world’s largest and oldest news organization.

The present profile of the Associated Press will be updated periodically, as CRL further examines AP archiving practices and strategies, past and present.

---

<sup>1</sup> Researcher, consulting archivist and onetime telegraph editor.

## Introduction

In an increasingly connected world, the Associated Press is redefining the meaning of the word “ubiquitous”: The latest edition of its venerable *Stylebook* claims that half the world’s population has access to AP’s content every day.<sup>2</sup> **As the oldest and largest news-gathering organization in the world, AP content shows up on websites, Blackberrys, cell phones, TV-screen news crawls and street-corner digital tickers—wherever up-to-the-minute breaking news is a sought-after commodity.** But equally important is AP’s continuing presence in its historic domain—the traditional media of radio, television and especially newspapers. Today, AP boasts more than 240 bureaus and 4,000 employees worldwide, an unparalleled global reach.<sup>3</sup> The only Western news organization in Pyongyang is, of course, AP.

With a history that stretches back to the middle of the nineteenth century, AP journalists have written and continue daily to write about historic events beyond number. Not only are its print and paper archives a trove of unique, primary source material, but the archives of other news organizations that were and are AP subscribers comprise a vast amount of AP content as well. **The famous “AP bug”—a device dating back to the days of the linotype<sup>4</sup>—identifies AP content in literally miles of newspaper microfilm all over the world.**

Since the early 1960s, the exploitation of electronic delivery has expanded AP’s text- and image-centric data stores to include sound, moving images, animated Web graphics, and sophisticated tabular data.<sup>5</sup> Massive storage and computing installations around the globe have on hand at any moment more than 100 terabytes of news in formats ranging from text to animated graphics.

Despite this history and ubiquity, though, AP’s own news archives do not comprise an uninterrupted historic record. As will become clear, there is historic AP material in a lot of places, but most of it isn’t readily accessible. This study investigates why that is, and how—with new initiatives in place and more on the horizon—recent history may fare better in the AP record.

---

<sup>2</sup> Associated Press. “Filing the Wire.” *Associated Press Stylebook and Briefing on Media Law*. Ed. Norm Goldstein. New York: Basic Books, 2007, 419.

<sup>3</sup> <http://www.ap.org>

<sup>4</sup> In its original form, the letters A and P were combined into a single glyph enclosed in parentheses. It existed as a discrete character on a linotype machine. As hot lead type gave way to photocomposition in the 1970s, the “bug” lived on briefly as a glyph but was eventually abandoned in favor of the two letters and parentheses.

<sup>5</sup> The Associated Press is the official aggregator and distributor of election data for the United States.

I have focused on the text archives in this report, but the archives in photography, audio, and video collections would show many of the same processes at work. Video in particular represents a much more recent segment of AP's archives and is perhaps less valuable to interrogating the characteristics of long-lived databases than AP's text and still images. Nevertheless, the long-term sustainability of these newer formats will undoubtedly present AP with new challenges to existing models.

**It is important to realize that news organizations maintain internal archives not as recorded history but as a quick reference tool for their journalists.** The value of the archives to these stakeholders drops off surprisingly quickly after a news event: News is “old” within 90 days, and queries against older material are correspondingly infrequent. Yet the back files are never quite “worthless”—until, inevitably, a triage situation tests their value against some more urgent contingency. This is the historic backdrop against which AP's present archives must be understood.

The electronic delivery of news, as this study will also show, is inextricable from AP's history. In journalism, AP has *written* the history of electronic delivery, and its influence has, in some instances, extended beyond the news business—in the area of digital photography, for example. Yet until recently this innovation has not translated into a concern for AP's own historical legacy. In fact, what was true in AP's 1894 annual report is true today: “The management has been so pressingly occupied with the making of Associated Press history as to leave but little time for recording it.”<sup>6</sup>

### **A Note on Sources**

Associated Press managers have been enthusiastic and cooperative about participating in this study. For one thing, no one within AP has ever quantified the range and depth of the text archives and they're interested in finding out what they actually have. What documentation exists is fragmentary, incomplete and in some cases conflicts with other evidence, so much of this process has been an exercise in ferreting out the facts.

As a journalistic enterprise AP takes very seriously its long-term role in providing primary source material to future historians. Even as other, more pressing projects occupy AP's focus (and capital), I heard from virtually everyone I talked to that they are now deeply interested in how they will retain their archives for the future. What these stakeholders mean by “archives,” though, is not always clear.

---

<sup>6</sup> My thanks to Corporate Archivist Valerie S. Komor for drawing my attention to this quote.

The information in this report is based on a series of interviews with AP's senior management and other stakeholders as well as on primary source documents from the AP Corporate Archives and editorial guides. In addition, secondary sources such as books and magazine or journal articles were consulted. I also was fortunate to find former AP employees who offered valuable information about the evolution of the text archives over the last four decades, some of which current AP staff were unaware of. Finally, Corporate Archivist Valerie S. Komor has been instrumental in helping me validate various findings as well as contributing many of her own.

As will be explained below in the section, "Funding Model and Business Activity," The Associated Press is a not-for-profit cooperative that holds annual meetings for members and publishes an annual report. These reports, in one form or another, date back to AP's formal incorporation in 1892 and provide information about technological innovation over a span of more than a hundred years.

In addition, I was provided with all of the speeches made by AP's president and chief executive officer, Tom Curley, whose arrival in 2003 signaled the beginning of AP's ongoing project to unify its disparate media types under a single search and delivery umbrella called Electronic AP, or eAP. These speeches provide an overview of the cooperative's recent strategic planning.

For reasons of confidentiality, certain internal strategic planning materials were not made available to me, but I am confident that I had access to everything crucial to documenting the archives past, present and future.

Finally, while covering news, AP is continuing to make news with eAP, copyright protection activities, fee structures and ongoing initiatives. This report, therefore, describes something of a rapidly moving target.

## **1. Overview of the Associated Press**

Since 1846, the Associated Press has been providing its members a constant stream of electrons over "the wire." Starting with the telegraph and evolving through the telephone system, satellite transmission and now the Internet, AP has exploited each successive tide of technological change to provide its members breaking news as quickly as physically and humanly possible. Indeed, this driving force—serving its members with speed and accuracy—informs every innovation out of AP from 1846 up to the present. In the sense of the current study, these members and subscribers remain AP's most influential stakeholders.

However, as that core newspaper membership struggles with a wholesale reshaping of traditional revenue models, the Associated Press, too, has had to branch out. CEO Tom Curley, who joined AP after his highly regarded tenure as founding publisher of *USA Today*, warned staff in June, 2004, that, “Ninety percent of our funding comes from organizations that are losing the audience that pays the rent.” The focus for AP, he went on to say, would be managing “a shift of revenue from comfortable, long-term partners to new players serving news on emerging channels in markets just forming.”<sup>7</sup> That reality informs all of the recent initiatives that will be discussed in this study.

While the numbers change constantly in response to global news events, some close approximations will suit. **AP has bureaus in about 100 countries producing text, still, and moving images, still and moving graphics, audio and video feeds to members and subscribers all over the world. Primarily an American organization, the cooperative is owned by some 1,500 daily newspapers.<sup>8</sup> Its total reach in the U.S. extends to 1,700 newspapers and several thousand television and radio stations, plus hundreds of nonmember newspaper, radio, television and internet distributors internationally.** Its news feeds are historically in English, but Spanish, French, Dutch, and German feeds were added in some regions in 1998 and 1999.

### *History and Mission*

The Associated Press was created to solve a nineteenth-century technology infrastructure problem—the limited and expensive availability of the telegraph during the Mexican-American War. Moses Yale Beach, publisher of the *New York Sun*, agreed to distribute incoming telegraphic news from the war to four other newspapers in New York City. Typical of the monopolistic practices of American business of the day, AP formed a tight, exclusionary relationship with Western Union that permitted it to dominate the limited telegraph capacity of the mid-century Northeast.

In short order, AP broke the lock held by the dominant rapid-distribution technology of the day—ships, carrier pigeons and fast horses—and created a virtual monopoly over telegraphic news. As it matured, it developed the form of narrative known as wire-service journalism, which has shaped American news writing ever since. In his history of wire service news and its influence, scholar Menahem Blondheim credits AP with creating at least the notion of a unified national narrative.

[AP’s] structure as a national institution—impersonal, non-local, unselfconscious, and hidden—gave wire service news, however partisan, the appearance of objectivity. The Associated Press

---

<sup>7</sup> Tom Curley, remarks to senior AP management, June 14, 2004.

<sup>8</sup> Hoover's Company Records. *The Associated Press*: Hoover's Inc., 2008, and the AP Web site, <http://www.ap.org>.

helped Americans accommodate to a common information environment. By giving news that impressed the minds of Americans a national orientation, it fostered the integration of American society.<sup>9</sup>

Daily journalism is competitive; wire-service journalism even more so. Over the decades American newspaper readers enjoyed the fruits of the cutthroat competition among AP, United Press, and the International News Service in the United States, and the Reuters agency overseas. Various collusions and cartels wreaked a certain amount of economic havoc among the wire services; the original UP went bankrupt in 1897, only to emerge ten years later as a group of regional services created by publisher E. W. Scripps with the primary intention of busting the AP monopoly.<sup>10</sup> But there was plenty of newsprint space to go around: During World War I, the domestic wire services were supplying nearly 2,500 newspapers in the United States.

From regional, national and overseas reporting, AP branched into financial news and sports in the 1920s. In the 1940s and 1950s, respectively, AP added radio and television reports—text written for on-air news readers. Television broadcasts were included beginning in the early 1990s with a merger that created AP Television News.<sup>11</sup> Efforts since 1995 have been focused on maximizing delivery from Web platforms and expanding AP's subscriber base through more niche content and vertical content packages, which arm AP subscribers with multiple options for displaying a particular news story—text, photography, graphics, audio, video, etc.

For 80 years, the mission and bylaws of The Associated Press were printed annually in full in the cooperative's report to members. Slightly revised and produced in booklet form in 2006 to mark AP's 160th year, the core objective is stated in language that has barely changed in the intervening century:

The Associated Press is a mutual and cooperative association formed to gather with economy and efficiency an accurate and impartial report of the news... [The AP] shall be as objective and complete as human endeavor can make it.<sup>12</sup>

---

<sup>9</sup> Blondheim, Menahem. *News over the Wires*. Cambridge, Mass.: Harvard University Press, 1994, 195.

<sup>10</sup> Reporters of the Associated Press. *Breaking News: How the Associated Press Has Covered War, Peace and Everything Else*. New York: Princeton Architectural Press, 2007, p. 409

<sup>11</sup> [APTV, a global video newsgathering agency, was launched in 1994; four years later, APTV merged with WorldWide Television News, forming APTN.]

<sup>12</sup> Associated Press. "Charter and Bylaws, 1846-2006." New York: Associated Press, 2006.

## 2. Funding Model and Business Activities

As noted above, The Associated Press is a “mutual and cooperative association” comprising regular members drawn from daily newspapers exclusively, and associate members from non-daily newspapers and broadcasting. It draws revenue primarily from annual assessments placed on its member newspapers, associate member radio and television stations, and subscriptions by thousands of newspapers, radio and television broadcasters in 121 countries worldwide.

The wire service was incorporated in 1900 following a protracted battle with local press barons in Illinois over its purported monopolistic practices. After the Illinois Supreme Court ruled that AP was a public utility guilty of restraint of trade under state law, AP decamped to New York City, where state law was more favorable to nonprofit cooperatives.<sup>13</sup>

The language of its bylaws in 1900 stated that the corporation is “not to make profit or make or declare dividends,” language that prevailed for a century until it was updated to reflect its status as an entity under section 102 (a)(5) of the New York Not-for-Profit Corporation Law. **In the course of updating the bylaws in 2004, AP broadened the definition of “information and intelligence” it gathers and disseminates to include “any and all kinds of news, information and intelligence; literary property of all kinds including that which is informative, educational or otherwise of public interest; news pictures, pictorial news and art of any and all kinds.”**<sup>14</sup>

The updated incorporation language also provided a bit of detail about how a surplus of income is to be used, recognizing that AP might, in fact, turn a profit.

The Corporation may charge fees or prices for its services or products and shall have the right to receive such income and, in so doing, may make an incidental profit. All such incidental profits shall be applied to the maintenance, expansion or operation of the lawful activities of the Corporation, and in no case shall be divided or distributed in any manner whatsoever among members, directors, or officers of the corporation.<sup>15</sup>

And, in fact, after several years of lackluster income or outright losses, AP finally turned profitable as a result of a round of major expense-cutting under CEO Tom Curley. Its revenue in 2007 totaled \$710.3

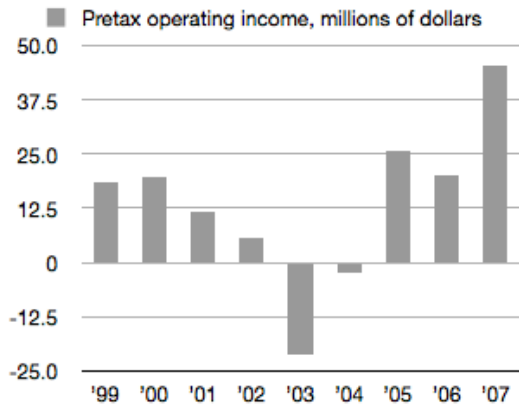
---

<sup>13</sup> *Breaking News*, 408.

<sup>14</sup> *Bylaws*, 3.

<sup>15</sup> *Bylaws*, 3, “Amended and Restated Certificate of Incorporation of the Associated Press.”

million on pretax operating income of \$45.8 million, more than double 2006 pretax income of \$20.4 million and a clear turnaround from losses in the 1990s and middle of the decade.



The improving financial picture, however, stands in stark contrast to its member newspapers, which are facing an alarming decline in print advertising and having to take round after round of painful cost-cutting measures, including editorial layoffs. Given that disparity, some members have questioned the fairness of AP’s current fees and what they’re fueling—technology improvements that would seem to benefit AP more than its members, and an expansion in foreign bureau operations at a time when newspapers are shuttering their own.

### *Fee System*

**For most of its history, The Associated Press has been funded by assessments determined by the Board of Directors and levied individually on each member newspaper, based on a number of variables including a member’s circulation, geographic region and discrete services provided.**<sup>16</sup> Increases must be approved by two-thirds of the directors, and, once established, members are not supposed to contest or even question how fees are apportioned.<sup>17</sup>

That prohibition hasn’t stopped fees from being a source of controversy. In April, 2005, the AP Board imposed a new fee on the “repurposing” of wire stories and images on members’ Web sites. Until then, newspapers and broadcasters had been free to use and reuse their AP feeds as they saw fit within their own publishing domain; AP’s new structure was an “online licensing fee” on top of what members were already paying. At the time, Curley, the CEO, said that it was necessary to “preserve the value and enforce the rights of our intellectual property across the media spectrum.”<sup>18</sup>

In response to member protests, AP rescinded the increase within a few weeks. One concern was that revenue-strapped members would stop posting AP content to their sites in order to save money. In

<sup>16</sup> As a rough guide, six Ohio members of AP reported earlier this year that their fees together totaled about \$4 million. See E&P.

<sup>17</sup> Bylaws, 25, “Apportionment of Expenses.”

<sup>18</sup> Liedtke, Michael. “The Associated Press to Impose Online Licensing Fees.” Associated Press. BC Cycle ed. New York, April 18, 2005.

announcing the return to the status quo, AP announced that the fee increase for 2006 would be 2.2%—one of the smallest in 30 years.

The AP Board returned with a new fee structure in the spring of 2007, aimed at giving members more flexibility to choose content they were apt to publish, which it announced the following October. The new structure is seen by AP management as pivotal to its strategy of enabling members to meet the demands of cross-media publishing. Instead of a constant stream of news pouring into member databases with only limited differentiation—which members were finding increasingly confusing and hard to sort through—the flexible fee structure permits niche delivery of specific packages, tailored feeds, vertical media (text for Web and print, with images, audio, video, and graphics all related to a specific news event, tied together with specific metatags) and even *a la carte* purchases of single items. Members pay fees based on what they use. At the AP annual meeting in April, 2008, CEO Tom Curley told members that their aggregate savings would total more than \$20 million.<sup>19</sup>

This, too, was not well received by some members, who argued that paying separately to use the same or similar AP content in print and on the Web was going to present an undue financial burden at a time when many newspapers are struggling financially. But AP held firm, insisting that this approach will instead reduce fees paid by members, and allow them to better manage the expense of content that is actually published.

#### *Adding Value for Publishers*

As another inducement for members to accept the new fee structure, AP offered a reduction to individual members who agree to submit their content for tagging and even storage by AP in its new metadata and term indexing structures, part of a major text enhancement initiative (described below).<sup>20</sup>

While there has not yet been a rush by members to take advantage of this service, AP hopes to enable members eventually to manage their own Web sites more efficiently by using this tagged content in conjunction with AP-enabled browser software, cutting down on the local labor involved in maintaining the site and enabling members to swap content among themselves. It also allows AP, with the members' agreement, to aggregate members' better-formed local content in AP's own feeds.

---

<sup>19</sup> "Editors at Odds with AP," *Editor & Publisher*, Vol. 141, No. 7.

<sup>20</sup> Personal interview with Lorraine Cichowski, March 4, 2008.

This is a powerful example of the influence of AP as a standard setter. The concepts and structure behind routing codes, story names and geographic identifiers that have added value to AP content for a century is the basis for the much richer tagging that AP now employs. **As more and more members submit their content to AP's tag factory, the prospect of a large body of fairly uniformly structured news text in widely distributed locations is no longer unrealistic.**

As a cooperative, the members of the Associated Press have certain obligations as well, the chief among which is the sharing of news.

Newspapers are obligated to provide the Associated Press with "all news that is spontaneous in its origin, but shall not include news that is not spontaneous in its origin, or which has originated through the deliberate and individual enterprise on the part of such member."<sup>21</sup>

In other words, AP has the right to expect members to provide it copies of their breaking, or "spot," news coverage for distribution to other members over the wire. A member is not, however, obliged to provide special, exclusive or "enterprise" reporting to the cooperative and its members, allowing members to preserve a competitive edge; a member can be the first to report on a case of government malfeasance uncovered by its reporters, but it would hardly be reasonable to keep a local earthquake under wraps.

However, as the newspaper membership base struggles with falling income, AP has been exploring other sources of revenue, a trend that began in the mid-1990s and the dawn of the Internet. **Newspapers now account for about 30 percent of AP's revenue.** The cooperative is less purely a cooperative and more of a direct competitor in some arenas, not just journalistically but as a provider of content to other competitors.<sup>22</sup>

Among AP's nonmember, or "commercial," customers are Web sites, wireless companies, database aggregators, companies and government, all of which pay fees that average 50 percent higher than those paid by member newspapers for the same material.<sup>23</sup> The recently enacted package pricing model permits members and subscribers alike to license specific products from AP.

Needless to say, this evolving model has surfaced new worries about competition and exclusivity of news. Addressing these concerns among members is important to maintaining mutual trust within the cooperative; indeed, the oldest corporate documents from the early twentieth century, the annual report to

---

<sup>21</sup> *Bylaws*, 23-24. Other sections of Article VII regulate sharing of news between members, limitations within a paper's circulation area, proper crediting of bylines, embargoed material, and so on.

<sup>22</sup> "Q&A—Jane Seagrave." *AP World*, Spring 2006, 10.

<sup>23</sup> *Ibid.* The interesting commercial and nontraditional uses of AP content range from risk assessment and corporate intelligence research to news screens on transit buses and entertainment video in retail stores.

members, record instance after instance of purported violations by members against each other and the cooperative. The ubiquity of AP's content in new media channels and sales of content to big portals like Google, Yahoo and CNN has challenged the definition of the cooperative, and AP has been careful to limit what it makes available to nonmember subscribers, even if it means missed revenue. One solution that allays member concern is limiting subscribers to national and international feeds. Only members are able to access and post state wires to their own websites, and they have the assurance that the local news they in turn submit to AP remains within the cooperative. For its potential to create links for local advertising, local news is considered the crown jewel of the wires.

**Of even greater concern to AP members and subscribers alike is the unauthorized aggregation and use of AP content,** which reduces the value of AP's material by wiping out exclusivity. Web technology has made news aggregators out of individuals and companies; amateurs with blogs alongside corporate "knowledge" vendors have readily seized upon syndication engines to harvest AP content and send it on to readers. Ironically, the same mechanisms that are giving AP its expanded reach are also seen as threats to its member and subscriber relationships. A subscriber won't pay for content that is duplicated by unauthorized sources.

The Associated Press is combating this on a number of fronts. First, increasingly sensitive digital "sniffers" roam the Web looking for sites that have picked up AP content and determining how it is used. Photos are fingerprinted and watermarked for ready detection. Even the adept cut-and-paste activities of individual bloggers are no longer invisible to discovery. When a violation occurs, AP lawyers seek remedial action, from a simple request to cease and desist to taking legal action; in October, 2007, AP took its first legal action in ninety years against a copyright violator,<sup>24</sup> and more recently threatened infringement action against a blogger for including short phrases of AP content on his website.

Along with digital tracking and legal strikes, AP is actively working to streamline and automate copyright registration processes, a concern for proper protection of tens of thousands of stories and images emanating from AP each week.<sup>25</sup> As a driver of standards, as well, AP is working with others in the industry on a scheme called Automated Content Access Protocol (ACAP), including hosting a meeting of the project group at AP headquarters last November.

---

<sup>24</sup> Srinandan Kasi, vice president and corporate counsel, personal interview, March 3, 2008. In October, 2007, AP sued VeriSign and subsidiary Moreover Technologies for misappropriation of AP content. An industry group including AP negotiated a settlement with Knowledge Networks, a market research firm.

<sup>25</sup> Singhania, Lisa. "Intellectual Property: A.P. Looks for New Ways to Protect and Maximize Content." *AP World* Fall 2007: 6-7.

AP's CEO Tom Curley acknowledged the opportunities amid the threats in copyright with a remark to members in 2004:

We are confronting the serious threats to the value of AP news that arise continuously in the wired world from information pirates and copyright infringers who seek to beat us to our own customers with our news. Some of these could become customers or partners. Others may become defendants. None will be ignored.<sup>26</sup>

### *New AP Markets*

While sophisticated tagging algorithms and intellectual property protection efforts certainly strengthen AP's revenue model, AP is also working constantly to identify either new or expanded content opportunities. A look at AP's archives of press releases for the last five years shows a steady stream of initiatives designed to benefit members and attract more paying subscribers, both in terms of wider news coverage and opening up new revenue options. The market for news in the past two decades has tilted toward personal lifestyle coverage—health, personal finance, entertainment—while seeing huge growth in business/financial and sports coverage. These topics, of course, often rate the flashiest display on newspaper websites.

Table 1 shows the range of initiatives announced by AP since 2003, the year that also saw the beginning of the eAP project. It lists partners, if any, and describes the intended benefit to AP and its customers. The announcements reveal a pattern of both exploiting existing content areas and expansion into new ones, playing to AP's acknowledged strengths in those areas. The majority are not outright investments but partnerships or service expansions of the kind AP has done throughout its history.

*Table 1. Five years of AP initiatives, culled from AP's online archives of press releases.*<sup>27</sup>

<b>Year</b>	<b>Principals</b>	<b>Description</b>	<b>Domain</b>	<b>Value</b>
2003	McClatchy, Newspaper Network, Vertis	Sale of McClatchy's Newspaper Network to AP and Vertis	Planning, pricing and placing of advertising in newspapers	\$12 mil
2003	Ipsos-Public Affairs	Partnership on public opinion polling	New service. Polling data available to members and subscribers as regular feature	Service
2003	AP	Enhanced election coverage for members and subscribers	Customizable or packaged pre-election content and election-night results	Service
2004	AP-VII	Distribution of content produced by French news photo cooperative	New service.	Service
2004	AP Financial News	Expanded coverage to include more markets,	Expanded service aimed at diverse markets beyond print and broadcast journalism, including market	Service

<sup>26</sup> Curley, remarks at annual meeting, April 21, 2004, Washington, D.C.

<sup>27</sup> AP's archives of press releases can be accessed at <http://www.ap.org/pages/about/pressreleases/preleaseindex.html>

		earnings, executive changes, regulatory actions, mergers new product development	intelligence, knowledge aggregators, etc.	
2005	AP, Microsoft	Technical platform for OVN network, revenue sharing among MS, AP and members	Establishes paid advertising model for AP- and member-produced video using MS video player and advertising infrastructure	Partnership
2005	AP, NewsCorp	Formation of Stats LLC, combines AP's multimedia sports news coverage with NewsCorp's Stat Inc., sports stats analysis firm	Expanded sports and statistics coverage for existing and new markets including agents, teams, fantasy sports, game developers	Partnership
2006	AP	AP planner	Database of future events, research and planning tool of 35,000 events in 100 plus categories. Service to members and paid subscribers	Service
2007	AP, Bolloré Group	Creates French news agency	News coverage plus sales rep for AP's English language media in France. Builds on existing language translation service	Partnership
2007	AP	AP Money and Markets Extra	Expanded personal investment news	Service
2007	AP, Johnson Publishing	Distribute content from <i>Ebony</i> and <i>Jet</i> magazines	New service, including magazines' historic images	Service
2007	AP, Google	Google begins hosting AP and Agence France-Presse content instead of displaying snippets and links	Resolves infringement dispute over unauthorized use of AP content harvested by Google from licensed sources	Protection
2008	AP, smart phone developers	Mobile News Network	Multimedia delivery of news and advertising from AP and 100+ member publishers. Optimized for iPhone platform, users can localize content	Development
2008	AP	Ask AP	Journalists answer queries from the public for a column that can be picked up by members	Service

### *New Value for the "Back Story"?*

A particularly intriguing initiative at AP is an effort to understand how younger consumers—those who have not grown up reading newspapers—use news in the Digital Age. Toward that goal, AP commissioned a group of research anthropologists to study news consumption patterns among 18- to 34-year-olds in the United States, the U.K. and India, and reported the findings at an international forum of editors in Sweden in early June 2008. What they found was “news fatigue”: the overload of facts and updates that prevented them from “connecting with more in-depth stories.” The atomized nature of headline news delivered to various receiving devices led one subject in the study to observe, “News [today] is not like the full story, but more like a preview—it’s kind of annoying sometimes. I don’t like to get bits and pieces of information.”<sup>28</sup>

One key finding is that popular genres like sports and film reviews are freestanding, i.e., they have a beginning, middle and end, and don’t require an ongoing frame of reference for understanding. This is not the case with general news and especially with politics. The study suggests that building in access to a

<sup>28</sup> Associated Press, and Context-Based Research Group. "A New Model for News". New York, June 2008. PDF. <http://www.ap.org/newmodel.pdf>.

“back story,” the background and/or recent developments in an ongoing news story, “could have an impact on the audience’s engagement level.”

In other words, offering news consumers easy linkage to earlier stories could provide them with the context they seek—and, more to the point, keep them occupied on the Web site where they can also encounter revenue-generating advertising. If, for example, a breaking news story refers back to a speech made by a political candidate in 2007, the user might access the actual coverage from the earlier event by launching a query against the text archives—in reality, by clicking on an embedded hyperlink. This is dependent on the “back stories” being uniformly tagged and structured in the current scheme; if the “back story” predates the new eAP system, there will be a diminishing scale of availability and potential links the further one goes into the oldest material in the digital text archives.

#### *A Wider Window on Content*

Even as the AP study was being released to the public in June, AP technologists finally answered a crucial question that had been unresolved since the launch of eAP in 2003. **The contents of the text archives, back to 1986, would be fully migrated to the new system.**

This suggests that the “back story” will be available to digital news consumers eventually, but how that happens won’t be known for some time after the material is migrated. In the meantime, the question of what “back” means remains: Does the potential value of the “back story” extend to the more limited pre-1996 material? What about historic material only in paper form or on film? Does the interest in the back story justify the expense of digitization? If it did, will “old form” news narrative from the twentieth century have any relevance for new users accustomed to aggregating their own information in RSS feeds and mashups?

AP notes as well that young consumers are often oblivious to the source of information and routinely surf away from an initial link. For the back story, they’ll want the best source of information, regardless of who has it, and, the study indicates, they think getting there should be seamless:

Where online consumers once surfed and bookmarked news sites, users now wonder why a logical trail through the news can’t simply unfold, link by link, across a multitude of sources.<sup>29</sup>

---

<sup>29</sup> “New Model for News,” 64

This is a very tall order. AP’s program inviting members into its tagging infrastructure at least sets the stage for a cross-domain news environment in the United States among member sites, but as the report notes, it will take “significant human cooperation on a very large scale” to link news globally.

### 3. Databases and Systems at AP

**Assuming that they are engaging in archiving activities (and all but the smallest do), news organizations, whether print, broadcast, online, or all three, all adhere to a set of basic processes that result in a flow of “privileged” content to a “permanent” archives.** These archives may be physical or digital, but as newspapers have long since stopped clipping articles and saving bound volumes, the archives usually comprises one or more databases. The steps in this process are:

1. Reporting, writing (or shooting) and editing the news. This is an iterative process that at a very early stage takes place in a production system. Depending on the medium, these systems allow for “data entry” by the journalist, shaping and refining by an editor, and placement into a publishable form, be it a page layout or video format.
2. Publication. This is a triggering process that sets off a series of other processes, but in concept it is the decision that an article or image is ready to appear to its audience. In a production system, it may consist of a set of keystrokes or clicking on a button labeled “send,” “file,” “compose,” “upload,” “publish,” etc.
3. Copy to the archives. Where the volume of articles and images is high, as at a daily newspaper, the production system is often programmed to automatically send a story or image to its archival database. The “archival object” at this stage is ingested along with a fair amount of publication metadata, such as date/time, section or topic, author or producer, page number, and so on—with greater granularity in more sophisticated systems. Relationships between objects—an article and a photograph, an article and a page PDF—may also be established automatically or manually at ingest.
4. Quality control. Typically, the production and archives databases are two different systems, often built by two different vendors, and to ensure data integrity, human beings (sometimes with machine assistance) check the archived files to ensure the metadata is complete and correct and that the fields are populated properly between systems. How thorough this is depends on local economics.<sup>30</sup>

---

<sup>30</sup> Amid ongoing downsizings, a growing number of newspapers are foregoing quality control altogether and sending data feeds to aggregators directly from their production systems. There is abundant anecdotal evidence that commercial databases that sell

The processes at the Associated Press are essentially the same. In the early days of telegraph and teletype, the initial system was a couple of journalists with typewriters and copy pencils, and the publication system was the Morse Code operator or teletype man. The edited paper copy went onto a metal spike or hook and most of the time was lost to history. As will be described in detail in the Text Archives analysis (page 26, below), this system evolved into today’s digital production, publishing and archiving systems, but the processes are the same. AP enjoys a crucial advantage, though. The developers of its in-house text production systems were also the developers of the modern text archives, and the creation of a common search and retrieval interface assured high-quality export from live production to the archives.

*AP Data Operations*

There is no description of AP’s information architecture available on its public website or any print source that I was able to find. This description of the general and text archives systems is based on interviews and email but does not present a complete picture.

The text archives and eAP architecture, which are currently two separate systems, sit in a major facility in Kansas City with replication in Cranbury, NJ. (AP also has data centers in New York City, London, Frankfurt and Bangkok.) At any given time, eAP serves about 100 terabytes of data. The number of objects and how they are measured varies by format, but AP staff members were able to provide these approximations:

AP unit	Medium	Ingest rate or scope of holdings
APTN and OVN	Television and online video	About 3,000 hours per year (daily 4.5 hours of TV and 2 hours of online)
AP Photo	Photography (JPEGs)	1 million images a year (2,500-3,000 per day)
Text archives	Database back to 1985	Currently hold 65 million documents (daily ingest rate in the thousands, depending on news)

The text archives systems have been migrated several times.<sup>31</sup> **Although the first electronic archive system was VuText, described below, the current technology actually traces back to the news production system, a widely used writing and editing product called Atex.** The largest bureaus—the hubs that collected filed stories from smaller bureaus—had large databases known as the Mouse (plural, Mouses) that powered the editors’ “dumb” terminals. (Smaller bureaus made do with floppy disks, most

---

news packages are increasingly prone to errors such as missing headline fields, malformed or absent metadata and truncated articles.

<sup>31</sup> This information is based on interviews with Stan Miller, who has been in news technology at AP preceding and throughout the development of the Reporter’s Workbench (March 6, 2008) and Bruce Toll, another developer since departed from AP (June 3, and 11, 2008).

of which have been lost.) In 1987, the Mouses were replaced by VAX architecture that ran Atex's replacement, AP Edit, a DOS-based system.

The end of DOS and arrival of Windows (1995-96) caused problems with AP Edit and sparked the development of a new production system, which became Reporter's Workbench a year later.

At the same time, AP's news technology group decided to make the text archives searchable using the same search interface as the production system—in principle, a federated search engine over the production and archives silos. Historically, the size of AP's text database and its demands for real-time access to filed stories—"micro response time," as one technologist phrased it—made installing off-the-shelf software out of the question. In developing the Workbench, a commercial database package and search engine were highly customized into a single command-line interface, and the system continued to run on the VAXes. Once the Workbench was functional, VuText was abandoned. The VAXes were replaced by the current DEC Alpha-based architecture.<sup>32</sup>

**Details about the eAP environment (referred to by staffers as the "Very Large Database" or VLDB) are not available, except to say that it is a Windows/Microsoft .Net environment on an SQL server database.** One of the considerations driving the recent decision to migrate the electronic text archives to eAP has been the increasing difficulty in maintaining the DEC Alpha system, which in any event is expected to be phased out in the next several years by its current owner, Hewlett Packard. As Mike Alexander, one of the leaders of the migration team explained, the software code base is "very old and difficult to change," not well documented, and all of the original developers have left the organization. Moving it into the VLDB will improve support and make the data more accessible to a broader base of users, including new customers.<sup>33</sup>

#### *Associated Press' Technology Development: A Timeline*

The manner in which news is structured and disseminated is as much a part of the historic record as the journalism itself. In structure and technology, AP's influence clearly equals its journalistic impact in the way it confronted the challenges and opportunities of technology, even primitive technology.

Although the written histories of the cooperative tend toward the élan and adventure of the wire service correspondent, the enduring history lies in its response to the technology of the day—quite a remarkable

---

<sup>32</sup> AP technologists still refer to the system underneath Reporter's Workbench as the VAX, which makes tracing the hardware history somewhat confusing.

<sup>33</sup> Mike Alexander, personal communication, June 9, 2008.

record over 162 years. **Its corporate mandate to “gather with economy and efficiency” is manifest in an impressive level of research and development, while its virtual lock on subscribers’ sending and receiving equipment made AP a powerful standard-setting body within American news publishing.**

A selected timeline of AP’s technology shows why this is so.<sup>34</sup>

1844—Point-to-point transmission of news by telegraph.

1846--Moses Yale Beach institutes cooperative newsgathering and the Associated Press is born.

1897—AP contracts with Burrelle’s Clipping Service (founded 1888) to provide it with AP content from New York newspapers.

1904—News library is formed and begins indexing clippings and later cable copy with subject and geographic indexing based approximately on the Dewey Decimal System.

1914—Morse Code transmission (which has to be translated) is replaced by teleprinters, which produce eye-readable copy. Members install teleprinters in their newsrooms.

1921—AP tabulates, codes and transmits stock tables.

1935—Using telephone circuits and modulators that convert light waves to sound, AP transmits the first photograph (a small plane crash) by wire. Members install Wirephoto receivers in their newsrooms.

1951—Teleprinters are replaced by teletypesetters, which generate punched tape alongside wire copy. The tape is used to drive members’ linotype machines and eventually phototypesetters in the 1970s.

1960—First computers: AP introduces IBM mainframes to automate stock listings.

1965—AP, along with other news organizations in North America and Europe, form the International Press Telecommunications Council (IPTC) to establish standards for the transmission of text (and, later, other formats) over the wire.<sup>35</sup>

1972—First CRTs (cathode ray tube) desktop computers replace typewriters for AP journalists, providing basic text entry and editing capability and allowing electronic filing of stories.

1976—News transmission goes even higher speed with introduction of DataStream, a 1,200-baud phone-line transmission rate. Member newsrooms install updated printers and receivers.

1976—Wirephoto becomes Laserphoto with the first laser-scanned and amplified photo transmission. AP had partnered with scientists at MIT to develop the technology. Members install the newly developed laser receivers.

---

<sup>34</sup> I have compiled this timeline from AP promotional literature, annual reports, issues of *AP World*, and my own experience in newsroom technology in the 1980s and 1990s. In the interest of casting it toward a thorough understanding of the environment that created and sustains the text archives, many developments in the broadcast domains have been omitted.

<sup>35</sup> The IPTC was established in 1965 by a group of news organizations including the Alliance Européenne des Agences de Presse, ANPA (now NAA), FIEJ (now WAN) and the North American News Agencies (a joint committee of Associated Press, Canadian Press and United Press International) to safeguard the telecommunications interests of international publishing media. Since the late 1970s, IPTC’s activities have primarily focussed on developing and disseminating industry standards for the interchange of news data. See <http://www.iptc.org>.

1979—Photo editing goes electronic with introduction of AP’s Electronic Darkroom, which allows cropping of pictures and editing of captions. Participating members install specialized editing units.

1982—AP leases satellite bandwidth to transmit Laserphotos without wires. Members install satellite dishes and updated Laserphoto receivers. Two years later, AP is the first news organization to own a satellite transponder.

1985—As personal computers are starting to enter households, Knight Ridder (a major newspaper chain) launches an experiment in online news delivery to the public. AP joins the effort, installing a VuText system and contributing content. While the online aggregation business never takes off, VuText forms the genesis of the electronic text archives ten years later.

1986—Charts and maps go digital with the introduction of the Macintosh computer, and AP responds with GraphicsNet, satellite transmission of information graphics. Members install new graphics receivers to hook up to dedicated Macintoshes. AP’s choice of software for creating graphics sets the standard for the newspaper industry for almost the next 20 years.

1987—Photo transmission goes from analog to digital with PhotoStream service.

1988—Satellite capacity doubles and DataStream service transmits at 9,600 baud.

1990—AP introduces the prototype digital asset management system, AP Leaf Desk, for receiving and managing incoming digitized photography. Members install Leaf systems in their newsrooms. This development also gives impetus to Adobe’s fledgling Photoshop program, which becomes the *de facto* standard for image software, and Adobe cooperates by embedding the IPTC standard for photo transmission into its header.<sup>36</sup>

1994—AP introduces a filmless, digital camera designed for photojournalism in partnership with Kodak on a Nikon body. At between \$25,000 and \$40,000, AP News Camera 2000 features removable storage, enabling photographers to deliver the photographs to editors while continuing to shoot an event. Super Bowl XXX is shot entirely digitally. Better-heeled members invest in a camera or two. (Digital camera prices, of course, fell precipitously over the next several years.)

1994—Responding to a new challenge—interoperability—AP partners with Adobe and graphics software developers to integrate the new Portable Document Format (PDF) into its graphics products. This frees AP and its members from forced software upgrades and accommodates members that had opted to use a different software platform. It is a sign that the once closed environment of AP standards and AP equipment is opening up to third-party technology in AP’s and its members’ own newsrooms.

1996—Meeting member demand for Internet content, AP launches “the WIRE,” a 24-hour, continuously updated online multimedia news service. Although wire feeds until now had been nearly a 24/7 operation, they were still tied to the AM-PM publishing cycles of member newspapers. By linking to the WIRE from their own sites, AP members can offer readers up-to-the-minute breaking news.

---

<sup>36</sup> Another beneficiary of this development was the JPEG standard for photo compression. The TIFF format, prevalent now as a digital preservation standard, was never adopted in publishing because of the size of the files—a major cost factor in the early 1990s when storage was expensive; photo management systems typically store hundreds of thousands of images at any given time. Further, the commonly used page design, color prepress and production software did not accept TIFFs as a valid format. Thus digital news image collections today are almost 100 percent JPEGs.

1996—Wanting to give reporters and editors faster access to the archives, AP leverages its production system to create an in-house text database. It is initially populated with about four years of archives from VuText and eventually reaches back to 1985 for some wires.

1999—Video and audio is streamed to members.

1999—Your AP is launched as an Internet-based delivery service for AP content. It allows members to customize and control how they view incoming AP feeds.

2000—Recognizing that existing metadata is not up to the challenge of the latest newspaper production systems, AP undertakes a program to normalize formats and tagging for text and images.

2000—AP solicits member contributions to its AP/Worldwide Photo archives in a revenue-sharing arrangement.

2000—AP Digital is launched: umbrella organization for Internet and digital products aimed at both members and commercial subscribers.

2003—Work begins to create Electronic AP (eAP), unifying all media formats in a single delivery platform, a project involving both more robust technology infrastructure and enhanced intellectual controls on data—more and better metadata and controlled vocabularies.

2003— CEO Tom Curley joins AP in June; first corporate archivist is hired in July.

2005—AP launches concerted effort to protect its intellectual property, through the use of sniffer technology that identifies unlicensed use of AP content followed by aggressive pursuit of violators.

2006—Launch of AP Exchange, a browser-based window on AP content for members and subscribers that unifies advanced searching across all media—text, still and moving images, audio and graphics. Enhanced metadata allows for vertical packages in popular topics such as health, finance and technology. Members (and public users of members' websites) may also search “archival” material back some number of months, all on a customizable subscription basis.

2007—AP adjusts its fee structure to accommodate niche and “a la carte” pricing, a controversial move that members insist will raise their costs.

2008—Members are encouraged to submit their own content for tagging in AP's enhanced metadata suite, and as an inducement can expect a break on their fees to the cooperative. It offers the eventual prospect of a large body of uniformly structured and described, cross-media news reports.

2008—AP decides to migrate the complete electronic text archives into the eAP environment.

#### 4. Challenges: The Evolution of News Text

**One often overlooked area of AP’s influence on technology is the very structure of journalistic narrative.** Until the middle of the 1800s, newspaper articles typically were chronological narratives: If something really important or interesting had happened, the reader might not discover it until he was many column inches into the report. The economics of Morse Code over the wire, on the other hand, required a minimum of data and a maximum of efficiency—not unlike managing bandwidth today—so there quickly evolved a compressed written form known as “cable-ese.”

**The most important, salient facts of an event were blasted out in a shorthand style that lacked even the embellishment of a’s, an’s or the’s.** With breaking news, a story would be telegraphed in short chunks as events unfolded in real time, requiring an editor on the receiving end to reinsert the articles and assemble the pieces for typesetting.<sup>37</sup>

This urgency is recognizable in the well-known “who, what, where, when, why” of the modern newspaper lead (or “lede” in cablese, to disambiguate it from the different readings of “lead”), with less-important facts added on in descending order until the story peters out altogether. Not only did this unyielding structure, codified in 1903, allow for fast editing on the transmitting end, it had benefits on the receiving end as well. The physical limitations of hot-lead type demanded that a story would often have to be truncated according to convenience rather than sense, so if a story was written in this proper “inverted pyramid,” the reader was guaranteed of getting at least the most important news and some subsidiary facts.

Other conventions, such as datelines, wire designations and story codes, which developed over time to route stories across regions and publication cycles, are simply the nineteenth century equivalents of geocodes, metatags and unique identifiers, and were invented to solve similar problems. How they have informed AP’s current efforts with metadata will be discussed in the section on Text Archives.

---

<sup>37</sup> The February, 1937, edition of *Fortune* magazine carried an amusing example of AP cable-ese: EARLIES ITALY TRIBUTED MILLIONS ITALIAN WIVES WHO YEAR AGO LAID WEDDING RINGS ALTAR PATRIOTISM LED QUEEN ELENA MUSSOLINI ESS STOP. Translated: “Italy paid tribute today to millions of Italian wives who, led by Queen Elena and Signora Mussolini, a year ago laid their wedding rings on the altar of patriotism” in Italy’s war with Ethiopia. This data compression, according to *Fortune*, kept AP’s telegraph tolls at half what they would have been uncompressed.

**The workflow model of a reporter handing off his typescript to a telegraph or teletype operator for transmission prevailed for a hundred years.** Reporters and editors didn't start working on computer terminals until the mid-1970s. But as the timeline shows, innovation has been nearly constant ever since.

*AP News Content Today: Text and Images*

AP provides a variety of categorized text feeds, the traditional core of its news coverage. **Meeting the need for text delivery for multiple platforms and an increasingly distributed, diverse readership, AP is developing spin-off products that build on existing processes.** One new example is NewsNow, a short (130 words or less) report compatible with traditional print formats but also mobile phones, Web, television and radio broadcasts or podcasts. In a breaking news situation, these are churned out first, followed by longer-form treatment of the same event with appropriate categorization.

Photography is another critical product for AP dating back to 1927. It holds many millions of photos in paper form in its New York headquarters as well as in bureau archives around the world and has undertaken a major digitization effort to allow more access via eAP. Global demand for pop culture images has focused attention on entertainment and sports (as it has for print), but there is solid demand for historical images as well. Facing stiff competition from Getty Images and Corbis as well as traditional rivals like Reuters and Agence France-Presse, AP is ramping up the availability of historic material with an annual digitization goal of about 150,000 images. Its collection is believed to have between 10 million and 15 million legacy photos, about 850,000 of which have been digitized to date. AP's daily intake of new images, meanwhile, ranges between 2,000 and 3,500, or about a million images a year. **While there is no firm estimate of how many images are stored in eAP, the total is into the millions.**<sup>38</sup>

*Broadcast*

Launched in London in 1994, **Associated Press Television News** is the world's largest video news agency with about 80 bureaus. In addition to delivering breaking video news, sports, and entertainment material, subscribers have access to half a million stories in its database. Its primary business is selling content for reediting and rebroadcast by subscribers (a familiar local television news anchor in many cases may be reading from AP's broadcast wire and showing AP video footage). APTN goes to more than 500 broadcast newsrooms, portals, Web, broadband and mobile customers worldwide. A separate unit, ENPS, sells electronic production systems to broadcast newsrooms.

---

<sup>38</sup> Estimates are derived from a personal interview with Ian Cameron, vice president of AP Images, January, 2007; email between Valerie Komor and Santiago Lyon, June 2, 2008; and current information on the Associated Press and AP Images websites.

A separate and somewhat different new video service, the **Online Video Network (OVN)**, provides prepackaged online video content for AP members. The service offers foreign, national and regional news in a wide variety of subject areas (finance, government, science, technology) as well as breaking news and features, edited and formatted for Web use. The OVN content allows members to enhance their own sites, which is seen as complementary to AP- or member-generated coverage. AP is also urging members to submit their own video content to AP for distribution—a reflection of newsroom “convergence” that now has print- and photojournalists shooting video as a routine part of their assignments.

In the audio arena, **AP Radio** has provided news wires written for radio since the early 1940s. Since 1974, AP Radio Network has provided hourly newscasts, sportscasts and business programs to member radio stations and today has more than 250 members (which have “associate” status in the cooperative). Voice feeds and actualities supplement the broadcast wire. All News Radio was started in 1994, a packaged newscast offering subscriber stations an all-news, 24/7 format.<sup>39</sup>

#### *Content and Services in the Digital Age: eAP*

Just three years ago, CEO Tom Curley declared that the wire services was finally

... on the cusp of a historic shift from the telegraph model that has characterized our business for a century and a half to a new database model that will give [members] easier access to our news content, rationalize pricing for different uses of our content and open up new business opportunities for all of us.<sup>40</sup>

It seems surprising that, given AP’s track record in innovation and the new realities of Web-based news delivery, the wire service is only now poised to leave the era of telegraph. This attitude is centered on what is known around AP as the “fire hose,” an apt description of the high-pressure stream of data flooding hourly into members’ newsrooms. And in time of reduced staffing and budgets, it is increasingly hard for them not to be swamped.

#### *Content Enrichment Project*

**One of the problems that AP identified starting in the late 1990s and moved finally to correct was the lack of metadata.** The “fire hose” pushed packaged content in different forms, but members had few choices among packages and took in much more content than they could possibly use—another key consideration as newsprint is literally shrinking. The local packaging and mounting of AP content by a member is labor intensive, requiring a layer of personnel to hunt down content in different siloes—text

---

<sup>39</sup> This material is now being collected by the Corporate Archives.

<sup>40</sup> Tom Curley, speech to Michigan members of Associated Press, June 17, 2005.

feeds, photo, video and audio databases. By introducing rich tagging into its various types of content, AP hoped to offer members greater latitude in searching for, selecting, customizing and publishing AP material, with fees more closely reflecting what they actually used.

Curley described the basic approach:

We ... intend to tag all the important people, places and things in the text, so [members] can link additional resources to them—stock prices and charts for public companies, statistics for athletes and profiles of the rich and famous.<sup>41</sup>

This rich mix of stories, pictures, broadcasts and graphics, tagged to permit easy combination of related material, is also known as vertical marketing, and beneath it is a major, multiyear effort to expand on AP's robust, if limited, metadata standards, which are founded in wire transmission codes. AP hired a team of taxonomists in 2005, who continue to build out various authority lists under its proprietary metadata schema, APPL (Associated Press Publishing Language, rendered as an XML DTD).

As in the past, where AP goes, members follow. Curley articulated the power of AP to determine a standard for news delivery:

As a cooperative, we have the opportunity to set the standards for how all this gets done, just as we did for print formatting decades ago, and we look forward to working through it with [members].

And, in its latest (2007) annual report, AP goes further by offering members a reduction in their assessments if they participate in its Content Enrichment Initiative, which permits members to submit their own text to AP and have it returned to them enriched with the APPL tag set. Not only will this make it easier for them to manage their own content, AP promises, but “at full scale the program will produce a comprehensive index of AP and member content that can be used to guide search engines.” **In other words, over time AP will create a pool of uniformly structured and tagged news content—not merely text, but at full scale still and moving images and audio as well.**

As will be discussed in the section on preservation policy, there is no system in place at AP for deleting or expiring content. So even if AP's goal is immediate access and dissemination of breaking news in multiple media types and packages, the outcome of this effort is potentially an ever-growing *multimedia* archives of objects normalized around AP's new standards. There are complications and hurdles to be overcome, but this is unprecedented in other organizations—public, private or not-for-profit.

---

<sup>41</sup> Ibid

## 5. AP Archiving

Before proceeding with a discussion of the depth, breadth and history of AP's text archives, it is useful to understand the nature and structure of wire service news feeds, which have evolved very slowly, as Tom Curley pointed out in his remark about the end of the telegraph age.

### *How Wire News is Distributed*

Anticipating the “fire hose” concept by about 70 years, *Fortune* magazine in 1937 had invoked the metaphor of rushing water to describe how news is channeled to AP members. While the process became much more complex and sophisticated as delivery technologies evolved, the fundamental concept remains the same. The core wire system sat in Kansas City:

... where 35 editors and 65 traffic men, situated on a sort of transcontinental news divide, edit and valve the news washing in from either coast...<sup>42</sup>

*Forbes's* “flow,” “flood” and “tributary streams” of news transmission depend today, as they did in 1937, on a system of codes and category identifiers that determine what stories a member receives, how and when he can use them, an individual story's instantiation in a breaking news saga, and, since the dawn of the Web, how stories are displayed on the member's own Web pages. **The codes, taken together, provide a highly abbreviated but surprisingly deep layer of descriptive metadata**, which is described with an example in Figure 1, below.

Historically, a simple wire transmission was highly structured and rich in metadata. The coding that describes the wire category, its urgency and its geographic origins date from protocols in telegraph and were adopted by IPTC in 1965. Textual description provides additional information about a story's place in a chronological sequence as well as advice for editors about its evolution during the news cycle. Related material is also specified.

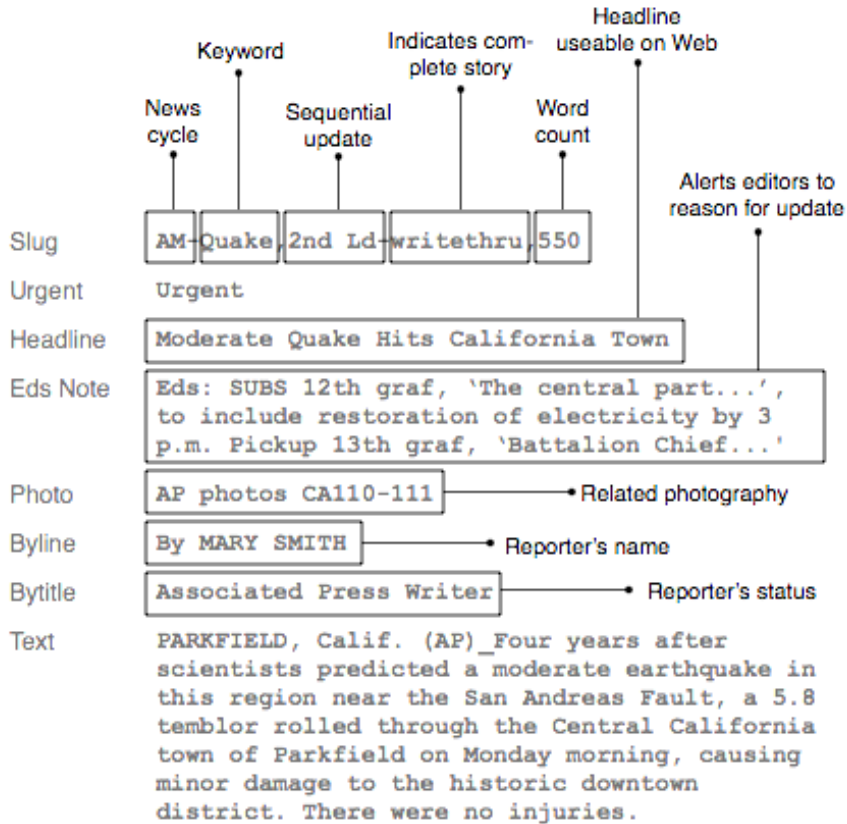
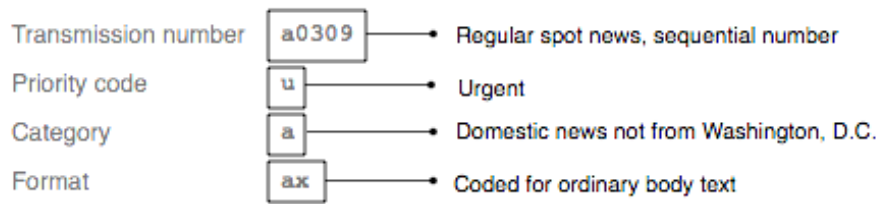
While these protocols are evolving in the Web era, they are native to the vast majority of AP's legacy content. Newspaper production systems, moreover, are programmed to interpret and act on AP/IPTC standard coding, both in print and in Web publication. For example, the current Associated Press *Stylebook* suggests that the headline be written so that it can be used automatically, without re-editing, on a member's Web site, taking advantage of automated processes embedded in the systems.<sup>43</sup>

---

<sup>42</sup> “(AP).” *Fortune*, February 1937, 148

<sup>43</sup> *Stylebook*, 403.

Figure 1. Inside wire text coding and structure



### *Transmission Codes and Categories*

Most news organizations use some kind of computerized production system (or desktop publishing software) to enable their journalists to write and edit their copy electronically and send it on for publication and archiving. **The larger the paper, the more sophisticated and automated the system. Regardless of manufacturer, these systems come enabled with software and scripts tailored specifically to manage wire service feeds; in other words, accommodation for AP code is embedded in publishing systems worldwide.**

The first layer of code describes the type of service available for transmission. The most common ones are:<sup>44</sup>

A	Spot DataStream news
B	Most advances
C	Weekly Features
F	Business and financial Datastream
P	Limited Datastream
S	Sports Datastream
T	Expanded sports report
U	Expanded business and financial report

Known in shorthand as the “A-wire,” “F-wire” and so on, these codes group wire stories in such a way as to allow members to purchase different packages, which give them more or less content as dictated by their publishing needs and “news hole” (column inches on a page not reserved for advertising).

The mass of data that each of these feeds provide can be filtered programmatically on the receiving end by two more categories of code. The first is the priority code, which alerts users to items of special interest, including major breaking news, notification of corrections, clarifications or kills (when a story is withdrawn from the wire), etc.<sup>45</sup>

F	Flash; highest priority <sup>46</sup>
B	Bulletins, alerts, kill notes
U	Urgent, high priority, corrections

---

<sup>44</sup> Associated Press. "Filing the Wire." *Associated Press Stylebook and Briefing on Media Law*. Ed. Norm Goldstein. New York: Basic Books, 2007., 390

<sup>45</sup> Ibid., 391

<sup>46</sup> Teletype machines were equipped with bells set to ring according to priority.

R	Weekly Features
D	Business and financial Datastream
A	Limited Datastream
S	Sports Datastream

The other is the category code, a list that has been expanded recently to allow computers to channel specific news topics to the appropriate pages on an AP member’s or subscriber’s Web page, as well as representing AP’s expanded coverage of lifestyle and entertainment topics. The subject categories serve as facets, permitting a hierarchical vocabulary underneath, e.g. Sports—Baseball—Dodgers.

A	Domestic, general news not originating in Washington, D.C.
B	Special events
D	Standing advance features about food and diet
E	Entertainment news and features
F	News for financial pages originating anywhere
I	International news
J	Lottery results
K	Commentary
L	Weekly Features packages
N	State or regional stories
O	Weather information, including forecasts and tables
P	National politics
Q	Sports scores
T	Sports news
W	Stories originating in Washington, D.C.

A final set of transmission codes sets up some basic filters that anticipate how AP feeds will appear on the printed page, allowing members’ production systems to introduce programmatically some basic, local typesetting data. Because AP transmits so much tabular matter (sports box scores, financial tables and election results), codes indicate the presence or absence of table tags among the data within the feed. Two other codes establish whether the data is destined for ordinary text fonts or small, sans-serif, or “agate,” type. The importance of these codes to the new categorization project is described below under “Metatagging.”

### *The AP Paper Archives*

As noted in the timeline, **the Associated Press started a news library in 1904, in order to provide reference services to the journalists.** This work probably began on the news floor itself, with the indexing of telegraph cables and newspaper clippings. AP also relied on Burrelle's Clipping Bureau for compilation of clippings of AP stories and stories about AP. Cables, clippings, and eventually wire copy were pasted onto stiff paper, which made the flimsy bits of paper easier to handle and kept discrete segments of wire transmissions together. These "pasted sheets," indexed by geography and subject, form the bulk of what was eventually microfilmed, beginning in 1948 and continuing intermittently through the late 1980s. To save space, most of the pasted sheets were destroyed after filming, but there are several years' worth of unfiled sheets in the Corporate Archives.

Selection criteria were obviously subjective, and, of course, there were no trained librarians or archivists at AP in the early years.<sup>47</sup> By the 1950s, general guidelines called for the selection of important stories and bylines of "star" reporters, but the selectors had a lot of latitude. In New York in the early 1960s, there were two selectors and four clerks performing pasting operations.<sup>48</sup>

Not all copy reached the formal "pasted sheet" stage of archiving; individual departments sometimes maintained their own collections of loose clippings and other reference material. In particular, the sports department indexed and maintained its own files, and this collection, dating to the 1940s, complements the material kept in the news library. Card indexes still survive for many of the collections, although various efforts to save space have taken their toll on the indexes and pasted sheets.

Another interesting part of the collection comprises bundles of copy compiled almost exactly as it came off the teletype—in reverse order. In the workflow at transmission centers and the larger bureaus of the day, copy clerks would cut the wire as it rolled continuously off the teletype (in rolls or fan-fold), and parcel it out to the appropriate editors. **Once the editors finished with the copy, it would go onto a metal spike,<sup>49</sup> and when the spike was full, the whole lot was pulled off and prepared for its next use as reference matter.** Known as "booking," this process consisted of threading a long brad into the spike hole along with a cardboard cover, then appending a label naming the wire and a date stamp. The "book"

---

<sup>47</sup> The evolution of the "morgue" in the news industry is hazy, but the introduction of professional librarians in newspapers is usually dated to the early 1920s. The Newspaper (later News) Division of the Special Libraries Association was created in 1923. See Semonche, Barbara, "The History of News Libraries," at <http://parklibrary.jomc.unc.edu/newslibhist2.html>, which is excerpted from *News Media Libraries: a Management Handbook*, Westport, Conn.: Greenwood Publishing, 1993.

<sup>48</sup> I am told by retired journalist Sue Avery, who began her career as a sheet paster for AP in New York in 1960, that a less stellar reporter who complained loudly enough might rise to having his material selected, but full runs of wires were obviously too voluminous to be processed more than minimally by a staff of six. Personal communication, May 25, 2008.

<sup>49</sup> These were eventually outlawed by OSHA in the 1980s.

then went into a slot in a set of pigeonholes, where it served as a reference copy for the two most frequently asked questions: “Have we had this story already?” and “What did we write last time?” The books accumulated for several weeks or until the slot became full. Off they went to longer-term storage and an uncertain future; some of these books still exist.<sup>50</sup>

**In rare cases where editors were cognizant of an extraordinary news event in the making, the relevant wires might be set carefully aside.** The most stirring example of this is the collection of wire copy documenting the first reports out of Dallas of the Kennedy assassination in 1963. The copy, its lines of text distorted by the rapidity with which the paper was yanked from the teletype machine, bear the telltale spike holes and copy pencil markings of history being written moment by moment.

#### *AP Electronic Databases*

Wire service news was paper-based until the early 1970s. The Atlanta bureau led the beginnings of the digital revolution with the introduction of editing terminals and an electronic hub that fed the new high-speed DataStream service. But even with the advances of CRT terminals and electronic editing, paper continued as the medium of archiving for at least 20 more years.<sup>51</sup> Library staff continued to paste paper copy, a bulky proposition. The inevitable storage issues were solved with another flurry of microfilming by ProQuest/UMI starting in 1979. But even as the library struggled with the paper workload, new developments in the newsrooms were paving the way for born-digital databases.

**In 1977, AP began sending nightly feeds of its A-wire to Nexis, and in exchange received a discount on its subscription. But the most significant step in the development of the electronic text archives came in 1979, when AP implemented a system called VuText.** This early attempt at electronic distribution of daily news to the public was aimed at those tech-savvy early adopters who were beginning to purchase personal computers for home use. VuText, owned and developed by the Knight-Ridder newspaper chain (now defunct), created partnerships with news organizations to provide content and test the market. The test lasted only a few years, doomed by clunky interfaces amid rapidly developing alternatives. But the newly created databases did offer a useful way of replacing laborious clipping and pasting operations.

---

<sup>50</sup> They are known by AP veterans as “bales of hay” for their friability and golden-yellow color. Personal interview with Jim Kennedy, senior vice president for strategic planning.

<sup>51</sup> Most of this history comes from Bruce Toll and Tim Gallivan, another technician and developer AP’s news technology department through the period.

The VuText database was populated by reporters and editors working on the newsroom editing system. When an editor filed a finished story (including updates and writethrus in the case of breaking news), a copy would be passed to VuText. Beginning in 1985, the library was given access to VuText as a research and reference tool, along with responsibility for maintaining the content. Metadata and story structure did not always parse accurately between the newsroom writing and editing system and VuText, so the librarians would spend a large part of their shifts cleaning up the ingested data. Nor were all wires represented initially in VuText; local content was omitted as too voluminous. The library continued to produce pasted sheets for a year or two into the VuText implementation as system problems were worked out.<sup>52</sup> (Paper copy also exists from the wires up through the mid-1990s but its provenance is not clear, and it overlaps with what was in VuText as well as Nexis and the microfilm through that period.)

VuText continued to accumulate content for ten years. As an access system, however, newsroom users found that it fell short of expectations. Like Nexis, VuText required a mediator—a news librarian—to perform a search; in the opinion of the journalists, this was a hindrance in a deadline-driven environment.<sup>53</sup> **In 1992 AP began an effort to integrate archives functionality into its in-house production system.** The goal for the Reporter’s Workbench was to allow a reporter working on a story to perform an archives search (again, “Have we had this?” and “What did we write?”) without exiting the Workbench program and by using the same commands used in searching and retrieving within the production database. It was a major technical challenge at the time, requiring heavy customization of a commercial search engine and nearly three years of development.<sup>54</sup>

**1996 is considered the beginning of the digital text archives in the form they exist in today,** because at that point they began to be populated by stories and other content flowing from the production system. When a journalist filed a story to the wire, it triggered the archiving ingest process about a day later, and was available in the meantime in the Reporter’s Workbench. The initial 24-hour lag before content was archived was due to the volume of daily material and the technological difficulties of ingesting and indexing material in real time. AP has since overcome that hurdle, and now once a story is filed, it is available in the archives immediately.

According to AP’s internal guide to the text archives, they comprise:<sup>55</sup>

---

<sup>52</sup> Barbara Gellis Shapiro, AP library director from 1979 to 1994, provided me with historic background on the news library and her recollections of the VuText implementation. Personal communications, May 12-28, 2008.

<sup>53</sup> “Useless to the newsroom,” in Gallivan’s words. Personal interview, May, 9, 2008.

<sup>54</sup> Ibid.

<sup>55</sup> Compiled in 1999. AP news librarian Susan James that additional series of text have undoubtedly been added but are not documented.

- All A, F, and T wire copy from 1985 to the present
- The national and state wire copy from January 1, 1995
- International wire copy from March 18, 1996
- AP Broadcast copy from January 1, 1997
- AP Online copy from September 30, 1997
- Foreign language copy (French, Spanish, German) from 1998
- Radio and TV wires

The problem of ingesting *retrospective* material—i.e., that which had originated in the VuText system—was another thorny challenge at the time, not only because of the volume of files, but also the complication of parsing the old VuText content into a form recognizable by the new search engine, and then loading it into the new database. The project team eventually decided to parse an initial three- to four-year block of archives (back to about 1992), and then add older files in one-year increments back to 1985. Influencing this decision was the measurably low usage by journalists of older material; as part of the project developers tracked retrieval and found that files were considered “old” after about six months.

The sheer size of the database was an issue at a time when data storage was still considered a significant expense. Developers hadn’t reckoned with the permanent requirements, which turned out to be about 80 gigabytes the first year—a new \$80,000 expense. However, the benefits were immediate. A reporter in Rome could be dispatched to a hot spot in Africa and, thanks to the database, be an “instant expert” on the story’s background. The new budget item was accepted without hesitation.<sup>56</sup>

### *Metatagging*

The taxonomy project, launched in 2005, has resulted in APPL, AP’s proprietary publishing language in XML. The power of eAP, according to AP’s 2006 annual report, is to allow members to “create valuable topical categories of news drawn from all AP content, state wires and international English language wires.”<sup>57</sup>

Because it is a proprietary language, details about APPL are not available. However, it is informed to some extent by a parallel effort by IPTC and its development of NewsML, which specifies how legacy categorization codes and the packaging of related news objects are to be handled in the NewsML document type definition. AP’s metadata project has proceed much more quickly than IPTC’s and is deeper and richer than that envisioned for NewsML and its text protocol NITF (news industry text format). For AP, content structure has clearly evolved beyond a standard for transmission into a

---

<sup>56</sup> Ibid.

<sup>57</sup> Annual Report 2006. New York: Associated Press, 2007, 19.

marketing tool, allowing it to offer members and subscribers the advantage of well-formed and richly described content.

In APPL, high-level facets, called “channels” by AP, denote broad topical areas (health, technology, sports, entertainment, news topics) with hierarchical subject terms, personal names, company names and geographic locations. Reporters and editors assign the first level of tagging as the story is written and edited. At the point of filing, the story then undergoes auto-categorization and is then piped to eAP, where members view the tagged content. A member submitting material for tagging by AP transmits his stories to AP where they enter the AP Pipeline for tagging and are returned in APPL format.

#### *Migrating the Text Archives into eAP*

According to AP technologists, the text archives will be migrated in two phases: first, extraction from the database of more than 65 million text files (expected to take 2-3 months) and loading them into eAP, and second, converting the files into APPL (Associated Press Publication Language) XML documents and performing metatagging. To ensure that the transformation doesn’t put an undue processing load on a live database, the project will use a low-priority input queue to give news priority to computing resources. The extracted files will consume roughly 1 terabyte of disk space and will be copied to a large network share drive. Again, the newer documents will be processed first, working backward to the earliest content.

**One of the crucial decisions that AP made for the migration was to include the auxiliary and supplemental material in the archives: the writethrus, budgets or digests (still known by the cabese term “bjt”), corrections and advisories, and certain internal communications.** Preserved in the metadata in eAP will be the provenance of each file—i.e., what database it originated in—as well as linkages between writethrus and other versioning structures. While it greatly adds to the complexity and size of the migration, this material is unique to AP and does not appear in any other sources of legacy AP data, such as news aggregators and member newspapers that haven’t filtered AP content from their own archives.

For the past two years, text files have been flowing into eAP, along with photos and graphics as the VLDB is built out. Users, including AP members and subscribers, use a tool called AP Exchange to access up to several months’ worth of older material. One of the issues still to be resolved in the migration project is how to handle the overlap between the approximately two years of text already piped to eAP and identical content in the Reporter’s Workbench.<sup>58</sup>

---

<sup>58</sup> Mike Alexander, personal communication, June 11, 2008.

Reporters and editors will continue to use the Reporter’s Workbench as their production tool and window into the text archives, but instead of running a search against the text silo, a query will now be directed to eAP. The change is expected to be seamless for the newsroom.

*Prospects for Digitization of AP Paper Files*

As for the rest of the material still in their analog “siloes”—paper, pasted sheets and film—the incentives to digitize aren’t obvious. It is a tempting notion to imagine a digitization project that would unlock access to historic files and apply AP’s enhanced metadata and tagging to the results, but any such efforts are likely to be in discrete niches and collections, if ever. There are markets for old data beginning to develop; while I was in New York one of the AP Digital staffers was starting on a project to understand how current metadata would line up with the archives of the financial wire back to 1937. **Driving the project is not historic interest but computational text analysis and data mining in search of market trends.**

On the following page is a timeline of the text archives from the 1840s to the present, indicating some milestones along the way that had an impact on the retention or loss of the material. A table below describes what is known to be in the archives, regardless of format, categorized by wire or other source. The microfilm, again, is of the series of pasted sheets, which were selected out of the much larger daily wire feeds.

## Associated Press Text Archives: timeline and triggers

Growing cooperative changes location frequently. 'Pasted sheets' made but not systematically retained. Some corporate documents kept after 1900 incorporation

1938: AP settles at 50 Rockefeller Center; earlier files discarded

1940s: First microfilming of pasted sheets begins: important stories and selected authors.

1970s: Additional microfilming.

1980s: VuText introduces data commercial data aggregation and limited electronic archives at the AP

1990s: Newsroom federates archives search with production system.

2000s: 'Digital Cooperative,' headquarters move, creation of corporate archives; metatagging initiative, migration to eAP



1846

1930

1940

1950

1960

1970

1980

1990

2000

Text Archives	Coverage	Format	Contents
A and B wires	1937-present	Microfilm (1937-76, 1983-85, partially indexed. 1977-82 missing) □ Nexis (Incomplete from 1977) □ Wire copy duplicating contents of VuText (?) stored in different locations (1987-1995) □ Electronic text archive (1985-present for A wire, 1996-present for B)	Wire copy, clippings, partially indexed by subject and writer, organized chronologically by country. Subfiles include important news categories ("Europe Disorders")
AP Newsfeatures	1940-present	Microfilm (1940-1979) □ Pasted sheets (non-filmed wire copy on heavy paper) (1980-86) □ Electronic text archive (1996-present)	Longer-format stories of special or general interest, some related to news events
Sports wire (s and t)	1937-present	Wire copy, some pasted up (1964-88), indexed □ Microfilm, 1937-85 in different locations □ Sports Reference (1947-90) □ Electronic text archive (1985-present)	Organized by type of sport (baseball, tennis) and sports figure, with additional subject files such as Baseball Lawsuits, Olympic Games. Sports maintained its own backfiles and developed a filing and categorization scheme. Reference collection is primarily personalities and preparedness for eventual obituaries
New York Bureau	1967-87	Microfilm (1967-87) □ Electronic text archive (1996-present)	Local news produced by New York-based journalists
Financial	1937-present	Film 1937-85, missing 1977-82. Electronic text archive 1985-present.	Financial news minus stock tables and tabular data

## 6. Current Policies, Strategies and Vulnerabilities

In more than a dozen conversations with AP stakeholders in New York and elsewhere, I heard many times that the surviving archives of the last 162 years are a tremendous historic treasure—but one of uncertain value to the Associated Press as a business. **Typical of news organizations, AP has never had a preservation policy per se. As the analysis of the text archives will show, we owe a great deal to benign neglect as an explanation of why there are any surviving archives at all.**

Recall that news archives, for news organizations, do not serve the classic definition of providing evidence of organizational activities; they document news events and thus comprise more of a “library” of discrete accounts of tens of thousands of events annually.<sup>59</sup> The so-called archives are created almost entirely as a research and reference tool for the journalists to use in the day-to-day writing and editing of new material.<sup>60</sup> Answering the two frequent questions, “Have we had this story already?” and “What did we write last time?” for many stories seldom requires a search deeper than several months.

The “appraisal” of news in archival terms is a highly subjective exercise, as generations of copy-pasters demonstrated. It is not so straightforward to determine whether a particular text or image is expendable, and any discards are controversial. In many organizations, not just news, digital appraisal is seen as expensive and virtually unnecessary, as long as storage can be added to accommodate the growing terabytes.<sup>61</sup> It is so much easier to allow material to accumulate than it is to make the hard decisions, any one of which could mean lost opportunity.

Clearly, however, material does get discarded. But any activities by news archives that could be vaguely classified as appraisal, accession, retention and disposition are still ad hoc, idiosyncratic and low-priority.<sup>62</sup>

---

<sup>59</sup> The notion of a news item in a database as a record of a “transaction” has not been explored in the literature, although there are certainly instances where a file might be important for identifying, say, the date or location of work by a particular journalist irrespective of the news. News items are occasionally used by third parties in courts of law to establish whereabouts or time frames, but these incidents are isolated. Copyright concerns are another aspect of the “transaction” model that have led to much more robust metadata in each “record.”

<sup>60</sup> Tim Gullivan, conference call, May 15, 2008

<sup>61</sup> In news, this has been sharpened into focus by a freelance photographer named Dirk Halstead, who had among his thousands of film negatives a photograph of President Clinton embracing a young woman in a crowd. She turned out to be Monica Lewinsky, and Halstead achieved instant fame himself for his canny management. The case is widely cited by photographers as a reason to store digital outtakes permanently.

<sup>62</sup> McCargar, Victoria, and Shannon Supple. “News Archives Survey.” InterPARES 2 Description Cross Domain Group Report to SSHRC. Ed. Anne Gilliland. Los Angeles: University of California, Los Angeles. 26-29 plus appendix. Publication in process.

### *Decision-making*

**Historically, relocating offices appears to have been the most predictable trigger for taking file-management action at the Associated Press.** The earliest wire copy in the microfilm collection is 1937, which coincides with preparations for AP's move in 1938 to its new headquarters at 50 Rockefeller Plaza in New York City. It is reasonable to surmise that a major weeding project was undertaken ahead of the move and a lot of content discarded in the process.

Valerie Komor, the corporate archivist who herself was hired in anticipation of AP's 2004 move *out* of Rockefeller Center, has observed this pattern in AP's bureau offices. A move to new quarters has often been the occasion of extensive file destruction. For example, when Saigon Bureau correspondent Peter Arnett changed offices in 1972, headquarters instructed him to destroy the bureau files. Instead, he paid to have some of them shipped to himself, and they followed him around the world until he finally passed them on to AP in 2006. His Saigon material is now an important collection of the Corporate Archives.<sup>63</sup>

The extent of bureau archives seems to be dependent, to a certain extent, on the habits of whoever is bureau chief. According to Komor, some long-established bureaus will have very complete local files—much of it unpublished material—while others much less. Of course, in some instances there are few decisions to be made; Arnett's is an example of the kind of situations that trump strategic accession plans in the news industry. Temporary bureaus set up for a particular event—a war, a coronation, the Olympic Games—did not systematically keep their files. In the paper era, what became of the material is often unknowable. In the early days of AP's digital production network, the Mouse, the only storage alternative for many smaller bureaus was floppy disks, most of which disappeared. State and local stories that weren't important enough to warrant the national "A" wire are thus missing from the electronic databases prior to 1996.

Further, a policy of permitting departing journalists and retirees to take their personal notes and files with them has placed unpublished AP content in the hands of manuscript collections—and garages—around the world. **But the sensitivity and confidentiality of reporters' notebooks (and photographers' outtakes) and specter of a legal discovery action add layers of complication to corporate appraisal and have long made news librarians and their lawyers leery of retaining unpublished material.** It is safer to let historic material walk out the door, the thinking goes, than to try to keep it.

---

<sup>63</sup> The Saigon files are not complete, however. According to Komor, Arnett attempted to retrieve the remainder of the files in 1975 after the city had fallen to the Viet Cong. He got as far as having the boxes loaded onto a ship, but it was looted by the North Vietnamese while it was docked and the files destroyed. Personal communication, June 16, 2008.

These considerations seem to add up to a policy of benign neglect in combination with individual initiative and inertia; material in some instances appears to have survived at AP because no one bothered to throw it away. However, it is difficult to argue that, at least for the first hundred years, the Associated Press had a compelling reason to think about its legacy news content. **Simply put, AP content was everywhere when the dominant form of mass communication was the daily newspaper.** The front page of the *Houston Post* for July 8, 1920, is typical: of 18 stories, 16 carry an AP byline.<sup>64</sup> When more than a thousand other daily newspapers carried many of the same stories, where was the rationale for clipping and keeping an archival copy at AP? Furthermore, when the Library of Congress began accessioning microfilm of these same U.S. newspapers in the 1930s, a “permanent” record was already being acquired elsewhere.

The notion of keeping some its content for *reference* seems to have started early, when the foreign editor—whose minions would race to the wharfs to meet incoming ships for news—began to paste clippings in a scrapbook with handwritten indexing marginalia. The intent seems to have been creation of a reference library of articles, not a historic or corporate archives. As described above, this process and those that followed, including microfilming, resulted in a collection of wire series that are now seen as “archival,” and are indeed referred to as the “text archives,” and their preservation in some form is no longer a question—only how it will be effected.

Komor has promulgated a mission statement that says, in part, that it is in AP’s interest “to expand the Archives and to assure their preservation and access in perpetuity.”<sup>65</sup> Yet to be determined, though, is whether the Corporate Archives’ purview extends to the news archives, especially the digital archives—and eAP. Over the 162 years of AP’s existence, a great many news transmissions have taken on the value of historic artifacts, thus folders of fragile cables—teletypes were always stocked with bad paper—become a matter of physical curation as much creating access to an intellectual resource. The question of whether the mandate to “assure their preservation and access in perpetuity” can be applied to sophisticated curation of multimedia databases will be fascinating to watch in the coming decade or two.

#### “Organic Convenience”

If a policy of benign neglect has sufficed to bring 70 years of AP content on paper and film successfully to the present, what is at work today in electronic archives?

---

<sup>64</sup> The two non-AP articles carry the source line “from exclusive leased telegraph.”

<sup>65</sup> “The Associated Press Corporate Archives Mission Statement,” December, 2005.

The AP's digital migration history is relatively recent. As described above, the most common wires—general news, finance and sports—are retrievable back to 1985. From 1996 forward, everything is kept in increasingly large silos and now in the Very Large Database. **In other words, eAP seems to be taking on some attributes of a repository: unified formats, rich, uniform metadata, a global search strategy and robust backup.** The eAP and Content Enrichment initiatives, which offer advanced automated subject indexing and tagging, hold promise for *future* AP text archives—for content produced from about 2005 forward.

The eAP initiative since 2005 has meant that incoming material is normalized to format with well-formed metadata, but compliance with the Open Archive Information System (OAIS) framework is incomplete insofar as the VLDB lacks the formal layer of digital preservation management that actively monitors format longevity, bit-level integrity, etc.

At present, the operative policy is one of “organic convenience,” in the words of a storage expert in AP's information technology department. **It is an acknowledgement that “we save everything [that goes into the VLDB] forever.”**<sup>66</sup> To be sure, this is an untested strategy that only its ultimate success will validate—10, 50 or 100 years from now. While the Associated Press is not in the business of being a repository, it is effectively creating one, and its ability and willingness to support material of unknown value will be tested.

Moreover, AP's solicitation of members to contribute *their* content for tagging and storage is another aspect of the repository model—the reaching out to external contributors who wish to have their heritage or research material housed in a formal setting. These are interesting questions for AP but certainly not in active discussion amid its many other preoccupations. As was the case in 1894, AP is too busy making AP history to worry about how it is going to capture it. Formalizing its Corporate Archives, if a bit overdue, has at least indicated some institutional interest going forward.

Through “organic convenience,” AP's digital archives will continue to grow without any upper limit on volume or complexity—at least as things stand now. **The process of ensuring access to valuable content—“value” being continuously redefined—will also ensure access to a lot of content of dubious value, because the process of separating the two is simply impossible.**

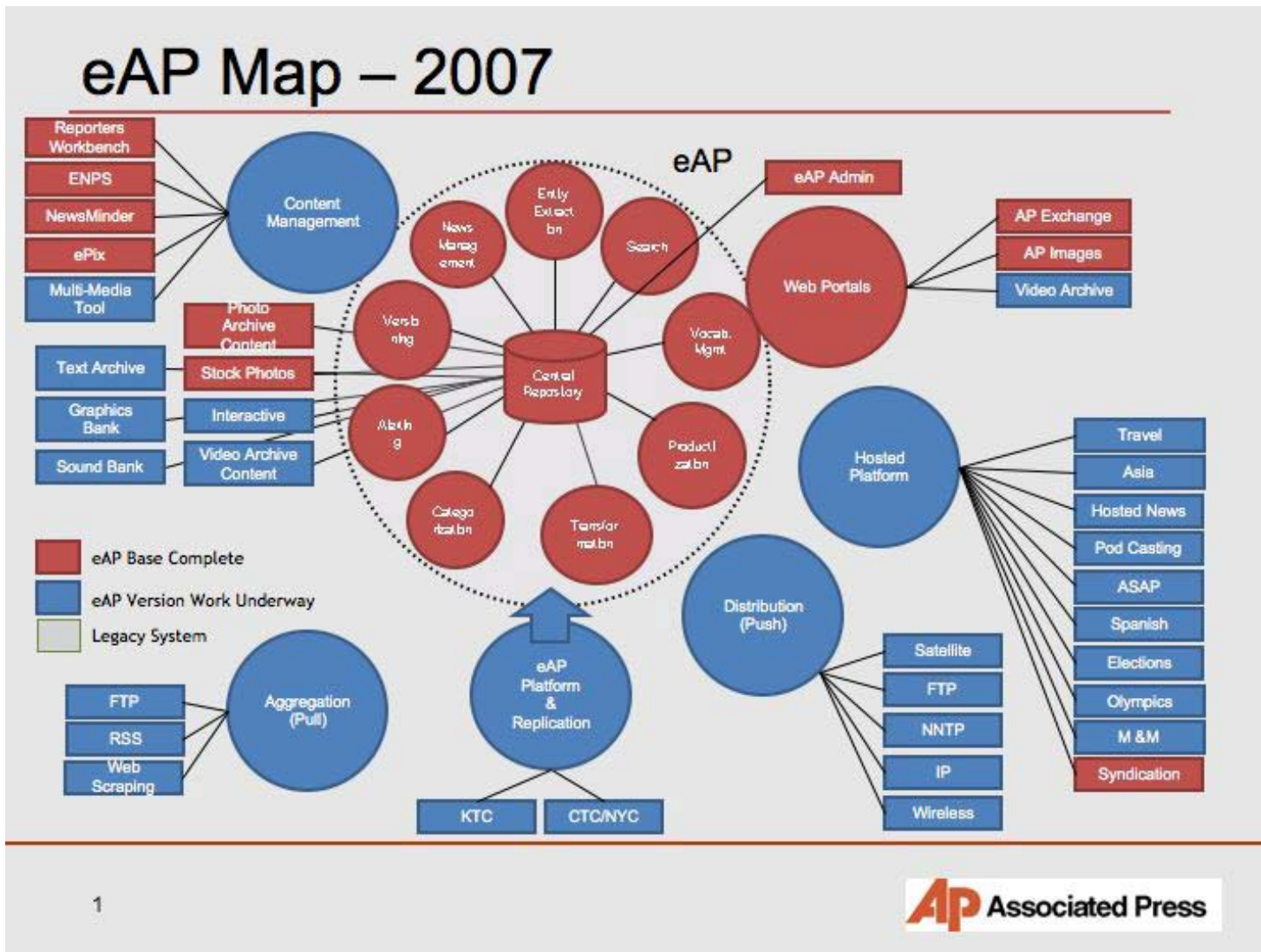
---

<sup>66</sup> Komor, quoting Dan Raju. Email communication, May 2, 2008.

In its paper archives, AP has shown that selected material, like the JFK assassination reporting, has a reasonable chance of being preserved, while the more prosaic content gradually disappears. While many at AP regret that so much has been lost, most recognize that it would have been impossible to keep all of it, and look to doing a better job with the archives in the future. **In a company whose continuing operations depend upon robust systems and the preservation of “value” content (however defined at the moment), the historic gems will help ensure the survival of the prosaic matrix of day-to-day journalism.**

The implications for searchability in this pool of data are dependent on AP’s being very intelligent about metadata, but there is really no other entity in news archives so equipped or so advanced in the process of tagging for the future. Those efforts alone should go a long way toward keeping this written “proto-history” viable and available across data horizons unknown.

Appendix I. AP's "bubble map" of the eAP build-out, latest version. Completed integrations to the VLDB are red.



## Appendix II

Interviews and personal communication, March-June, 2008

### *Corporate initiatives and strategies*

Valerie S. Komor, Director, AP Corporate Archives  
Jim Kennedy, vice president for strategic planning  
Lorraine Cichowski, senior vice president for technology  
Ellen Hale, senior vice president for corporate communications  
Srinidan Kasi, vice president and General Counsel (First Amendment, copyright and licensing)  
Todd Martin, vice president for technology development

### *New product development*

Ted Mendelsohn, AP Digital  
Claire Wachter, AP Digital  
Jay Duquette, AP Digital

### *Current newsroom technology*

Chad Schorr, director of newsroom technology

### *Electronic text archives—VuText, Reporter's Workbench, and eAP migration*

Barbara Gellis Shapiro, former library director  
Bruce Toll, former staffer  
Mike Alexander, Workbench migration project leader  
Stan Miller, Workbench development team  
Dana Bloch, VuText development  
Tim Gallivan, Systems Editor, VuText and Workbench development

### *Text archives—history and development*

Valerie S. Komor, Director, AP Corporate Archives  
Susan James, Deputy Director, News Research Center  
Barbara Gellis Shapiro, former library director  
Sue Avery, former sheet paster

## Bibliography

- Associated Press. "Charter and Bylaws, 1846-2006." New York: Associated Press, 2006.
- Associated Press, and Context-Based Research Group. "A New Model for News". New York, June 2008. PDF. <http://www.ap.org/newmodel.pdf>.
- Goldstein, Norm, ed. "About the A.P." In *Associated Press Stylebook and Briefing on Media Law*. New York: Basic Books, 2007.
- . "Filing Practices." In *Associated Press Stylebook and Briefing on Media Law*. New York: Basic Books, 2007.
- . "Filing the Wire." In *Associated Press Stylebook and Briefing on Media Law*. New York: Basic Books, 2007.
- "Q&A: Jane Seagrave." *AP World*, Spring 2006.
- Bagli, Charles V. "Associated Press to Move from Rockefeller Center." *New York Times*. Metropolitan Section. January 3, 2003.
- Blondheim, Menahem. *News over the Wires*. Cambridge, Mass.: Harvard University Press, 1994.
- Carvajal, Doreen. "A.P. Is Feeling Pressure for Change". New York. *New York Times*. Business Section. June 20, 2005.
- Hoover's Company Records. *The Associated Press*. Hoover's Inc., 2008.
- "(AP)." *Fortune*, February, 1937.
- "AP Leverages Content Tagging Software." *Newspapers and Technology*, September, 2007.
- Liedtke, Michael. "The Associated Press to Impose Online Licensing Fees." *Associated Press*. BC Cycle ed. New York, April 18, 2005.
- Madore, James T. "The Associated Press Drops Plans for Fee." *Associated Press*. July 23, 2005.
- Martin, Shannon E., and Kathleen A. Hansen. *Newspapers of Record in a Digital Age: From Hot Type to Hot Link*. Westport, Conn.: Praeger Publishers, 1998.
- McCargar, Victoria, and Shannon Supple. "News Archives Survey." *InterPARES 2 Description Cross Domain Group Report to SSHRC*. Ed. Anne Gilliland. Los Angeles: University of California, Los Angeles. 26-29 and appendix.
- Reporters of the Associated Press. *Breaking News: How the Associated Press Has Covered War, Peace and Everything Else*. New York: Princeton Architectural Press, 2007.
- Ricchiardi, Sherry. "Covering the World: As U.S. News Organizations Have Backed Away from Foreign News Coverage, the Associated Press' International Report Has Become Increasingly Pivotal." *American Journalism Review*, October/November, 2007.

Singhania, Lisa. "Intellectual Property: A.P. Looks for New Ways to Protect and Maximize Content." *AP World*, Fall 2007.

Strupp, Joe. "Editors at Odds with AP." *Editor and Publisher* Vol. 141 No. 7.

Strupp, Joe. "Two Groups of Editors Pen Strong Complaints About New A.P. Fees, Other Practices." *Editor and Publisher*, January 29, 2008,  
[http://www.editorandpublisher.com/eandp/article\\_brief/eandp/1/1003703492](http://www.editorandpublisher.com/eandp/article_brief/eandp/1/1003703492)

Trott, Nancy. "NewsNow: Fast News for All Formats." *AP World*, Fall 2007.

**In other words, eAP seems to be taking on some attributes of a repository: unified formats, rich, uniform metadata, a global search strategy and robust backup.**