Preserving Digital Local News

Robert B. Allen and Kirsten A. Johnson
College of Information Science and Technology,
Drexel University, Philadelphia, PA, 19104, USA

Keywords:
Archives, community model, digital local news, newspapers, preservation, radio and television news, selection

Abstract:

Purpose
Much local news -- important documentation of local history -- is being lost. The fact that a lot of news media is now available digitally, presents new opportunities but also new challenges for such preservation. We explore issues and approaches for collection and management of born digital local news.

Design/methodology/approach/findings
We examine several specific bottlenecks for implementing this project. For instance, we estimate the size of the problem by estimating how much local news is generated in one US state. We then consider the difficulties in capture, storage requirements, selection, access, and sustainability. We focus on difficulties in selection. Finally, we explore a number of business models for handling these challenges.

Originality and value
While none of the business models offers an ideal solution for the preservation of local news, especially not for multimedia sources, we believe that some of them provide partial answers which should tried. Newspaper Web sites and streaming radio stations should be downloaded. The possibility of capturing video from cable distribution points could be explored.

Preserving Local News

If news is “the first draft of history”[1], what should we save for the future historians to tell them about our times and ourselves? This is an urgent task given that increasingly news is available only digitally and that digital content is ephemeral. Thus, digital preservation has triggered intense activity. However, the proliferation of news and the ease with which it is now disseminated are making it difficult to preserve. While some attempts are underway to save national news media, attempts to save local news are sporadic.

There are many reasons to have primary sources for local history (Gottschalk, 1969; Lowenthal, 1995). Such a historical record can help local citizens in understanding their local community. It can be a resource for genealogists who track the activities of their ancestors. In some cases, local news can have national or international impact. For instance, records of local reports are invaluable in writing the history of the U.S. Civil Rights movement. At a more generic level, evidence such as that collected from local news, can be useful for statistical analysis of social trends. For these reasons, we believe that there is a modest but steady demand for such historical materials. Further, a comprehensive collection would probably exhibit a “long-tail”; that is, a steady demand for particular items, any one of which is of low priority, but cumulatively they are sufficient to justify the preservation effort.

Preservation of local news has long been recognized as a role for libraries (Phillips, 1995). The simplest approach is to collect and put old newspapers on a shelf. More ambitiously, the U.S. National Endowment for the Humanities (NEH) and the Library of Congress (LC) have supported microfilming old newspapers under the United States Newspaper Project (USNP). Recently, there has been a effort to digitize the

[1] This quotation is attributed to Philip Graham

microfilm as part of the National Digital Newspaper Program (NDNP). However, little attention has been paid to the preservation of news that is born digital. There is a lot of digital news but there are also many challenges. Clearly, for a digital news repository, we would like to have broad coverage across media type and geography.

What is News Anyway?

The traditional view is that news is:

> information about recent events or happenings, especially as reported by newspapers, periodicals, radio, or television. A presentation of such information, as in a newspaper or on a newscast (from dictionary.com).

To a media professional, such criteria as timeliness, proximity, and impact play central roles as to whether or not an event will be covered. Even after an event or story is covered the journalist must decide which information will reach the audience and what will be omitted.

> Newspapers…. provide a unique and readily accessible glimpse of the unfolding nature of events. They indicate that state of knowledge or of public opinion at a given time that no amount of subsequent analysis and more considered reflection can provide. Newspapers are not merely historical sources for academics but have an equally important role in education and for all that are interested in the past. Of course any reasonably sophisticated reader knows that all newspapers are at times inaccurate or else select, interpret, and at times distort the events they report. Indeed some newspapers will print what amounts to little more than barefaced lies (Stoker, 1999).

Beyond traditional news articles, many other types of content appear in newspapers, on TV, or on-line that attracts audiences' attention. These include, but are not limited to, entertainment news, opinion pieces, sports, weather, movie reviews, and advertising. Similarly, television and radio news-broadcasts include a variety of soft news stories. While not minimizing the value of journalistic professionalism and ethics, news is a social artifact and necessarily reflects some bias in its collection, presentation, and distribution. Although, there is some debate about the value of what is officially recorded as news, it is still a significant social record. A pragmatic, operational definition of news is "the product of news organizations".

News, of course, has been affected in several ways by new technology. The nature of news is changing because of the digital world's 24-hour news cycle. Fresh and novel content is needed. Moreover, because it is easier and cheaper to produce and distribute news, there are an increasing number of news outlets such as citizen journalism. Here, however, we emphasize traditional news outlets which still provide the broadest and most systematic coverage.

Preserving Traditional News Media

The value of preserving news has long been recognized. It is relatively easy to preserve a small number of individual newspapers; they can be put on a shelf. However, "just binding the newspapers currently received by the Library of Congress is estimated to produce 50,000 volumes per year" (Cox, 2002, p. 76). Because of the microfilming effort for newspapers under the USNP, the emphasis on saving paper copies of newspapers has diminished. However, Baker (2001) has criticized this by pointing out that the original newspapers have some qualities that are not well captured by microfilm (e.g., occasional color plates). This was answered by Cox (2002) who responded that there is not sufficient warehouse space to save the physical papers, and that does not include the difficulty in providing access to them.

As old film reels decay and the technology to play back those reels becomes obsolete the opportunity to preserve traditional tape and film media is slipping away

> The modern media upon which information is recorded are short-lived. Poor quality paper decomposes and color photographs fade. Electronic impulses become illegible due to deterioration

> of the medium or irretrievable because the system that created and supported the record has become obsolete (Ham, 1993).

The Library of Congress and news stations have saved some national network television news, but already much of the local video has been lost.

> The most devastating losses have already occurred among files of news film and videotape produced by local television stations throughout the United States. Devastation that has taken place because station owners and executives, often remote from daily broadcasting operations, failed to see much cost-benefit value in keeping recordings of old news (Library of Congress, 1997).

In the 1970's many television stations stopped using 16mm film and changed to a 3/4 inch video format. Fewer than 10% of stations transferred their film to public archives; in fact, most of the film was destroyed. Now that videotape is in use at many local television stations, the story is still the same; most tapes are saved for a week, and then the tape is recycled. The National Historical Records and Publications Commission (NHPRC) is trying to help save local television history by financing several local television collections, including a grant to 11 local television news projects aimed at improving preservation and access (Library of Congress, 1997).

The television news archive collection at Vanderbilt University may be, "the world's most extensive and complete archive of television news." The collection contains more than 30,000 news broadcasts from the major U.S. broadcast networks since 1989. This archive allows visitors to the site to search for individual stories as well as entire newscasts (Vanderbilt University). Although this is a large amount of material, it is small compared to the total amount available.

Can we save Everything?

Because so much disk space is now available, could we, perhaps save everything? To explore this M.E. Lesk asked "How much information there is in the world?".[2] To answer that, he estimated how many bits it would take to digitally capture all information recorded in the Library of Congress (e.g., by scanning the books) and he concluded that it would be possible to save all the information there by using all the disk capacity available in 1997. In an analysis of the total amount of information that people produce, he concluded, "There will be enough disk space and tape storage in the world to store everything people write, say, perform or photograph." A similar reasoning has led to talk about petabyte storage and, again, the idea of saving everything (Bell, et al., 2006): "My life bits is a lifetime store of everything."[3]

One interpretation of "store everything" would be to "store digitally everything that everyone remembers." To get an estimate of this, Lesk adopts Landauer's 1986 estimate of the capacity of human memory; however, that estimate has been widely debated. As a result of all these analyses, Lesk concluded that "the implication is that in the year 2000 we will be able to save in digital form everything we want to." This general claim about saving all information is, of course, broader than any of the supporting analyses. Although Lesk makes useful points, we feel that literal interpretation of these conclusions about being able to literally "save everything" is not justified. Some of the difficulty seems to stem from not having a clear definition of "information". For instance, it is unclear that the information in human memory could ever be distinguished from the computing substrate of the human mind and the contextual experiences it captures. For instance, how many bits does it take for a person to know how to ride a bicycle? How many bits to be able to translate languages? Similarly, even if we could video record all encounters of a person, such recordings might document a person's responses to his or her environment but they would not be a useful store of general information that has been acquired, because much of the context for these would be lost and the person's attention and thought processes would not be known. Similarly, the analyses do not fully consider what level of quality is necessary for the stored materials. For instance, even MPEG1 is a lossy

[2] Lesk, M.E. (1997) How much information is there in the world?
http://www.lesk.com/mlesk/ksg97/ksg.html

[3] http://research.microsoft.com/barc/mediapresence/MyLifeBits.aspx

compression technique so that it does not save “everything”. Even MPEG2, which uses about 30 times more data for preservation does not record everything.

Lesk excludes information generated by machines for consumption by machines. Though it’s unclear why that shouldn’t be considered “information”. Although most astronomical data, which primarily automated telescopes generate these days, will not be looked at that doesn’t mean they shouldn’t be saved. Astronomical discoveries are often confirmed by examination of databases of previous observations. While most of those recordings are never examined, having them all available is essential for the few cases that are accessed. Similarly, the save-everything approach does not seem to include human activity that is supported by computing tools. For instance, when a person creates a document with a word processor, we could try to save all keystrokes but apparently not save all versions of the document. We would then have to re-construct all versions from the keystrokes. To do that, we would need emulations of all word-processing software. There are also pragmatic difficulties in Lesk’s analyses. We would have to somehow convince everybody to carry a video camera with them all day, the possibility of disk failures is not considered, and there seems no viable business model for allocating storage in this way. Even if the necessary disk capacity exists, it is not clear that it would be economically possible to use it in that way. A final difficulty with the “save everything” notion is that the volume of digital information grows more rapidly than our ability to process, index, and make searchable that information. In other words, we can propose a corollary to the Moore’s-Law-like increase in disk capacity, that the ability to produce bits is also increasing at an exponential rate. So, the ability to save content remains roughly constant. Technologies such as cell phones, digital cameras, and video camcorders are producing bits at an increasing rate.

Policies for the selection of content to include in an archive are a critical issue for archives (e.g., Jenkinson, 1965). Indeed, “… the archival community must take responsibility for fashioning from this new world of recorded information a manageable historical record” (Ham, 1993). Table 1 summarizes some of the major factors that are frequently considered in selection. Alternatively, we could control the quantity of material to be stored by restricting the quality or the digital objects.

Factor	**Description**
Legal/policy	Requirements to save specific material.
Historical value	Material that reveals aspects of our times and culture.
Economic value	What may have the most commercial value?

Table 1: Some broad criteria which have typically been claimed for selecting materials to be preserved.

Bottlenecks for a Comprehensive Local News Preservation System

Moving past the broad issues, there are several the specific difficulties we will encounter in developing a comprehensive news repository.

Capture

Because local news production is highly distributed, just obtaining copies of the content would be a challenge. There are several options: Capture from broadcast, capture online, or obtain materials directly from the producer. The least obtrusive would be to capture from published (printed or broadcast) news product. However, considerable human effort will probably be required at each of the collection points. Among the tasks for the staff would be quality control, and to be on the look out for novel news sources which need to be added. It would easier to capture online material at a central location. Perhaps, online versions of newspapers and radio broadcasts could be captured, but Web versions of other media offer only limited context compared to the original broadcast.

Storage Requirements

The main issue for storage is the amount of content. To illustrate storage requirements, the table below provides a summary of the disc space for a single state’s news output in one year. In Pennsylvania there are about 50 daily newspapers and about that many more weekly newspapers. In most cases, page images should be available from the publishers. In Pennsylvania there are about 20 news-talk format radio stations and about 20 public radio stations. There are about 40 television stations in Pennsylvania as well as several cable community access channels[4]

Rough estimates for disk storage requirements for digitized newspapers, radio, and television are shown in Table 2. We calculated the bytes required for capturing one television station for a year via MPEG1 compression (1.5Mb). Our analysis demonstrates that it is impractical to save all the television broadcasts at MPEG1 quality. If we save only 4% of the video, approximately, the amount of original news programming on many stations, as shown in the bottom row of the table, the storage would still require 10^6 GB, the equivalent of 1000 times the 1 terabyte usually claimed as the amount of storage required to hold all the text from the Library of Congress.

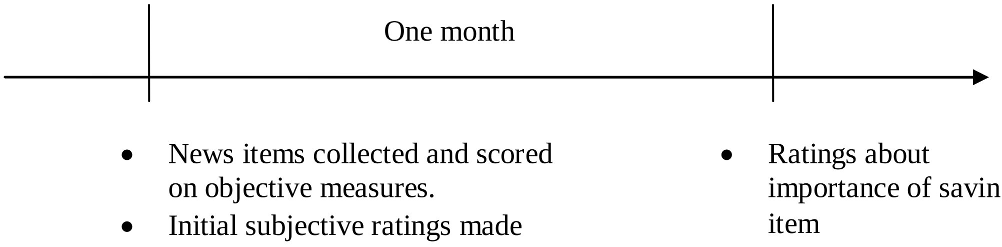
$$\frac{1.5\text{Mb}}{\text{sec}} x \frac{86400\,\text{sec}}{\text{day}} = \frac{2000\text{GB}}{\text{day}} x \frac{365\text{days}}{\text{year}} = \frac{730000\text{GB}}{\text{year}}$$

Medium	**Annual Storage Requirements per Outlet**	**Outlets**	**Total**
Newspaper	353 GB	50	1.2 x 10^5 GB
Radio	250 GB	40	1.0 x 10^5 GB
Television (100%)	730000 GB	40	2.8 x 10^7 GB
Television (4%)	33000 GB	40	1.0 x 10^6 GB

Table 2. Estimated annual storage requirements per year for major news outlets in Pennsylvania

Selection

Because it is not practical to save everything, especially not all video, we need to consider how to select items in a way that is both principled and practical. For instance, we might not save all news items; or, we might use higher resolution for only the most important. As noted above selection of content has been a major debate in the archival community. Here, we follow the judgment of news professional about what is most important. Even then, we ask whether there is reasonable stability across time in the judgment. Ideally, we would aim for same type of automated selection by assigning points (Shirky, 2005) to different attributes. Thus, we could compare the effectiveness of automated (objective) methods to subjective methods requiring human judgment.


One month
News items collected and scored on objective measures.
Initial subjective ratings made
Ratings about importance of saving item

[4] For this analysis, we do not require an exact figure. This estimate in based on web sites such as http://www.mondotimes.com/1/world/us/38

Local news from the Harrisburg, Lebanon, Lancaster, York Designated Marketing Area in Pennsylvania was collected. Specifically, the articles were taken during one week in August 2005 from the *Lancaster Intelligencer Journal*, the local newspaper in Lancaster County, and we recorded 6 PM newscasts from WGAL-TV, the NBC affiliate in Lancaster County, Pennsylvania. These included 74 newspaper articles and 56 TV news stories. We categorized them as: community/feature, government, business, crime, and education.

The participants were students at Elizabethtown College. The first group consisted of 19 undergraduate students enrolled in one of two communications classes in the Department of Communications at Elizabethtown College during the fall of 2005. The participants were given check-sheets containing the criteria for preserving local television news stories and the subjective criteria sheet. The use of the check sheets was explained. Each student was then given randomly two 6 PM newscasts from WGAL-TV and also given randomly two issues of the Lancaster Intelligencer Journal from August 2005. The students watched the newscasts and read the papers individually while filling out the criteria check-sheets. The number of articles rated by a single participant varied depending on the news of the day but each television news story, as well as each newspaper news story, was rated by more than one participant. Direct ratings of preservation priority were made by 11 students in advanced journalism courses who also had some professional media experience. This "expert" group was formed to validate the ratings given by those in Group I. Each of these 11 students was asked to rate five television news stories and five newspaper stories chosen at random.

Table 3 gives some examples of newspaper articles that were rated particularly high or low on the 25 point scale. The ratings for articles divided by category are indicated in Table 3. Articles dealing with the government were rated as being more worth saving. An analysis of variance (ANOVA) was performed on the objective and subjective criteria and the news categories. The differences in categories for the subjective measures were found to be significant ($F(4, 125)=5.29$, $p<0.01$) with "Government" articles being given the highest priority. Perhaps not surprisingly, the materials deemed most worth saving dealt with significant community accomplishments, expenses, and events.

Article Title	**Category**	**Preservation Priority Ratings**
Festival will aid make a wish	Community/feature	6.5
Local lawmakers target pay-hike	Government	19.0
Wohlsen craftsman nets awards	Business	4.8
Two school districts net kudos	Education	17.8
Stolen truck found	Crime	4.5
Coach charged in drug arrest	Crime	23.0

Table 3: Preservation priority ratings for several items along with the categories to which they were assigned.

Category	**N**	**Mean Objective Measures**	**Mean Subjective Ratings**	**Preservation Priority Ratings**
Business	24	5.75	8.00	6.88
Community and feature	54	5.89	5.89	5.89
Crime	23	6.15	7.09	6.41
Education	11	5.36	7.18	6.27
Government	8	5.13	10.88	8.01

Table 4: Mean ratings for items in each category.

Both subjective and objective measures proved to be significant predictors of preservation priority. For the objective measures there was a significant positive correlation ($r = 0.22$; $df=96$, $p<0.0001$). There was an even stronger correlation with the subjective measures ($r = 0.41$, $df=96$, $p<0.0001$). The objective measure was significant but smaller than the effect of the subjective measures.

A regression analysis compared the predictiveness of each factor compared to the objective and subjective scores. The question that explained the most variance for objective measures was, “Whether or not the story was teased at the beginning of the newscast” (R^2 =0.579). For the TV subjective measures R-squared values for all of the questions were R^2 =0.59 or above. The largest R-squared value was for the question, “Will the story be valuable to future generations of researchers?” (R^2 =0.624). All factors were significant at the 0.01 level.

Using the mean score for both television and newspaper stories on the objective and subjective criteria, we found a positive correlation between the two (r=0.285, df=129, p<0.01). There was also a positive correlation between objective and subjective scores for television and newspapers when examined separately. For television (r=0.344, df=55, p<0.01), and for newspapers (r=0.387, df=73, p<0.001). This suggests that the objective criteria can predict the subjective criteria for both newspaper and television. If the criteria could be automated, human judgment would not be needed to determine which stories should be saved.

Supporting Access: Indexing and Browsing

Indexing and providing access increases the cost; thus, “dark archives” – those archives with minimal access to the content – have been proposed. However, accessibility would encourage public support for the project but accessibility is partly determined by copyright. News is difficult to index due to wide subject matter, brevity of some stories, and new concepts and neologisms. Automated systems perform less well on news than on other types of content; for example, medical information about a specific field. Let us consider what navigating such an archive would be like. Indeed, ideally, there could be a fly-through interfaces (e.g., Allen, 2005b). For instance, access can be supported by timelines. Services could also be developed on top of the collection such as summarizing articles and providing access to the collection. In addition, some of the processing strategies come from other news processing projects such as “topic detection and tracking” and new-story detection. Ideally, we would have automated tagging of news stories to determine where there are threads.

Additional Content Processing

Beyond, indexing, the content will need to be processed especially if it is collected from public outlets. With copies, there will be a lot of duplicates. The duplicates would be determined and detected. Duplicate detection is employed by search engines that need to minimize duplicates in Web pages presented to users (Broder, 2000). Similar techniques have also been developed for video (Jaimes, et al., 2006). Unfortunately, redundancy is not a simple concept. While exact matches are relatively easy to identify there are gradations of context. For instance, what should it do if the audio track of a syndicated television show is interrupted by a severe weather advisory? Should the combination of syndicated show and weather advisory be treated as a distinct item? Detecting 90% of the duplicates would probably be relatively easy, the remaining 10% would be quite resource intensive. The duplicate detection problem would be greatly simplified if publishers and broadcasters added metadata to make it easier to identify content (e.g., with SMPTE codes for video). However, that would be a major project in its own right. Another complication is that once duplicates are detected they may need to be stored with extra care because they are re-used by many objects and losing them would affect many other resources.

However, finding duplicates would require extensive computation. Moreover, once a duplicate is found and one of the copies deleted, considerable cross-referencing would be needed to keep track of it. Such a complex system would be a concern to archivists who have found that simple solutions reduce the likelihood of orphaned file formats or outmoded viewing devices. First, news objects would need to be segmented (Hauptman & Witbrock, 1998). For instance, commercials would have to be separated from the news content (as TIVO does). More ambitiously, breaking news stories which interrupt normal radio or television broadcasts.

Risks to the Collections

Ideally, this material would be kept for hundreds of years so sustainability is a fundamental problem. While there is likely to be a continuation of the declining costs for disk storage, there would be an increasing amount of storage required. Moreover, periodically – perhaps every 10 to 20 years -- the entire set of disks would need to be replaced. Because of the large amount of material, even with a small chance of disk failure, there would be large change of losing some material.

There would be both technological and human risks. Some of the technological risks include disk failure and some of the human risks include censorship and targeted attacks for political purposes. The amount of effort that should be put into preventing these problems should be proportional to the probability that they will occur and the damage they are likely to do. 24-hour reliability against difficulties such as disk failure, accidents, disgruntled employees, and targeted attacks for political purposes would be needed.

Business Models

There are several possible business models for funding this preservation project. Each of them presents a different configuration of the issues considered in the previous section and each has distinct challenges with respect to scope and sustainability based on issues such as the difficulty of capture, storage requirements, and rights.

Government

Beyond the intellectual property rights discussed above, the Library of Congress has special rights to demand deposit of all copyrighted works (Besek, 2003). To the extent that the distributor wants to maintain copyright, they would need to send a copy to the Library of Congress. If they were not deposited, we could assume that the copyright owner did not want to exercise copyright; then we could capture it and freely distribute it. The national news may be deposited but it is likely that little of the local news would be deposited. A library can save the paper copy of a newspaper because it has actually purchased that copy. However, it is less clear that a local library could record and save broadcast news for its region. Although there is considerable controversy about copyright, it probably means substantial limitations permissions for capturing and saving digital content. The most direct solution would be to obtain permission from the copyright owners, but other options can be explored. Under U.S. copyright law, the principle of Fair Use provides for limited use of content. Furthermore, under U.S. copyright lay, the Library of Congress has special rights to obtain anything under copyright from the publishers (Besek, 2003) although the files obtained in this way may not be able to be widely distributed.

Because copyright would not be an obstacle, all the material could be collected directly upon publication. However, the cost would be substantial. In Table 5, we provide a rough budget for such a service. This could be funded if we had $0.01/year for every person in the U.S. to devote to preservation of news, a total of about $3M/year. To put the value of this preservation into perspective, the budget for the Smithsonian Museum of American History is about $35M/year.

Central Management	$500K
Central Data Coordination	$500K
Software Development.	$500K
5 Regional Data Centers ($300K each)	$1500K

Table 5. Rough budget outline for a set of news collection centers.

Commercial Providers: News Databases

Most news is produced by commercial entities and, indeed, most local newspapers and radio and television stations are now part of larger media companies. Many of these media companies already coordinate with news database services such as NEXIS (http://www.nexis.com).. Presumably, the rights-holders approve contracts for distribution of their content. However, at least in the past, such services have not been comprehensive; for instance, they have included minimal amounts of audio and video. Nor do they tend to include items such as death notices and editorials. Perhaps the content of these services could be greatly expanded and an explicit commitment made for long-term preservation.

Commercial Providers: News Search Engines

Search engine services such as Google News capture and index news stories. When the user clicks on a link, Google links the user to either the story on the publisher's site or to some other page designated as the target. Of course, the responsibility would then be on the providing news organization to maintain the link. Google News has begun to present archives. That is promising, but still far from the ideal proposal. As with the news databases, much of the context is lost and there is as of yet, the there is no capture of audio or video. Furthermore, Google News remains a lightning rod for discussions about fair use.

Private Foundations

There are several foundations dedicated to preservation of media though, typically, these do not have sufficient funds for operational projects. The Internet Archive (http://www.archive.org/) has been successful collecting commercial web pages, but it has, reportedly, been cautious about recording commercial broadcasts because of the copyright issues. Some foundations have close ties to the media industry. The Museum of Television and Radio (MTR) in New York and Los Angeles whose mission statement is "To be the premier trust of radio and television heritage." In the past, MTR has been focused on national media outlets. Another model could be The Newseum (http://www.newseum.org) is dedicated to First Amendment, news publishing, and journalism. The Newseum web site already displays front pages of many newspapers – although selected dates have been withheld.

Comprehensive Community Information Repository and Semantic Capture

News is an important record of community history but it is only part of a broader picture. News but news is interwoven with many other types of community content. Thus, this project suggests a larger need for a repository of formal and informal community information. This includes all records that are held by a community. For instance, the digital archive of local government will also include personal events (e.g., marriages, births, deaths) and business transactions (e.g., real estate sales)

Cox (2001) points out that many types of traditional records are kept in a community and those interlock. We propose that this is probably even more true of digital material. Such as: sports scores, high-school graduation records, selection records, the police blotter, and court records. Of course, there should probably be an ongoing community review panel to ensure that appropriate material is saved, that local records are well managed, and that there is coordination and consistency across the records. Allen, et al. (2007) describe the utility of such a model for processing digitized collections of historical newspapers.

Summary

Currently, there is no large-scale effort underway to preserve local television and newspaper news stories, and as a result this part of our history is being lost. We have found many practical difficulties to a comprehensive system but there would be value even in a system which was not the ideal. Newspaper Web sites and streaming radio stations should be downloaded. The possibility of capturing video from cable distribution points could be explored.

Given limited resources, we cannot save everything so some selection must be made. We have explored two kinds of selection problems. What should be saved and whether the selection of that can be automated. We have examined both objective and subjective criteria can be used to predict which local news stories should be saved. While we cannot "save everything" we can save a lot. Because there is so much local news, we consider the "save everything" approach to preservation. However, we suggest that this is better characterized as "save lots". Because news media are themselves changing so quickly, selection criteria will have to be continually evaluated.

References:

Allen, R.B. (2005) Using Information Visualization to Support Access to Archival Records. *Journal of Archival Organization*, 3(1), 37-49.

Allen, R.B., Japzon, A.P., Achananuparp,P., and Lee, K-J. (2007) A Framework for Text Processing and Supporting Access to Collections of Digitized Historical Newspapers. *HCI International Conference.*

Anderson, S. and Allen, R.B. (in preparation) Documentation and Structuration: Appraisal, Social Theory and the Archive. To be submitted to *Archivaria.*

Baker, N. (2001) *Double Fold: Libraries and the Assault on Paper.* Random House, New York.

Bearman, D. (1999**)** Reality and Chimeras in the Preservation of Electronic Records**.** *D-Lib.*

Bell, G., Gray, J., and Szalay, A. (2006) Petascale Computational Systems: Balanced CyberInfrastructure in a Data-Centric World. Letter to NSF Cyberinfrastructure Directorate, *IEEE Computer,* 39.1, 110-112.

Besek, J.M. (2003) *Copyright Issues Relevant to Creation of a Digital Archive*, CLIR, #112. http://www.clir.org/pubs/reports/pub112/contents.html

Broder, A.Z. (2000) Identifying and filtering near-duplicate documents. *Symposium on Combinatorial Pattern Matching*, 1-10.

Cox, R.J. (2001) *Documenting Localties.* Chicago, Society of American Archivists.

Cox, R.J. (2002) *Vandals in the Stacks? A Response to Nicholson Baker's Assault on Libraries.* London, Greenwood Press.

Dale, R. (2003) Trusted Repositories for Preserving Cultural Heritage. *ERPA Workshop, Rome.* http://www.erpanet.org/php/Rome/dossier_final_version.pdf

Dozier, D.M., and Rice, R.E. (1984) Rival Theories of Electronic Newsreading? The New Media: Communication, Research and Technology (ed.) Rice R.E. and Associates, Sage Publications, Beverly Hills, 103-128.

Gottschalk, L.R. (1969) *Understanding History,* 2nd edition, New York, Knopf, 1969.

Ham, F. G. (1993) *Selecting and Appraising Archives and Manuscripts*. Chicago, The Society of American Archivists.

Hauptman, A. and Witbrock, M.J. (1998) Story Segmentation and Detection of Commercials in Broadcast News Video, *Advances in Digital Libraries*, Santa Barbara CA.

Ide, M., MacCarn, D., Shepard T., and Weisse, L. (2002) *Understanding the preservation challenge of digital television.* Washington, D.C., Council on Library and Information Resources and Library of Congress: 67-79.

Jaimes, A., Chang, S.F., and Loui, A.C. (2006) Duplicate Detection in Consumer Photography and News video, *ACM TOIS*, 1-50.

Jenkinson, H. (1966) *A Manual of Archive Administration.* Rev. 2nd ed. (London: Percy Lund, Humphries & Co.; org. published 1922).

Jimerson, R. C. (ed.) (2000) *American Archival Studies: Readings in Theory and Practice.* Chicago, Society of American Archivists.

Library of Congress (1997) Television and Video Preservation: A Report of the Current State of American television and video preservation. 3 vols. Washington, D.C.: Library of Congress.

Lowenthal, D. (1995) *The Past is a Foreign Country*. Cambridge, UK, Cambridge University Press.

Phillips, F. (1995) *Local History Collections in Libraries.* Libraries Unlimited, Englewood, CO.

Shirky, C. (2005) AIHT: Conceptual Issues from Practical Tests, *D-Lib,* http://www.dlib.org/dlib/december05/shirky/12shirky.html

Stoker, D. (1999) Should Newspaper Preservation be a Lottery? *Journal of Librarianship and Information Science,* 31, 131-134.

Vanderbilt University Television News Archive, http://tvnews.vanderbilt.edu/.

Robert (Bob) B. Allen is at the College of Information Science and Technology at Drexel University where he teaches courses and does research on information retrieval, question answering, digital libraries, multimedia, and information management. He was the first to develop and publish the basic algorithms used by recommender systems. He is currently working on visualization tools for interacting with events and narratives and he is the PI for grant to develop a model curriculum for the management of digital information. Bob has been Chair of the Publications Board of the ACM and was Editor in Chief of the ACM *Transactions on Information Systems*. Before moving to Drexel, he was a Professor of Practice at the College of Information Studies of the University of Maryland and before that he was Senior Scientist in the Information Science Research Group at Bellcore. He received his PhD in Experimental Psychology from UCSD and joined Bell Laboratories in 1978.

Kirsten A. Johnson is a Doctoral Candidate in the College of Information Science and Technology at Drexel University. She is also an Assistant Professor of Communications at Elizabethtown College. Kirsten has a B.A. in Broadcast News from Drake University, and a M.S. in Telecommunications from Kutztown University. Her research interests include citizen journalism, user created content, and issues of credibility and trust.