Comments to Section 108 Study Group: News Archives

April 28, 2006

Submitted by: **Victoria McCargar, M.A., MLIS**
Independent researcher and news archivist

Research affiliations:

- The Center for Research Libraries
- International Research in Permanent Authentic Records in Electronic Systems (InterPARES)
- Preservation Metadata Implementation Strategies (PREMIS)
- News Division, Special Libraries Association (SLA)

Biographical note: Victoria McCargar is a veteran print journalist, librarian, archivist and preservation researcher. Her newspaper technology experience began with the first newsroom systems at the Los Angeles Times, where she worked for 27 years, and now comprises high-level research projects with aging formats, emerging best practices, and criteria for creating long-range repositories. She is currently working with the Center for Research Libraries on audit criteria for trusted news repositories. She is conducting a survey of news libraries under the auspices of InterPARES (International Research in Permanent, Authentic Records in Electronic Systems) and served on the Preservation Metadata Implementation Strategies (PREMIS) committee 2003-2005. The News Division of the Special Libraries Association has just presented McCargar with its David Rhydwen Award for scholarly contributions to the field. She holds an M.A. degree from the Missouri School of Journalism and an MLIS from UCLA.

Submitted by: **Peter F. Johnson, MLIS**
Independent researcher and news archivist

Research affiliations:

- California State University, Los Angeles
- News Division, Special Libraries Association (SLA)

Biographical note: Peter Johnson has more than 30 years' experience in journalism and multimedia digital news archives. Most recently he was an image archivist at the *Los Angeles Times* with experience in various digital asset management installations. His extensive background includes newspaper editing, writing, photography and production, radio/television production, online research, digital text, graphics and photo retention, information literacy, small- and large-group motivation, personnel management and project management. He holds a B.A. in journalism, a B.A. in radio-television production, and an MLIS, all from the University of Arizona.

Introduction: The nature of news libraries

By Victoria McCargar, M.A., MLIS

Journalism is fondly referred to as the "first draft of history," and indeed, newspapers are often the first iteration of what becomes the historic record. However, for an industry that gives solemn lip service to this responsibility, newspapers pay little relatively attention to what happens to that historic draft in the days or years after publication. The advent of digital workflows and archives has only made the situation more ominous.

Our comments to the Section 108 Study Group are intended to present a voice that is seldom considered in discussions of digital preservation: for-profit news libraries. While academic libraries and government archives are beginning to develop serious solutions to the problems of long-term preservation, newspaper libraries have neither the institutional impetus nor the financial resources to pursue solutions on their own. A large body of historic content dating to the early 1980s is thus at risk.

We wish to present the key considerations of preserving digital newspaper content that have a specific bearing on Section 108. These have been developed in the course of my research since 1998 and are based on my experience in editorial technology, digital archives and rights management at the *Los Angeles Times* (roughly 1985-2005). The comments here touch on a 2005 survey—results as yet unpublished—of emerging best practices in news libraries that will be presented in June to the News Division of the Special Libraries Association. A bibliography is included in this submission.

In this multimedia age, it's important to remember that newspapers are not only about text and photography. The typical news collection may now comprise vector software, HTML, animation, and Web video alongside ASCII, Unicode, and JPEG. It is also an industry that is still waiting for comprehensive standards for content description, rights management, subject vocabularies and bibliographic metadata.

Further complicating the picture, news archives share characteristics with both for-profit and nonprofit institutions: the need to take measures to preserve content of protected or uncertain provenance, and must maintain tight control over their own intellectual property. Greater minds than ours will ultimately have to weigh costs and benefits, but in the meantime we are convinced that this formidable digital "orphan" deserves inclusion in the Section 108 evidence.

A final observation. It is worth noting that these issues would not be unfamiliar to an archivist at one of the film or recording studios. They, too, are struggling with an explosion of format diversity, complex provenance and lack of standards, and preservation is driven almost entirely as a bet toward profits from the preserved material. Ironically, studio attorneys are so focused on protecting intellectual property that they are overlooking the threats to their own content. So in considering the needs of news archives, we might also strike a blow for the seldom-seen archivists from other creative industries.

Section 108 Topics

I. Eligibility for Section 108 exceptions

News archives and libraries should be considered for Section 108 exception because:

1. News archives perform many of the same formal functions of other traditional/brick-and-mortar institutions: organization and preservation of intellectual material (indexing, cataloging, disaster recovery) and retention of material in perpetuity for use in research.
2. As for-profit institutions, news libraries nevertheless hold deep collections of invaluable – and often unique – cultural heritage material, considered by historians to be the most important and most frequently used primary source material.[1]
3. In the digital era, newspapers are now producing material in a wide variety of formats besides print, any of which holds value to historians and other researchers:
    a. Web pages; material may or may not duplicate print editions
    b. Page PDFs; the only resource for historic advertising content other than microfilm
    c. Digital photography in JPEG format
    d. Orphan works (articles and photography where rights are unknown and creator cannot be located)
    e. Outtakes (unpublished images)
    f. Web video (sometimes the source of still photography)
    g. Information graphics (maps, diagrams, charts and graphs)
    h. Flash animation (unique content, no print equivalent)
4. Microfilm is under threat; papers are seeking to eliminate it as a way to cut expenses; material is seen as duplicative of content in PDF "archives." It is not unreasonable to anticipate that all news material will be electronic in the near future.
5. In an era of stagnant or declining corporate revenues, news archives focus resources on print and photography, ignoring other formats but not discarding them. Thus, news archives are accumulating unprocessed stores of obsolete formats that will eventually require archival intervention (reformatting, refreshing, reverse engineering, etc.).
6. Newspaper JPEGs are aging, approaching 12-15 years old (dating from earliest digital archives in the industry) and will require archival intervention.
7. Access to the archives varies by publication; some are open to the public while others are closed to all but editorial staff. Public-access requirement would arbitrarily exclude many repositories for no reason other than the organization's own policy or its ability to handle outside researchers.

II. Three-copy limit

[1] Helen R. Tibbo, "Primarily History in America: How U.S. Historians Search for Primary Materials at the Dawn of the Digital Age." *The American Archivist* 66 (1), 19.

We agree that the three-copy limit should be replaced with a standard more appropriate to the nature of digital obsolescence. There are two considerations for news archives:

1. News archives are often required to maintain old versions of software in order to maintain access to old files in proprietary formats. Some software licensing prohibits successive migrations of a program onto later PC hard disks, which, if honored, would render portions of archives inaccessible.
2. News archives maintain copies (both analog and digital) of orphan works. Where provenance is unknown or the creator cannot be located, successive copying for preservation should be permitted. It is impossible to determine how many times copying (which includes media refreshing or reformatting) might be required over the indefinite lifespan of a digital object, or when such copying might be required.

III. Triggers

1. News archives contain digital files of uncertain provenance as well as published objects to which the archive does not hold permanent copyright (object may, in fact, be licensed from the creator but are retained as a record of publication). These are almost always stored alongside a news organization's own works and may lack the metadata to differentiate them from the organization's own material. They are thus migrated or reformatted in large batch processes in the course of routine system maintenance and preservation activities. Awaiting a trigger such as the imminent digital demise of an obsolete object is untenable and puts the material at risk.

IV. Offsite access

Market-driven consolidation is paving the way for virtual, aggregated libraries of news content; e.g., a media company with a dozen newspapers may decide to maintain a single digital archive for all 12 properties, and limit access via a digital permissions interface. When high-access failover technology is taken into account, such a repository may exist simultaneously in two or three physical locations, invisible to the users. All of the same considerations apply: multimedia formats, uncertain provenance, orphan works, propriety standards. To deny Section 108 exception because of physical location would arbitrarily put news content at risk.

V. Preservation-only repositories and the Web site exception

There is an argument to be made that some news content is not being preserved by the news organizations themselves and that a third-party repository may be the optimal solution to ensure that this content is not lost.

1. A survey of news archives we conducted in the fall of 2005 shows that more than 80% of newspapers are not formally archiving their Web sites. Some are making an effort to extract content for archiving in text databases, but the original presentation is lost.
2. Because news archives exist primarily to serve newsroom research (according to 100% of survey respondents), redundant material is seen as a hindrance to efficient

searching. Identical or similar content from a Web site is likely to be omitted from the archives.

3. The for-profit business model does not encourage preservation activities. In order to leverage user metadata and maximize brand marketing, most news Web sites require registration at initial log in. These interfaces preclude automated harvesting of the sites by the Internet Archive.

4. Content of unknown future value (such as unpublished photography) is gradually reformatted to nearline or offline storage, where it eventually becomes obsolete.

5. Digital preservation is still an evolving concept for news archives. The term more often refers to digitization of microfilm or old photography for sale. In our survey, almost two-thirds of respondents said they did not know whether any preservation activities were prohibited by the copyright act, by which one can infer that they are either engaging in unlawful procedures or failing to intervene for preservation where it is permitted.

6. As an industry, the news media lack formal mechanisms (i.e., standards and best practices) for digital preservation.

VI. Recommendations

Based on our knowledge of preservation issues for news archives, we have concluded that this niche should be afforded exceptions on a several fronts.

1. News libraries should be granted a section 108 exception based on their unique need to take precautions against losing culturally important material. Their for-profit status should not outweigh the needs of researchers and historians for access to this material. Their ability to sustain content in aging or obsolete formats is dependent on archival interventions not currently allowed.

2. Third-party repositories should, in some cases, be allowed to capture copyright news material where the content may otherwise be lost. There are business-case decisions that work against retention of some types of content (such as Web pages) that is nevertheless important to the historic record.

3. In an industry/discipline that lacks standards and best practices, a Section 108 exception may afford enough protection to encourage news organizations to cooperate on preserving material that they might otherwise allow to become obsolete.

Bibliography

Data Dictionary for Preservation Metadata: Final Report of the PREMIS Working Group. Dublin, Ohio and Mountain View, CA: OCLC and RLG, May 2005. www.oclc.org/research/projects/pmwg/premis-final.pdf

Martin, Shannon E. and Kathleen A. Hanson. Newspapers of Record in a Digital Age: From Hot Type to Hot Link. Westport, CT: Greenwood Publishing Group Inc. 1998.

McCargar, Victoria. "Statistical Approaches to Automatic Text Summarization," *Bulletin of the American Society for Information Science and Technology*, 30 (4) April/May, 2004 http://www.asis.org/Bulletin/Apr-04/mcargar.html

McCargar, Victoria. "News That Moves: Accessioning Video for Newspaper Archives," The Moving Image: Journal of the Society of Moving Image Archivists, 4 (2) Fall, 2004, 22-37

———————. "Following the Trail of the Disappearing Data," The Seybold Report, 4 (21) February 15, 2005, 7-14

———————. "No Pain, No Metadata," The Seybold Report, 5 (6) June 22, 2005, 10-12

——————— and Alexander Mikhalevitch. "Heart of Darkness: A Look Inside Aging JPEGs," The Seybold Report. 5 (22) February 22, 2006, 9-12.

———————. "Newspapers Online: From Promise to Practicality," presentation at Web Wise '06 Conference, February 15-17, 2006, Los Angeles.

Thompson, Mark. "To Fix or Not to Fix: Online Corrections Policies Vary Widely," Online Journalism Review, May 28, 2004, http://ojr.org/ojr/workplace/1091056600.php.

Tibbo, Helen R. "Primarily History in America: How U.S. Historians Search for Primary Materials at the Dawn of the Digital Age." *The American Archivist* 66 (1), Spring/Summer 2003, 9-50.