Digitizing California's Newspapers: A Guide and Best-Practices for Institutions Around the Golden State

Created by the Center for Bibliographical Studies and Research, UC Riverside for LSTA Grant 40-7696, August 2011

Newspaper Digitization – Brief Background

Standards and best practices for digitizing historical newspapers, whether from microfilm or print originals, have developed and matured relatively recently. Prior to 2005 newspaper digitization followed a variety of approaches with varying and, more often than not, unsatisfactory results. In some cases, digital services vendors used proprietary software for processing and display, and the resulting digital newspapers were tied to particular content management and presentation systems. Years later, libraries that worked with these vendors have often been unable to migrate their newspapers to different presentation systems, or have had to pay to have the data reformatted. In other cases, libraries lost control of their own content when commercial agencies claimed exclusive rights when digitizing their local papers. In still other cases, file formats were used that are no longer supported, rendering those assets unusable, at least without costly reprocessing. Unfortunately, all these practices continue today. We hope this documentation will help California institutions avoid these pitfalls in the future.

In North America, newspaper digitization standards and best practices took a leap forward in 2005, when the National Endowment for the Humanities (NEH) and the Library of Congress (LC) initiated the National Digital Newspaper Program (NDNP). The NDNP has two distinct but related goals. The first is to create a free online resource of searchable newspapers from every state for the period 1836-1922, available through the Chronicling America website (http://chroniclingamerica.loc.gov/). Newspapers hosted here are true digital facsimiles of the original newspapers, and are full-text searchable. Individual page images are retrieved and search words are highlighted.

The second goal of the NDNP is to create a set of best practices for newspaper digitization. Libraries have always recognized the importance of using open standards to facilitate collaboration and data sharing. Digitization projects are often expensive, and using standards and tested formats are essential to ensuring long-term preservation of and access to the digital assets these projects produce. NDNP specifications adhere to open formats as much as possible (see http://www.loc.gov/ndnp/). Specifically, image files are produced in such standard formats as TIFF, JPEG 2000, and PDF.

When it came to deciding on a metadata format for the NDNP, the Library of Congress joined a growing number of libraries around the world by adopting and developing a set of shared specifications for digitizing newspapers. Commonly called METS/ALTO, these two open XML standards, METS and ALTO, describe the structure of each newspaper issue. METS (Metadata Encoding and Transmission Standard) provides the overall structure of the single newspaper issue. ALTO (Analyzed Layout and Text Object) extends the METS and creates the structure for each page of an issue, embedding the text created from OCR (optical character recognition) into an XML structure that allows for full-text searching and highlights search terms within a page. Today, METS/ALTO has become the internationally accepted foundation for newspaper digitization and is used, for example, by the National Library of Australia (http://trove.nla.gov.au/newspaper), the National Library of Singapore (http://newspapers.nl.sg/), and the National Library of Finland (http://www.digi.lib.helsinki.fi/sanomalehti/).

The California Digital Newspaper Collection

The Center for Bibliographical Studies and Research (CBSR) at the University of California, Riverside established the California Digital Newspaper Collection (CDNC) (http://cdnc.ucr.edu) in 2006 with content submitted to the NDNP and content created with LSTA (Library Services and Technology Act) grants from the California State Library. Over the last 5 years we've continued to add NDNP- and LSTA-funded content, as well as newspapers digitized with funds raised by local institutions and contemporary PDFs collected from newspaper publishers. The CDNC currently contains more than 450,000 pages and many of California's most significant historical newspapers, selected over the years by an advisory board. The project adheres to NDNP specifications, with additional processing to create article–level access. The CDNC uses Veridian Digital Library Software (Veridian) for its search engine and presentation system, and displays both page- and article-segmented data. The project recently became the first North American newspaper portal, and the second in the world that we are aware of, to implement user correction of computer-generated text. With the ongoing addition of local and regional newspapers, the CDNC will continue to serve as the portal to digitized California newspapers.

Best Practices: Microfilm and Newsprint

The following guidelines derive from the CDNC's participation in the NDNP and our work with newspaper digitization programs at institutions around the world.

When selecting newspapers for digitization, one of the first decisions libraries or other cultural heritage institutions face is determining the quality and quantity of available content. Many communities have had several newspapers, some of which were published concurrently; other communities primarily have a long run of one title. Given limited funds, local institutions will have to determine which title or titles, and potentially which years, contain the most valuable content. Papers published before 1923 are in the public domain and can be digitized without restrictions. Libraries intending to digitize papers published after 1922 should get permission from the copyright holder (often a current publisher) or perform due diligence in searching copyright history. Because of possible copyright restrictions, all post-1922 papers presented at the CDNC will only be digitized with page-level access.[1]

Established practice within the NDNP is to digitize from master negative microfilm, in part to achieve economy of scale since scanning microfilm is more efficient and less expensive than scanning original newsprint. The program does, however, allow for digitization from newsprint. Master negatives or newsprint are required because the microfilm reels library patrons use to read and print historical newspapers are often scratched and damaged from that use. Furthermore, NDNP specifications require that institutions create duplicate negative reels from the master negatives and use those duplicates for digitization. The CBSR is less stringent about digitizing from duplicate negatives. Best practice is still to use duplicates in order to protect the master negatives from possible damage during scanning. However, in cases where creating duplicate negatives becomes a substantial expense for a project, the CBSR would consider waiving this requirement for microfilm from the California Newspaper Microfilm Archive (CNMA; see next paragraph) if an institution works with a vendor that has a reliable track record of duplicating film without damaging it.

Locating and confirming availability of the master negative reels, or good quality original newsprint, must be a first step in planning a newspaper digitization project. CBSR manages two projects that can help institutions locate copies for digitization, and our staff is available to help with the process. The first is the California Newspaper Project (http://cnp.ucr.edu), the most comprehensive bibliography and union catalog of newspaper titles and their associated holdings around the state. The second is the CNMA, the largest archive, nearly 100,000 reels, of newspaper master negatives in the state. The film is stored in UC storage facilities and a searchable database for the CNMA (http://cnma.ucr.edu) will be completed by the end of 2012. In addition, ProQuest and Heritage Microfilm store or own negative microfilm for many California newspapers, and the CBSR has relatively up-to-date lists of their holdings. ProQuest told the Center that they would consider making duplicate negatives for scanning for California projects on a title-by-title basis. Heritage has digitized many of the titles in their microfilm vault and made them available at NewspaperArchive.com, for a fee.

---

[1] Providing online access to digitized in-copyright newspapers is still something of a grey area. The CBSR has followed two relevant court cases in determining that it can legally present post-1922 papers with page level access. In "Tasini v. the New York Times" the Supreme Court ruled that the NYT could not license the works of free-lance journalists for inclusion in electronic databases. In a follow-up case, "Greenberg v. National Geographic," the Eleventh Circuit Appeals Court ruled that National Geographic's effort to digitize its archive was consistent with "Tasini" because National Geographic had "maintained the context of its prior collective works." We assume that ruling applies to newspapers also, as long as the copyright holder has granted permission or has been determined to no longer exist.

Once an institution has located master negatives or newsprint, they must evaluate the condition and completeness of the film (or newsprint) to be scanned to make sure they are acceptable. The NDNP guidelines recommend that microfilm images be no more than twenty times smaller than the newsprint original (20X reduction) and that there be minimal variations in density within images and between exposures. The CBSR, however, has digitized many titles that fall below these specifications with satisfactory results. Digital services vendors can evaluate film. The CBSR can also assist with this step. See Appendix A: "Selection of Newspapers for Digitization" for more information.

The more a library knows about the newspaper issues on the film, the better. The CBSR recommends inventorying each reel of film that is to be scanned and collecting at least a minimal amount of metadata for each issue on the reel, in the order the issues appear. This information can be entered in an Excel spreadsheet or database program. Many vendors use this information in the digitization process. Also, with this information in hand, the library can compare the scanned images with the reel inventory as a quality assurance step. Two appendices are attached to this report that provide additional information. Appendix B: "Summary of Specifications used for the California Digital Newspaper Collection" outlines the recommended metadata and format specifications for newspaper digitization. Appendix C: "Sample Digital Newspaper Metadata Template" provides an example of the type of data that might be collected.

Recording additional information such as pages per issue, pages out of order, special edition information, missing pages and duplicate pages, and changes to page dimensions allows for better processing and better quality control, but requires much additional time and effort. Digitization costs are usually based on number of pages or number of reels to be digitized, so knowing this information upfront is useful. The orientation of the page images on the film should also be noted: 1A means 1 page per frame with the side of the page parallel to the film edge; 2B means 2 pages per frame with the top and bottom page edges parallel to the film edge. A full reel typically contains 650 pages at 1A or 1300 pages at 2B, but reels are often much smaller than that. The reduction of the pages on the film, such as 15X (15:1), is needed for scanning. If the reduction ratio is not noted on a reel information target, the dimensions of the original newsprint that was filmed can be used by the scanning operator to calculate reduction ratios. Other data that should be collected are: source of the master negatives, library responsible for the project, cataloging title from the 245 or 130 field of a WorldCat record, place of publication from the 260 field, and the LCCN from the 010 field; the CBSR can help collect this data.

Best Practices: Metadata and Image Files

Several digital service vendors understand and can meet NDNP specifications. Apex CoVantage, HTC Global Services, Northern Micrographics, Content Conversion Specialists (CCS), Backstage Library Works, and iArchives are all working with

NDNP state projects. The CBSR has worked with Backstage Library Works and CCS since 2005. It should be noted, however, that NDNP requires only page-level metadata, not the article-level used at the CDNC and by libraries around the world, and not all these vendors will be familiar with creating article-level metadata.

The CBSR strongly encourages institutions to follow NDNP metadata standards, particularly the creation and use of METS/ALTO and the associated image files described above and in Appendix B. Doing so produces archival-quality digital assets and provides the best assurance possible that those assets will be usable across platforms and into the future. Some institutions outside the U.S. no longer produce and store TIFF images, instead relying on the JP2s. At this time the CBSR still considers TIFF to be a better preservation file format and the costs associated with creating and storing the TIFFs, which are not used for display purposes, are minimal.

The costs associated with producing article-level metadata, rather than page-level, are still substantial, however, and institutions will have to decide whether those costs are justified. The CBSR has followed international, rather than NDNP, best-practice by producing article-level metadata and believes doing so produces more refined and better search and retrieval. Institutions can compare article- to page-level data for themselves by searching *The San Francisco Call* in both the CDNC, where it is available at the article level, and Chronicling America, where it is page level.

In contrast to the growing numbers of vendors that can produce NDNP-spec metadata, there are limited software options for hosting and displaying that metadata. The CDNC uses Veridian, which is developed by DL Consulting. The Library of Congress recently released their Chronicling America software open-source. The University of Oregon was one of the first institutions to adopt it (http://oregonnews.uoregon.edu/). Finally, there is OCLC's CONTENTdm, which both the Santa Monica Public Library (http://digital.smpl.org/index.php) and the Marin County Library (http://contentdm.marinlibrary.org/index.php) use to host their newspapers. CONTENTdm has the advantage of being able to host a variety of digital assets, not just newspapers or METS/ALTO. The CBSR cautions institutions using CONTENTdm, however, to make sure their version preserves the METS/ALTO metadata and associated files (TIFF, JP2, and PDF), rather than converts them into a proprietary format that will be unusable or difficult to convert for use in another system. (Alternatively, vendors digitizing newspapers for CONTENTdm can output two sets of files – one for CONTENTdm and one with METS/ALTO XML, TIFF, JP2 and PDF files, at minimal or no extra cost. Institutions should archive that METS/ALTO and associated files for use in other systems in the future such as the CDNC.

Another option for hosting and displaying NDNP-spec METS/ALTO and associated files is to present only the PDFs. The Whittier Public and Torrance Public Libraries had some of their historical newspapers digitized as searchable PDFs, but without the other associated metadata. See: http://digi.whittierlibrary.org/wpl/index.html, and http://www.torranceca.gov/libraryarchive. It is important to note, however, that the PDF was designed to be a cross-platform presentation file format. It was never designed for presenting search and retrieval results, and using PDF for those purposes often creates an awkward user experience and limits metadata enrichment, such as user text correction. For example, systems that rely on PDFs can guide a user to a PDF containing terms he or she searched for, but the user must then download the PDF and perform the search again in the file to find the terms on the page. The CBSR strongly advises institutions against producing only PDFs when digitizing their historical newspapers, particularly when search and retrieval is important or when the papers will be available over the internet. If you choose to produce PDFs without the associated TIFF, JP2 and METS/ALTO XML files specified by NDNP, you most likely will pay to reprocess those PDF files in the future.

Before undertaking a newspaper digitization project or committing to a vendor, institutions should find out from vendors what specifications they follow, what output they create, and what presentation systems can be used with that output. They should request a sample before proceeding with a full project and be able to evaluate that sample carefully or identify a third party that can. Finally, institutions should also make plans for backing up their digital assets to ensure long-term availability. It is important to keep in mind that the amount of data associated with newspaper digitization often exceeds what institutions are accustomed to with other digitization projects, and server and storage requirements must be planned accordingly.

California Newspaper Digitization – the Role of the CBSR

The CBSR strongly believes that researchers are best served by having one portal to digitized California newspapers, where they can browse and search across all available titles and dates. Institutions that want to host all or part of their newspaper metadata locally can and should do so, but those local archives should duplicate, not replace, a central California repository. The California Digital Newspaper Collection is the logical resource for this role. To date the CBSR has not charged to host content created by local institutions. In the future, however, we expect to reevaluate this and to charge institutions a fee to offset the hardware, software and staff costs associated with hosting their data. This charge will be substantially less than an institution would pay to license and maintain software on its own, and will most likely be based on the amount of data an institution contributes to the CDNC.

California institutions that contribute data to the CDNC will have to produce output that meets CDNC specifications, as outlined in this document and the appendices. There are at least two ways institutions might go about this. They could contract with the CBSR to have the Center assume the digitization work. The CBSR has a license for docWORKS, CCS's digitization software, which is installed on local servers. CCS employees remotely process the data on these servers while CBSR staff monitor workflow and perform quality review. On the other hand, institutions might want to work with vendors directly and submit the data produced to the CDNC. The CBSR is willing and eager to help institutions identify, evaluate and work with vendors, and review sample data before a project starts in earnest.

We hope our years of experience digitizing newspapers will be of help to other California institutions as they consider whether and how to digitize their own local content. The CBSR looks forward to partnering with other institutions around the state to increase the number of digitized California newspapers, to continue to add to the CDNC, and to make more of our state's newspaper content freely available to its citizens.