Appendix B: Summary of Specifications used for the California Digital Newspaper Collection

Guidelines for Digital Files

Ideally, a standards-based target film strip should be scanned at the start of each session, to monitor scanning equipment performance.

The archival master is the TIFF; the JP2 and PDF service copies used within the presentation system are derived from the TIFF.

Archival Master	**Service copy**	**Copy for download**	**Optical Character Recognition (OCR)**
Uncompressed TIFF v. 6.0	*JPEG2000*	*PDF*	*Full text OCR*
8-bit Grayscale	JPEG2000 image for each page image.	PDF Image with searchable Hidden Text for each page image.	One OCR text file for each page
300-400 ppi (depends on reduction ratio of film)	The JPEG2000 consists of 6 decomposition levels, and 25 quality levels.	The page image will be grayscale, downsampled to 150 ppi and encoded using a medium JPEG quality setting.	Text in UTF-8 character set.
De-skew images with a skew of more than 3 degrees. No other image processing on archival master (e.g., no sharpening).	JPEG2000 compression is at 8:1 (that is, the JPEG2000 will be 1/8 the size of the master TIFF).		OCR text ordered column-by-column in a natural reading order (left-to-right for English and European language papers).
1 page per image (pages filmed 2-up should be split into 2 files)			OCR conforms to the ALTO XML schema.
Image cropped to page edge so that full page is viewed.			OCR text file with "bounding-box" coordinate data at the word level (this facilitates display of highlighted search terms on page).

Metadata guidelines

Descriptive Metadata for Each Issue

- Title (use exact title from CONSER record, if available - e.g., MARC field 130|a or 245|a). *[Note: if scanning microfilm, check film to note if other titles are present on the reel.]*
- LCCN (MARC field 010)
- Date of issue (should be entered in ISO-8601 format, e.g., YYYY-MM-DD). If desired, there can be an additional field to write out a conventional date display, such as "January 3, 1912."

Optional information to collect for each issue:

- Volume number
- Issue number
- Edition label if present (e.g., EXTRA, Evening Edition, Special Edition, etc.)

Additional information about the source:

- Format of source material (microfilm, microfiche, print, electronic [e.g. PDF])
- Reduction ratio used in creating microfilm, or, if not available, the actual dimensions of the print newspaper
- Microfilm position (1A, 2B, etc.)
- Institutional source of the original microfilm, print, or electronic copy
- Institution responsible for digitization (may be the same as the source)
- Title place of publication (e.g., MARC 260|a).

Descriptive metadata can be encoded in any appropriate XML standard, such as MODS, Dublin Core, etc.

Technical Metadata (most can be automatically captured by scanning software)

- Format mimetype (e.g., image/tiff, image/jpeg)
- Mode (grayscale)
- Bits per pixel (8)
- Image width and length in pixels
- X and Y sampling frequency, in pixels per inch (ppi). For the CDNC, this would be 300-400 ppi.
- File size (bytes)
- Unique identifier for each image (assigned by institution)
- Checksum for each image (MD5, SHA-1, etc.)

Technical metadata is typically encoded in the MIX standard (NISO Tehnical Metadata for Digital Images).

Structural Metadata

For the California Digital Newspaper Collection (CDNC), every page of each newspaper issue is produced in the following formats: TIFF, JPEG 2000, PDF, ALTO XML (OCR). In addition, each issue is "wrapped" in structural metadata following the METS standard (METS = Metadata Encoding Transmission Standard). The METS/ALTO "package" incorporates descriptive and technical

metadata, provides structural integrity to every issue, and enables full-text searching with highlighted display of search terms on the image pages.

These guidelines are adapted from the specifications created by the Library of Congress for the National Digital Newspaper Program (NDNP). For more detailed information and current program specifications, see http://www.loc.gov/ndnp/pdf/NDNP_201113TechNotes.pdf

To learn more about the METS standard, see http://www.loc.gov/standards/mets/

To learn more about the ALTO standard, see http://www.loc.gov/standards/alto/

To learn more about the MODS standard, see http://www.loc.gov/standards/mods/

To learn more about the MIX standard, see http://www.loc.gov/standards/mix/