

EXHIBIT A.1
**DESCRIPTION OF GOODS AND SERVICES FOR TO PROVIDE NDNP-COMPLIANT DIGITIZATION,
OCR & XML METADATA SERVICES FOR NEWSPAPERS ON MICROFILM**

1. The Client Agency shall:

- a. Provide duplicate negative microfilm of newspapers that are free of copyright restrictions or is being used with permission
- b. Provide the Contractor with detailed information about every title and reel, and collation of each issue and page on each reel, to be used to create the NDNP-compliant metadata
- c. Provide portable hard drives for the digital files and pay for shipping film and hard drives from the Client Agency to the Contractor

2. The Contractor shall:

- a. Provide image and metadata files that comply with the National Digital Newspaper Program (NDNP) specifications (validated using DVV software from Library of Congress) and that can be loaded into the Library of Congress database
- b. Provide these files to meet the current NDNP specifications and, as time goes on, to meet the latest update to the NDNP specifications
- c. The files should also be re-useable in other systems or for other purposes
- d. Provide a mainland U.S. shipping address for receipt of microfilm and portable hard drives from the Client Agency

3. NDNP Microfilm Processing Specifications:

- a. Provide validated master digital page image format=Grayscale, scanned for maximum resolution possible between 300-400 dpi, relative to the original material, uncompressed TIFF 6.0
- b. Validated derivative digital page image format=JPEG2000 (JP2) using specified compressions options
- c. Validated OCR text file with bounding-box coordinates for associated bounding boxes for words= text file per page
- d. Validated PDF Image with Hidden Text, i.e., with text and image correlated = 1 PDF per page
- e. Structural metadata to relate pages to title, date, and edition, sequence pages within issue or section; and to identify image and OCR files
- f. Validated metadata using METS in accordance with guidelines (all digital assets in a METS object structure, Metadata Encoded Transmission Schema) according to an XML Batch template structure
- g. Deliverable files should organize the page images and related files for each newspaper title in a hierarchical directory structure sufficient for identification of the individual digital assets from the metadata provided
- h. Note: The four digital files associated directly with a newspaper page (TIF, JP2, PDF, and OCR) are expected to use the same file identifiers with distinct file extensions

EXHIBIT A.1
**DESCRIPTION OF GOODS AND SERVICES FOR TO PROVIDE NDNP-COMPLIANT DIGITIZATION,
OCR & XML METADATA SERVICES FOR NEWSPAPERS ON MICROFILM**

4. Digital Page Image Specifications:

- a. Images shall be 8-bit grayscale scanned from 2N silver negative microfilm at the maximum resolution possible, between 300 and 400 dpi, relative to the physical dimensions of the original newspaper, rather than the microfilm
- b. Each JPEG2000 will incorporate appropriate XMP metadata, will be 6 decomposition levels, and 25 quality levels, compression shall be 8:1
- c. A standards-based target film strip must be scanned at the start of each session. Target for NDNP project to be provided by the Client Agency. Target test images shall be delivered with page images
- d. Exclude the other informational targets, such as START or JANUARY 1902 or ISSUE(S) MISSING, that typically appear interfiled with the newspaper pages on microfilm. These targets do not need to be scanned, or, if scanned automatically, do not need to be processed and delivered
- e. Newspapers microfilmed two sheets per frame should be split into two separate image files and must be assigned appropriate corresponding metadata
- f. Images with more than 3 degrees of skew should be deskewed
- g. Page image files should be cropped to the page edge (not to the text block boundaries), retaining the actual edge and up to ¼ inch beyond
- h. All operations that change the image dimensions, spatial resolution, or orientation (e.g., cropping, deskewing) shall be made to the TIFF before OCR processing
- i. The grayscale master TIFF files must have the same dimensions, spatial resolution, cropping and deskewing as the images used for OCR, but no other enhancements (e.g. sharpening, contrast enhancement, etc.) may be used in the OCR-creation process

5. Uncorrected OCR Text File Specifications:

- a. One OCR text file per page image shall be provided. (Discrete files produced for each page)
- b. Each OCR text file name must correspond to the page image it represents. Text must be encoded in the UTF-8 character set
- c. No graphic elements shall be saved with the OCR text
- d. The OCR text shall be ordered column-by-column (i.e. in a natural reading order) with bounding-box coordinate data at the word level
- e. The OCR text shall be encoded using the ALTO (Analyzed Layout and Text Object) schema, Version 1-4 or greater
- f. Use of the SourceImageInformation\fileName element is required; and should be included in the path if path contains useful information
- g. If the OCR process generates coordinates for zones, the segmentation data shall be removed from the METS/ALTO object prior to delivery to Library of Congress
- h. All page images shall be accompanied by an ALTO XML file containing recognized text
- i. If possible, provide these additional elements for OCR files: Confidence level data at the page, line, character, and/or word level; point size and font data at the character or word level

EXHIBIT A.1
**DESCRIPTION OF GOODS AND SERVICES FOR TO PROVIDE NDNP-COMPLIANT DIGITIZATION,
OCR & XML METADATA SERVICES FOR NEWSPAPERS ON MICROFILM**

6. Image PDF with Hidden Text Specifications:

- a. A PDF Image file with Hidden Text shall be provided for each page image
- b. Each searchable PDF file name corresponds to the page image it represents
- c. The PDF files should incorporate appropriate XMP metadata
- d. Page image must be grayscale, downsampled to 150dpi and encoded using a medium JPEG quality setting
- e. Only the 14 standard Type1 fonts may be used. These fonts may not be embedded
- f. Text streams must be Flate encoded
- g. The page image may not contain any bookmarks, links, named destinations, comments, forms, Javascript actions, external cross references, alternate images, embedded thumbnails, annotations, or private data
- h. The PDF may not be tagged; must open to Fit Page sizing; open to single page layout; will open with neither document outline nor thumbnail images available; will open with the tool bar, menu bar, and user interface elements visible
- i. The PDF shall not open centered in the screen; will not be encrypted, digitally signed, or have any security
- j. The PDF shall be linearized (also known as “Fast Web View” and shall be compatible with Acrobat 5.0 or later. Except where conflicting with any of the other requirements of this profile, the PDF shall conform to PDF/A (ISO 19005-1)

7. Specifications for Metadata for all Image Files:

- a. Structural metadata for pages, issues, editions, and titles shall be organized by date to support a chronologically-based browsing interface
- b. Headers for all image deliverables (TIFF, JPEG2000, and PDF) should incorporate tagged metadata relating to the creation of the images
- c. Technical metadata describing the quality characteristics of the film used for digitization must be encoded in a XML METS object
- d. Target test images used in scanning shall be described in the reel metadata

8. Validation Specifications:

- a. All digital objects shall conform and validate to NDNP technical specifications
- b. Assets should be ready to be delivered to Library of Congress in a prescribed directory structure conforming to the “BagIt” specification, a hierarchical package format for transferring digital content