# SAMPLE
# Project Description

## 1. Project Title

NDNP newspaper pages

## 2. Description

The Library of Congress (LC) and the National Endowment for the Humanities are jointly developing the National Digital Newspaper Program (NDNP). Ultimately, over a period of approximately 20 years, NDNP will establish a national, digital resource of historically significant newspapers published between 1836 and 1922, digitized by consortia representing all states and U.S. territories. This searchable database will be permanently maintained at the Library of Congress and be freely accessible via the Internet. In order to create test data to facilitate development of the system architecture, LC will scan microfilm from its own newspaper holdings.

## 3. Objective

Deliver digital files for each original object supplied, including TIFF, PDF, JPEG2000 and optical character recognition (OCR) output files. Additionally, metadata at the issue/edition and page level will also be delivered.

## 4. Scope

Between x and y newspaper pages, scanned from 35mm negative microfilm.

## 5. Task Description

5.1 Overview

The contractor will create sets of digital images and metadata from supplied newspaper microfilm reels, to conform to the technical guidelines in this document and its appendices. Data and derivative files will be produced for all images: page images, informational target images and technical target images.

**Requirements for Scanning**

1. The spatial resolution shall be 400 pixels-per-inch (ppi) relative to the original newspaper. If that is technically impossible, due to high reduction ratio of particular reels, the spatial resolution for those reels shall be 300 pixels-per-inch (ppi) relative to the original newspaper.
2. 8-bit grayscale
3. TIFF 6.0 uncompressed.

4. Two-up film should be split so that there is one page image per file.
5. De-skew images with a skew of greater than 3 degrees.
6. Crop to edge of page**.**
7. Capture microfilm target frames. These image files to be identified as "targets" in metadata; will not be used for display.
8. Capture scanning resolution targets (supplied by the Library of Congress) at the start of each session, to monitor scan quality. These target images should be delivered with microfilm targets and page images.
9. TIFF headers shall incorporate data as described in attached file TIFFSpecs.pdf.

Note: the grayscale images delivered must have exactly the same dimensions, spatial resolution, skew, and cropping as the images used for OCR.

**Required elements for OCR files:**

1. One OCR text file per page image. (Discrete files should be produced for each page, rather than for a multi-page issue or entire title).
2. Each OCR text file name corresponds to the TIFF image it represents.
3. Font point size data at the character or word level is required.
4. OCR will be delivered in the format described in attached OCRSpecs.pdf and its associated alto_1-1-041.xsd.

Image sharpening software may be used in the OCR process, if it improves the accuracy of the OCR output. Additional sharpened versions of the TIFF files created as part of this process will **not** be delivered to LC.

**Required elements for PDF files**

1. PDFs should be version 1.4 or 1.5.
2. Each searchable PDF file name corresponds to the TIFF image it represents.
3. Files will have characteristics described in attached PDFSpecs.pdf.

**Required elements for JPEG 2000 files**

1. Files will be JPEG2000, Part 1, (or ISO-15444).
2. Each file name corresponds to the TIFF image it represents.
3. JPEG2000 files will conform to the specification in attached JPEG2kSpecs.pdf
4. As JPEG2000 is an emerging standard, the Library of Congress reserves the right to change the settings numbered 3-16 in JPEG2kSpecs.pdf.

**Required Metadata**

Metadata shall be created for each Issue/Edition and reel, conformant to the definitions in the attached metadata_dictionary.doc. There will be one XML record for each reel, and one XML record for each Issue/Edition on that reel. A single batch file (example batchTemplate.xml attached) will point to each Issue/Edition and reel record on the delivery media. Example METS templates are attached for Issue/Editions (issueTemplate.xml) and reels (reelTemplate.xml). Note: an Issue/Edition record contains both information about that Issue/Edition and pages

within that Issue/Edition. A reel record contains information about the reel and all target images on that reel. Issue/Edition records will also be created for issues indicated as missing by a target on the film.

Information about each reel's density and resolution will be provided by the Library of Congress for incorporation into the reel's XML file. This information will be provided in a Microsoft Access Database.

The PREMIS and MIX metadata referenced in the templates is **not** required at this time.

**File and Directory Structure on Delivery Media**

The name of the batch on the hard drive should adhere to the following standard: "batch_[MARCOrgCode]_[unique word for awardee]" – i.e., "batch_dlc_alpha". This name should appear as the folder name for the batch on the hard drive, as well as in the batch.xml file as the "name" attribute, as shown here:

```
<batch xmlns:ndnp="http://www.loc.gov/ndnp"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns="http://www.loc.gov/ndnp" name="batch_dlc_alpha">
```

In the root directory of the portable hard drive will be the single batch file pointing to each Issue/Edition and reel record on the delivery media. Example:
batch.xml

Each title will be contained in its own directory, name matching the title's LCCN.
Example:
/sn83045433/

Within each title directory, subdirectories will be created for each scanned reel, with names matching the barcode on each reel. Example:
/sn83045433/0010049324a/

Within each reel subdirectory will be one metadata file for the reel, named after its barcode number, the target images for that reel, and a subdirectory for each newspaper issue/edition, named after its date and edition order (typically 01).
Example reel metadata file:
/sn83045433/0010049324a/0010049324a.xml
Example subdirectory for the only edition from January 24, 1905:
/sn83045433/0010049324a/1905012401/
The metadata for that edition would be in:
/sn83045433/0010049324a/1905012401/1905012401.xml

In each issue/edition subdirectory will be the metadata file for that issue, and all TIFF, JP2, PDF and OCR files for the pages in that issue/edition.

If the issue/edition is one noted as missing from the guide to contents or other filmed targets, the proper issue/edition subdirectory will be created and contain the metadata file, but **no** TIFF, JP2, PDF or OCR files. Example for "issue missing November 29, 1905":
/sn83045433/0010049324a/1905112901/1905112901.xml

Files shall be named in four digit, one-up manner, according to the order of appearance on the scanned reel, whether the image is a target or page. For example, a reel might contain in order, a technical target (whose TIFF would be named):
/sn83045433/0010049324a/0001.tif
a title target:
/sn83045433/0010049324a/0002.tif
a guide to contents:
/sn83045433/0010049324a/0003.tif
and newspaper pages:
/sn83045433/0010049324a/1905012401/0004.tif
/sn83045433/0010049324a/1905012401/0005.tif
/sn83045433/0010049324a/1905012401/0006.tif
etcetera.

Each JP2, PDF and OCR file will have the same prefix as the TIFF file from which it was derived, and will be located in the same directory as its matching TIFF.
Example JP2, PDF and OCR files:
/sn83045433/0010049324a/1905012401/0004.jp2
/sn83045433/0010049324a/1905012401/0004.pdf
/sn83045433/0010049324a/1905012401/0004.xml

The scanning resolution technical target reel (the one to be run at start of each scanning session) is not associated with a single LCCN, so its directory structure will be delivered as:
/the-date-it-is-imaged/three-digit-one-up-number-in-case-more-than-one-per-day/barcode/
example for reel 00100493147 imaged on March 11, 2005
metadata file:
/20050311/001/00100493147/00100493147.xml
image files:
/20050311/001/00100493147/0001.tif
/20050311/001/00100493147/0002.tif
etc.

**6. Government Furnished Equipment (GFE)/ Government Furnished Information (GFI)**

The Library of Congress will furnish 500 Gigabyte magnetic storage media required for image delivery.

**7. Deliverables/Delivery Schedule**

7.1 Delivery Media

The Library of Congress will furnish 500 Gigabyte magnetic storage media required for image delivery. Contractor should keep a backup copy of files until they have been notified by LC that they can be deleted.

7.2 Delivery Schedule

Frequency of delivered batches will be determined at project startup, on a schedule mutually agreed to by the contractor and the Library.

**8. Inspection and Acceptance Criteria**

Images will be inspected and accepted in accordance with the requirements listed in Section E of the Statement of Work.

…