

Guidelines for preparing your historical newspapers for online access



If your library, historical society or other cultural heritage organization is planning a newspaper digitization project, you'll have many choices to make. Based on our years of experience, we would like to help you to understand these choices and offer some guidance.

The challenge of digitizing newspapers

Newspapers can be more challenging to digitize because of their larger format and, in many cases, their fragile condition. They also add an additional layer of complexity, in comparison to other primary source materials, due to their layout, the article format and the small fonts.

How do you decide?

As part of your newspaper digitization project, you'll have to decide the scanning and format specifications for your project, including:

- Desired bit depth (bitonal, grayscale or 24-bit color)
- Desired resolution (400 dpi, etc.)
- Image processing options (deskewing, etc.)
- Desired file format (TIFF, etc.)

Currently, there are not any clearly defined standards that are universally accepted to ensure the quality and longevity of digitized newspapers. However, many organizations look to the National Digital Newspaper Program (NDNP) for guidance.

What is the NDNP?

The NDNP or National Digital Newspaper Program is a partnership between the National Endowment for the Humanities (NEH) and the Library of Congress (LC). The goal of the NDNP program is to build a national, digital resource of historically significant newspapers from all the states and U.S. territories published between 1836 and 1922. Additionally, the NEH is awarding grants to nonprofit organizations to help them convert newspapers into digital files.

(Source: www.loc.gov/ndnp).

NDNP specifications

The NDNP-developed specifications are based on today's understanding of digital preservation needs. By following these specifications, the NDNP expects to create digital formats with a high probability of sustainability, as well as a set of specifications that considers the issues of cost and maintenance.

The NDNP technical specifications also were designed to provide page images that support OCR (optical character recognition) software, and to facilitate the link from digitization to online presentation via the Web.

Your project and the NDNP guidelines

Whether or not your organization pursues an NEH grant, you can still look to the NDNP program for scanning and metadata guidelines when starting your own newspaper digitization project. However, following every NDNP specification may not be necessary for your digitization project. We can work with you to tailor a solution that meets your needs.

While our approach considers the NDNP guidelines, it is specific to each digitization project. We can work with you to determine what your project's needs are in terms of resolution, processing options, file format, etc., and deliver an appropriate solution.

Please see the chart on page 2 of this document for an outline of the NDNP technical overview specifications. It provides a high-level view of the basic scanning, format and OCR guidelines.

NDNP scanning, format and OCR guidelines

This chart is an outline of the NDNP technical overview specifications. It provides a high-level view of the basic scanning, format and OCR guidelines.

| Archival Master: TIFF | Production Master: JPEG2000 | Derivative: PDF | OCR text: ALTO |
|---|---|--|--|
| <ul style="list-style-type: none"> ▪ Conforms with TIFF 6.0 ▪ 8-bit grayscale ▪ 400 dpi preferred ▪ Uncompressed ▪ Only deskewing should be applied ▪ Cropped to page edge ▪ Additional TIFF tags required | <ul style="list-style-type: none"> ▪ Conforms with JPEG2000, Part 1 (.jp2) ▪ Use 9-7 irreversible (lossy) filter ▪ Compressed to 1/8 of the TIFF or 1 bit/pixel ▪ Tiling, but no precincts ▪ RDF/Dublin Core metadata in XML box | <ul style="list-style-type: none"> ▪ Compatible with Acrobat 5.0 (PDF 1.4) ▪ Image with text behind ▪ Image will be a grayscale, 150dpi ▪ JPEG, using a medium (or 40) quality setting ▪ XMP/RDF/Dublin Core metadata | <ul style="list-style-type: none"> ▪ Conforms with ALTO (Analyzed Layout and Text Object) schema ▪ ALTO is product of EU-funded METAe project ▪ Mapping of OCR'ed text to image coordinates |

Helpful definitions

While working on your newspaper digitization project, you'll come across a lot of acronyms. Here are the more common ones and their definitions.

ASCII (American Standard Code for Information Interchange) - An ASCII or tab-delimited text file is one in which each byte represents one character according to the ASCII code and it will contain only text with no special formatting (outside of line breaks and tabs). When documents are scanned, OCR software analyzes the image and converts everything into ASCII or plain text.

DPI (dots per inch) - A measure of printing resolution. The higher the dpi, then the higher the resolution and the clearer and more detailed the output.

Dublin Core® metadata - An element set that provides a simple and standardized set of conventions for describing things online in ways that make them easier to find. Dublin Core is widely used to describe digital materials.

JPEG2000 - An image compression standard used for large digital objects, such as newspapers, maps, etc. The file extension is JP2.

METS/ALTO - METS is a metadata encoding and transmissions standard. An XML document format for encoding metadata is necessary for both management of digital library objects within a repository and exchange of such objects between repositories. ALTO (Analyzed Layout and Text Object) is an extension schema to METS, describing the layout and content of an item. It also includes the OCR results and word coordinates that will be used in search term highlighting.

OCR (Optical Character Recognition) - The ability of computer systems to translate images of typewritten text (usually captured by a scanner) into machine-editable text.

PDF (Portable Document Format) - A file format created by Adobe Systems that allows electronic documents to look exactly like the original pages of the document. It allows documents to be shared regardless of the software platform, the original application, or the availability of specific fonts using free Adobe Reader software.

TIFF (tagged image file format) - A file format mainly used for storing images, including photographs and art. This is a nonproprietary format of choice for the digital master files in your collection. From the TIFF the smaller Web access files get created.

CONTENTdm and the OCLC Preservation Service Centers
**Guidelines for preparing your historical newspapers
for online access** (page 3 of 3)



OCLC and CONTENTdm offer complete digital newspaper solutions

We can work with you to make sure your newspaper collections are digitized and available on the Web through CONTENTdm® more efficiently and cost-effectively.

The solution we deliver can meet the NDNP specifications for you and even go beyond them (for example, by providing article segmentation), or we can tailor a solution to meet your particular needs.

OCLC Preservation Service Centers. At OCLC® Preservation Service Centers, we use the *docWORKS* Newspaper Edition software to help with the newspaper digitization process. Developed by CCS (Content Conversion Specialists), *docWORKS* is an intelligent software application that was developed based on NDNP specifications. It meets current NDNP requirements and will be updated regularly to meet future specifications.

We can create high-quality, digital images from preservation microfilm (or originals). All image processing and content conversion is integrated with the *docWORKS* tool. The converted newspapers are exported in both METS/ALTO and PDF formats and then can be transferred to a variety of content management systems, including CONTENTdm and others.

Since the early 1990s, OCLC Preservation Service Centers staff members have been involved in newspaper digitization projects. And since 2005, we have processed newspapers using NDNP specifications. To date, we have scanned and processed more than 2 million newspaper pages, including student, daily and weekly newspapers. This experience, combined with our state-of-the-art facilities, has helped us become a leading provider of innovative newspaper digitization solutions.

CONTENTdm. CONTENTdm offers a range of solutions for your digitized newspapers, from a basic “do-it-yourself” solution to an outsourced, complete solution. We will work with you to select what is most appropriate for your institution, the size of your project, the digital file format you prefer and your budget.

We offer a complete standards-based solution for your digital newspaper projects. Your converted newspapers’ METS/ALTO, PDF or CONTENTdm OCR Extension formats can be loaded into the CONTENTdm software. Your digital newspaper collections then can be imported to your own server or a hosted server.

With CONTENTdm, you have full control over your digital newspapers, their descriptions, metadata, access and display. And users will be able to search and access your newspapers using any Web browser.

For newspapers, CONTENTdm also supports:

- Full-text searching, with highlighting of the search word(s)
- JPEG2000 for large format pages
- Full-article segment highlighting and extraction
- Use of the electronic newspaper article clipper.

In addition to newspapers, CONTENTdm lets you showcase—and lets your users search across—*all* of your digital collections, from photos and audio/video files to documents and newspapers.

For more information

To learn more about OCLC’s newspaper digitization services, please contact us at digitalcollections@oclc.org

To learn more about CONTENTdm’s newspaper solutions, please contact us at contentdm@oclc.org or call 1-800-848-5878, ext. 4301.