Guidance for Digitally Imaging Content for Submission to GovInfo

November 06, 2023

The following guidance for digital reformatting is based on current best practices. The results will provide the best image capture for the range of information products to be ingested into GovInfo.

The guiding principle for digitization rests in GPO's definition of preservation copy of record for digital content:

> The preservation copy of record for digital content is the preservation master file stored in a trustworthy repository. Derivatives of the preservation master copy are made available for access. *The digital copy of record should be produced to specifications that will allow the creation of a printed facsimile version, should one be needed.*

PRESCRIBED TECHNICAL SPECIFICATIONS

Specifications for preservation master, access derivative files, and metadata are described as "Preferred" and "Acceptable." The Preferred specifications follow the latest best practices and guidance for digital preservation and support GPO's operation of GovInfo as a digital repository. The Acceptable specifications are given to accommodate a variety of stakeholders and their technical capability. GPO may convert TIFF master files to JPEG2000 to reduce preservation storage requirements for content.

Digitize publications using the preferred JPEG 2000, or the acceptable TIFF master file format following the prescribed sampling rate of 300 to 600 ppi to accurately capture the original content. Each page scanned will result in a separate digital file in the prescribed master format. An access file in PDF format will be derived from the master image files. Optical character recognition software will be run to embed machine readable text into the PDF.

	PREFERRED	ACCEPTABLE
MASTER FILE	Uncompressed JPEG 2000, 300-600 ppi, 24 bit RGB color conforming tothe ISO/IEC 15444-1 standard for JPEG 2000	Compressed JPEG 2000 files at 4:1 ratio; Uncompressed TIFF 6.0 300-600 ppi, 24 bit RGB color
ACCESS FILE	PDF/A 2-b with embedded Optical Character Recognition	PDF/A with embedded Optical Character Recognition
TECHNICAL METADATA	MIX XML*	
BIBLIOGRAPHIC METADATA	MARC XML	MARC XML

* NISO z39.87 defines a set of metadata elements for raster images to enable users to develop, exchange, and interpret digital image files. These elements, such as information about the compression, color profile, resolution, scanner or digital camera make and model, can be recorded and preserved as technical information for still images. When possible, this information should be recorded as a set of XML elements conforming to the NISO metadata for images in xml Schema (MIX).

OTHER SPECIFICATIONS

SKEW

No software de-skew should be used. The skew variation should be +/- 1 degree at the vertical axis at the left side of the page.

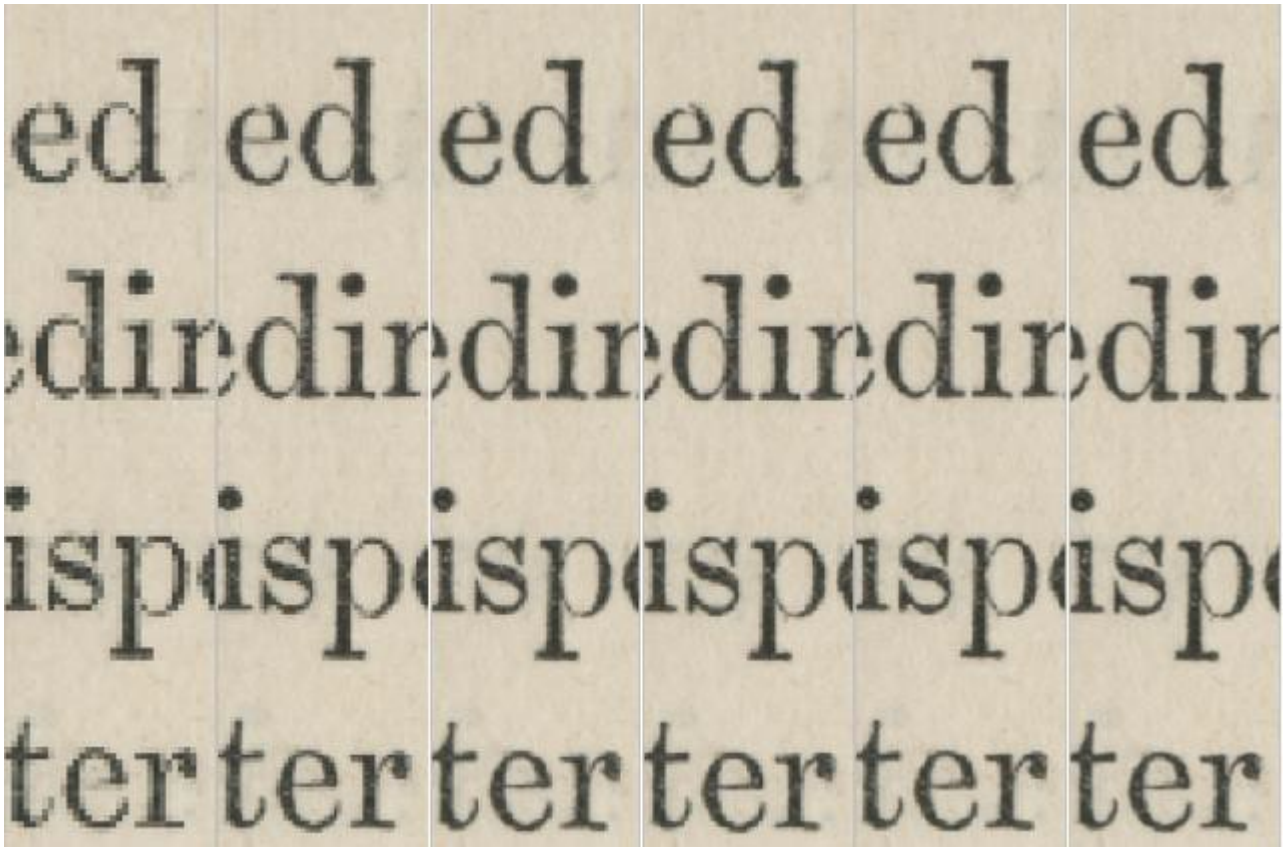
GUTTER MARGIN

If publications are digitized from bound volumes, sufficient gutter margin has to exist to avoid page curvature that may interfere with image clarity, text readability, and OCR accuracy.

PIXELS PER INCH

The sample page images below were digitized in a range from 200 to 700 pixels per inch (ppi). Note the image on the far left, it shows a sampling rate of 200 ppi is not high enough to produce a clear image. Four hundred ppi is the sampling rate where the image comes completely into focus and would be the ideal image sampling rate for this page. Some Federal Government publications use 6 point type for text in page headers, footers, or in tables. A page with type elements this small should be digitized at a minimum of 400 ppi to resolve the smaller type with clarity. Test different ppi to be sure all text is readable. Higher pixel rates, especially beyond 600 ppi, do not create a clearer image, just larger files due to the number of pixels per inch captured.

Examples of text at 200-700 pixels per inch (ppi)



200 300 400 500 600 700

DIGITIZATION BEST PRACTICES

- The digitization process will preserve the page image look and feel of the original source publication, preserving the original cover, blank pages, table of contents, pages and indexes in their original order. The source publication must contain all of the information content intended by its publisher(s) or creator(s).
- All information content from the original publication must be captured in the digitization process. The file format and digital sampling rate must be able to capture small print, line art, maps, and color.
- The physical condition of the source publication will not compromise the digitization of page image or the ability of optical character recognition software to convert the page image to machine readable text. An exception may be made when the source copy is the last known copy available.
- Examine the publications you intend to scan for general condition and whether the gutter margin will allow you to digitize the volumes without removing the binding.
- Conduct a pilot test digitizing a sample of pages representative of the overall pages in the document.
- Evaluate the results of the pilot test, including image quality control and adjust workflow as needed.
- Develop and document the digitization specifications you intend to use for the publications including ppi, workflow, and the computer workspace you will use for temporarily downloading and storing the digital image files while you are working with them.
- Assess the level of staff available and be knowledgeable about the equipment you intend to use for the work.
- Conduct a brief test each day to determine that all equipment is working properly.

REVIEWING PUBLICATIONS FOR DIGITIZATION

Review a representative sample of the publications you intend to digitize. This is to ensure that the general condition of the item will not impair the image capture.

- Volumes with mold on the book cases or pages should not be digitized.
- Volumes covered in dirt and dust should be vacuumed before digitization. This protects both the digitization equipment and the people performing the work.
- A tight gutter margin may make it impossible to create page images without curvature or distortion effects that may interfere with the user's experience in being able to read the text in the online image and the accuracy of the OCR software.
- Documents with tight bindings may require bindings to be removed for proper scanning as part of destructive digital imaging.

When digitizing a large number of volumes, it will be impossible to look at every single page to confirm its condition before the volume reaches the scanner. Turning the page when the book is on the scanner may reveal that the next page is missing, torn with missing text, or covered with marginal notes or underlining. While some situations can be remedied with a little work, others cannot.

- Dirt on the pages and marginal notes or underlining in pencil can be removed using a plastic art eraser.
- Pages underlined in pen or marginal notes in pen may be able to be erased using a plastic art eraser depending on the formula of the ink. If not, then replacement pages should be used as the source copy.
- Replacement pages will also have to be used as the source copy for missing or damaged pages.

YOUR DIGITAL CONTENT AND GOVINFO

Consider making your digital content accessible through GovInfo. The scope of content that is acceptable for ingest is found in Superintendent of Documents Public Policy Statement (SOD-PPS) 2016-2 Content Scope for GPO's System of Online Access (https://www.fdlp.gov/file-repository/about-the-fdlp/policies/superintendent-of-documents-public-policies/2738-content-scope-for-gpo-s-system-of-online-access).

By digitizing Federal Government publications, Digital Content Contributors provide a valuable contribution by:

- Increasing public access to legacy and historic Federal government information.
- Providing authentic digital surrogates of Copies of Record or surrogates for publications held in special collections.
- Preserving authentic digital copies of Federal government publications in a Trustworthy Digital Repository such as GovInfo.

We're interested in learning of your digitization interests and projects, particularly if you are interested in ingesting your content into Govnfo. If you have any questions about this guidance, or questions about digitization you may have already completed, please contact GPO at PreserveFedInfo@gpo.gov.

OTHER RESOURCES TO CONSULT

Federal Agencies Digitization Guidelines Initiative, "Technical Guidelines for Digitizing Cultural Heritage Materials: Creation of Raster Image Files." Washington: Library of Congress, September 2016. [Online accessed 02/4/2021] http://www.digitizationguidelines.gov/guidelines/FADGI%20Federal%20%20Agencies%20Digital%20Guidelines%20Initiative-2016%20Final_rev1.pdf

Preservation and Reformatting Section. Association for Library Collections and Technical Services. American Library Association, "Minimum Digitization Capture Recommendations."

Chicago: American Library Association, June 2013. [Online accessed 02/04/2021]
http://www.ala.org/alcts/resources/preserv/minimum-digitization-capture-recommendations

Version 2.1 (November 2023)
Prepared by: Digital Preservation Librarian, LSCM