LL Serial Set Full Digitization Specifications

Imaging and metadata requirements

1. **Imaging**
   Volumes shall be digitized according to the following instructions:
   a. Capture the entirety of each volume
   b. Blank leaves (recto and verso) at the beginning and end of volumes and between documents will not be captured
   c. Capture all pages at a resolution of 400 ppi
   d. Pages greater than 74 x 37 inches may have separate captures (panels) stitched together to create a 400 ppi single image
   e. Page image files shall be cropped to the text block boundaries. Foldouts and maps shall be cropped to the edge of the page
   f. Pages with color ink shall be captured in 24-bit color. All other images shall be captured as 8-bit grayscale
   g. Disbound book pages may be captured using either equipment with an automatic document feeder or an overhead camera at the contractor's discretion
   h. Bound pages must be captured using an overhead or planetary camera
   i. Foldouts must be captured under glass using an overhead or planetary camera
   j. Rotate pages to ensure the majority of the text is in the right reading orientation
2. **Master File Format**
   a. JPEG 2000
   b. Compression rate: 20
   c. Quality layer: 1
   d. Reduction level: 8
   e. Tile size full image size
   f. Progression Order: RLCP
3. **Derivative File Formats**
   a. GIF thumbnail image, 150 pixels on the long side
   b. METS/ALTO XML file one per page
   c. PDF/A 2b Searchable Text file, one file per document, 20:1 compression rate
   d. PDF/A 2b Searchable Text file, one file per volume, 20:1 compression rate
4. **Performance Level**
   a. Federal Agencies Digital Guidelines Initiative (FADGI) 3-star performance level is required for all images
   b. See Section 6 for reference to the FADGI technical guidelines
5. **Metadata Requirements**
   a. Provide uncorrected OCR output in a PDF/A 2b Searchable Text file, one file per volume
   b. Provide additional metadata for each volume as defined in the bag-info.txt file requirements in the TDL. An example bag-info.txt file is included below in Section 7 File Naming and Delivery
6. **Imaging Equipment Performance and Monitoring**
   a. Digitization shall be performed in accordance with the information capture specifications (sampling rate, tone/amplitude resolution, encoding, etc.), quality guidelines and content categories and digitization objectives recommended by the Still Image Group of the Federal Agencies Digital Guidelines Initiative ("FADGI"), which are found here: http://www.digitizationguidelines.gov/still-image/
   b. Digitization shall be performed in accordance with the latest approved version of the FADGI Technical Guidelines for Digitizing Cultural Heritage Materials to achieve a minimum 3 star-performance level for master images. The current Technical Guidelines are found here: http://www.digitizationguidelines.gov/guidelines/FADGI_Still_Image_Tech_Guidelines_2016.pdf
7. **File Naming and Delivery**

   Files shall be named to the following convention:

   Image Number: Page numbering within a volume shall begin with 0001 and continue in consecutive order through the end of the volume

   - Example: 0001-n

   Document Directory: Each document and the pages that make up the document shall be grouped in a document directory. Table of contents, indexes and errata are considered separate documents within a volume. If a document or property of a document falls outside of this naming scheme, Creekside shall contact LOC for guidance on naming the file. Document directories shall be named using the following format: {bag}-{sequence}-{doc}-{part}-{special}.

   Definitions of each component are:

   {bag} = Bag name

   - Bag name = SerialSet#_Part_Volume
     - Serial Set #: 00001-14277
     - Part: 00-77, 1a-7g, 0a-0h
     - Volume: 00-02, 0a-0c
   - Bag name examples:
     - 00559_00_00: Serial set number 559; No volume or part numbers
     - 11975_23_0a: Serial set number 11975; part number 23, volume a

   {sequence} = Sequence of the document within the Serial Set volume

   - sequence: 000
     - Three digits
     - Sequence numbers shall begin with 001 and continue in consecutive order through the end of the volume
     - LOC does not believe there are volumes with more than 999 documents in them
   - Sequence example: 042 = 42nd document within the Serial Set volume

{doc} = Document number assigned to the document by GPO at the time of printing

- doc: 0000
  - 4 digits, numbered consecutively throughout the Congress or later Congressional session (though the documents may not appear in order within a volume or across volumes)
- Doc example: 0023 = Document number 23

{part} = Part of the document if over multiple volumes

- part: 0000
  - 4 digits, as the part appears printed on the document
- Part examples:
  - 0002 = The second part of a document
  - 084A = Part 84A of a document
  - 84-1: Part 84-1 of a document

{special} = Special features of the volume - Not utilized since early in the project after discussion with the digitization vendor and GPO

- special = string
  - Text string describing the additional content
  - Covers things such as errata, index and table of contents
- Examples:
  - toc = Table of contents included in a volume
  - tod = Table of documents
  - idx = Index
  - err = Errata

<u>Document Level PDF Naming</u>: Multi-page searchable PDF/A 2b (PDF) files shall be created for each document in a volume using the same scheme used for document directory naming. If a document or property of a document falls outside of this naming scheme, Creekside shall contact LOC for guidance on naming the file. An example of a document level PDF file name would be:

- 00079_00_00-003-0056-0000: Serial Set volume number 79 – Third document in the volume – Document number 56
- 08739_01_00-005-0557-0000: Serial Set volume number 8739 – Fifth document in the volume – Document number 557

<u>Volume PDF Naming</u>: Multi-page searchable PDF/A 2b (PDF) files of the complete volume shall be created using the same naming scheme used for the bag, outlined above. An example is:

- 00079_00_00.pdf
- SerialSet#_Part_Volume.pdf

<u>Directory Structure</u>: An example of the directory structure of a bag would be:

```
00079_00_00\ - Bag name

     data\

          00079_00_00-001-0000-0000-toc\ - Table of contents

               0001.jp2

               0001.gif

               0001.xml

               …..

               00079_00_00-001-0000-0000-toc.pdf – PDF of the table of contents

           ….

          00079_00_00-003-0056-0000\ - Third document in the volume and house document number 56 the Congress

               0120.jp2 – Page 120 in the volume

               0120.gif

               0120.xml

               ….

               0079_00_00-003-0056-0000.pdf – PDF of the document

           00079_00_00.pdf – PDF of the complete Serial Set volume
```

If a document or property of a document falls outside of this naming scheme, Creekside shall contact LOC for guidance on naming the file.

<u>Bag Naming Scheme</u>: Bags shall be named as a combination of the Serial Set Volume Number, Part and Volume:

- Bag name: SerialSet#_Part_Volume

- Serial Set #: 00001-14277
- Part: 00-77, 1a-7g, 0a-0h
- Volume: 00-02, 0a-0c

a. Files will be delivered on external magnetic drives with USB connections provided by Creekside. Drives will be delivered to LOC on a delivery schedule as mutually agreed upon by Creekside and LOC and documented in the Project Plan. Drives may be retained by LOC for up to 90 days. File name instructions will be provided in the TDL
b. Volumes will be delivered in bags, one book per bag, according to the BagIt structure, which is specified here: https://tools.ietf.org/id/draft-kunze-bagit-13.txt. Details for the bag-info.txt file are provided below.
c. The bag-info.txt file will contain a list of the PDF files contained in the bag that includes a page count for each document and volume level PDF.

Details of the U.S. Serial Set bag-info.txt tags

Field	**Example**	**Comments**
Payload-Oxum:	409209311.53	
Bagging-Date:	2019-03-02	
Bag-Size:	22.7 GB	
LC-Project:	llss	Constant
LC-Bag-Id:	00079_00_00	[bag id], where “bag id” is listed in shipping manifest
LC-Content-Type:	textual	Constant
LC-Content-Process:	lc_conversion	Constant
LC-Content-Provider:	dlc	Constant
LC-Service-Provider:	tbd	Constant
LC-Items:	1 Book	Constant
LC-Master-Files:	500	Value indicates the number of JPEG 2000 files in the bag
LC-Task-Order:	LCFDL19XXXXX	Constant
LC-Doc-Type:	Disbound	Value is “Bound” or “Disbound”
LC-Foldouts:	10	The number of foldouts in this bag (to facilitate invoicing)
LC-Stitched-Panels:	4	The number of stitched panels in this bag (to facilitate invoicing)
LC-Documents:	10	The number of individual documents in this bag
LC-PDF-Page-Count:	00079_00_00.pdf:500; 00079_00_00_00-001-0000-0000-toc.pdf:10; ….. 0079_00_00-003-0056-0000.pdf:202……	[File name].pdf:[# of pages] for each PDF file in the bag, separated by ‘; ‘ (semicolon/space) If the length of this field reaches 5,000 characters before all PDF files are listed, the contractor shall add an additional field called “LC-PDF-Page-Count-2:” and include the balance of the PDF files in that field