Audit Manual for Metadata Output Validation

Tool: Raw Image Metadata Generator
Division: Georgia Newspaper Project (GNP), Digital Library of Georgia (DLG)

Overview

This audit manual outlines a step-by-step process to validate the output metadata file generated by the Raw Image Metadata Generator tool. Since the accuracy of metadata is critical to ensure the proper cataloging and archival of newspaper scans, this manual helps non-technical users perform detailed checks to detect and resolve inconsistencies.

This guide is divided into three parts:

1. Verifying Page Count Consistency
2. Verifying Descriptive Metadata (Date, Volume, Issue)
3. Rectifying Errors and Regenerating the Metadata

Each section begins with the technical premise, followed by an accessible explanation of how and why the step is necessary.

Part 1: Verifying Page Count Consistency

Technical Explanation:
The total number of images (pages) inside the scanned directory must exactly match the number of entries generated in the metadata file under the column `Reel_Sequence_Number`. If this number is off — even by one — it suggests that the input Excel file (the raw data) does not represent the real, scanned newspaper images accurately.

How to Audit:

- First, count the number of scanned image files in the relevant directory.
- Then, open the output Excel metadata file and observe the `Reel_Sequence_Number` column. This number should increment without gaps or duplicates and should match the

number of images.

- If there's a mismatch, it likely means one or more entries in the input file do not reflect the true number of pages for an issue.

How to Isolate the Error:

You can use a "divide and conquer" approach:

- Split the metadata file into halves (first half vs second half) and compare each with the scanned image set.
- Repeat this halving process until the specific row or issue causing the discrepancy is located.
- Check that the `Total_Pages` entry in the input file for that issue is accurate.
- Once found, this error can be corrected directly in the raw input Excel file.

Part 2: Verifying Descriptive Metadata (Date, Volume Number, Issue Number)

Technical Explanation:
While page count mismatches are often easier to detect, errors in fields like `Date`, `vol`, or `issue` are more subtle. These metadata fields represent descriptive attributes extracted from the newspaper itself, and an error here might go unnoticed unless carefully reviewed.

How to Audit:

- Use the metadata file's `Digital_Filename` or `Reel_Sequence_Number` columns to determine which scanned image corresponds to which row in the file.
- Open the first image of each issue in the image viewer (you only need one image per issue, usually the first page).
- Cross-check the visible header from the newspaper scan with the metadata entry in the Excel file:

- Date
- Volume number (vol)
- Issue number (issue)

Scope of the Task:

This might sound tedious, but it's quite manageable. For example:

- If the scanned directory contains weekly papers over a year, you might be reviewing around 50 images.
- Because you already know which file to open from the metadata, the process is linear and quick.

Part 3: Fixing Errors and Regenerating the Metadata File

Technical Explanation:
After errors are discovered in either page counts or descriptive metadata, the corrections must be made directly in the raw input Excel file. Once corrected, the script can be re-run to regenerate the metadata file. However, the tool cannot create multiple files with the same name in the same folder, so care must be taken when rerunning the script.

Steps to Fix and Regenerate:

1. **Open the Input File:**
   - This is the same Excel file you initially used to run the script.
2. **Locate the Error:**
   - Use the sequence number or filename to trace the problem.
   - Update fields like `Total_Pages`, `Date`, `vol`, or `issue` as needed.
3. **Save the File:**
   - Ensure you save the changes to the Excel file before proceeding.

4. **Prepare for Re-run:**
    - Option A: Delete the old metadata file from the output folder.
    - Option B: Change the output file name or path in `config.json` to avoid overwriting issues.
5. **Run the Script Again:**
    - The script will use the updated input file and generate a clean metadata file based on the corrections.

Important Considerations:

- Only one metadata file should exist in the output folder at a time.
- The correction process is iterative — you can repeat the audit and fix steps multiple times until the metadata aligns with the scanned images.

Final Remarks

This audit process ensures the integrity and usability of the metadata generated for the Georgia Newspaper Project. While the tool automates the generation process, the human verification step is essential to account for scanning irregularities, data entry inconsistencies, or source document variations.

Accurate metadata is vital for the long-term preservation and accessibility of digital newspaper archives, and following this audit manual ensures that quality remains consistently high.