NATIONAL DIGITAL NEWSPAPER PROGRAM

Digital Viewer and Validator (DVV) User's Guide

(DVV Version 2.3)

Library of Congress & National Endowment for the Humanities

April 7, 2023

Table of Contents

1. Introduction

National Digital Newspaper Program

The National Digital Newspaper Program (NDNP) is a partnership between the Library of Congress (LC) and the National Endowment for the Humanities (NEH). The goal of the project is to create a freely accessible, searchable digital repository of historically significant public domain newspapers published in the United States. An overview of the program can be found at https://www.neh.gov/divisions/preservation/national-digital-newspaper-program. Over a period of two years per grant, successful applicants will select newspapers—published in their state or territory—and convert approximately 100,000 pages into digital files (primarily from microfilm), according to the technical guidelines outlined at https://www.loc.gov/ndnp/guidelines/.

Digital Viewer and Validator (DVV)

The Digital Viewer and Validator (DVV) is a tool developed at LC for use by NDNP staff, awardee institutions (originators of newspaper content), and vendors to allow viewing of NDNP data and validation of select technical aspects of the files. It includes validation processing that can be run from a DOS command line or the DVV's own Windows Graphical User Interface (GUI). The DVV also allows visual inspection of NDNP image files, and viewing of header and metadata file content.

As part of the NDNP technical guidelines, for each newspaper page (and each reel target) that awardees submit, there will be three image files (TIFF, JPEG2000, PDF) and an XML file containing the OCR of the microfilmed page. Additionally, awardees supply METS XML files with descriptive, structural, and technical metadata for each newspaper issue and each microfilm reel digitized. The DVV is capable of displaying each of the three types of images as well as offering viewing capability for the OCR and metadata files. Awardees organize and submit the digital assets mentioned above in a logical grouping called a **batch**. The DVV has been developed to enable processing of batches of content. Examples of NDNP digital assets and batches can be found at https://www.loc.gov/ndnp/guidelines/examples.html.

Validation Library and the Validation Processor

The NDNP Validation Library is a set of rules that the DVV is programmed to enforce when processing NDNP data. Processing of NDNP batches comes in the form of validation and verification. The Validation Processor is a utility of the DVV that accomplishes the validation and verification processes. All available processes of the Validation Processor can be accessed via the DOS Command Line while the most commonly used processes are also available via the GUI.

Validation vs. Verification At-A-Glance

Validation: Validation by the DVV is the process normally completed by vendors (or awardees creating images or editing metadata) that checks data for adherence to the NDNP technical specifications. There are two types of validation processes available within the DVV: "**Validate and update**" and "**Validate with no update**". These will be described in more detail below. While validation is most commonly performed on an entire batch, it can also be used on discrete newspaper issues, reels, and individual objects.

> **Validate and Update:**
> This commands the DVV to check the data for adherence to the NDNP technical specification. If successful, the DVV creates digital signatures for each object being validated, and it records the signatures in XML files at the batch, issue, and reel level as necessary. These XML files are commonly referred to as the 'underscore 1' files (e.g. batch_1.xml). A batch without 'underscore 1' files has not been validated.
>
> **Validate With No Update**:
> This commands the DVV to simply check the data for adherence to the NDNP technical specification. No signatures are created, and no changes are made to the batch. This is often used as a test run validation prior to initiating "validate and update".

Verification: Batch verification checks the digital signatures of all files in the batch. This process can only be initiated after a batch has been validated. Typically vendors and awardees verify that a batch is valid before shipping. Verification is a process to help ensure integrity of data over time. Generally, it should occur anytime data moves across storage systems and anytime changes are made to a batch.

About this Guide

This guide was created by Library of Congress staff primarily for DVV users directly involved with the National Digital Newspaper Program (NDNP). Questions or comments about this guide or the DVV software can be sent to ndnptech@loc.gov.

2. System Requirements

Operating System – The DVV has been configured to run on Windows. It has been tested on Windows 10.

Memory - The validation of digital images, often many megabytes in size, is a resource intensive operation. A minimum of 1024 megabytes (1 gigabyte) of available RAM for the DVV is recommended and the default installation assumes this amount exists. Please note the use of "available RAM", means that even if you have a system with 2 gigabytes of RAM, but only 512 megabytes are available (not in use), then the DVV will fail to start. When in doubt, or when encountering memory errors, close all other programs and try again. If you continue to have problems, please contact the Library of Congress. This version of the DVV was tested with 16 gigabytes of available ram.

Disk Storage – For any sufficiently modern computer, disk space should not be an issue. However, for users that may be running older equipment, here are guidelines and gotchas to keep in mind:

> A minimum of about 60 megabytes of disk space is required to install (unzip) the DVV.
>
> XML and HTML files are generated during normal processing, will add to the space requirement, and are stored in the html directory of the install. These should be deleted automatically upon restart of the DVV.
>
> Log files, which may grow to large sizes, can be safely deleted any time the DVV is not running.
>
> This version of the DVV was tested with 100 gigabytes of available disk storage

Processing Power – As with any resource intensive operation, more is better. If you are running hardware that is older than the past 7 years, please contact the Library of Congress for directions on how to best tune the DVV for better performance. This version of the DVV was tested with an Intel i7 processor.

Java –Version 17 (64-bit) of the Java Development Kit (JDK) was used in the creation of the DVV. Java (JDK/JRE 11) or OpenJDK Version 11 or newer (64-bit) is required to run the DVV. The DVV run script assumes that the environment variable JAVA_HOME is set, and it is build version 11 or newer. If it is not set you will receive an error. If you have problems setting your environment variable please see your system administrator before contacting the Library of Congress for help. **Note: Java is not pre-packaged with the DVV software. It must be installed on your local system.**

PDF Viewer – A PDF viewer is needed in order to view PDF files. We recommend Adobe Acrobat Reader 7.0 (or later). It is free and can be downloaded from https://get.adobe.com/reader/ .

3. Download the Software

The DVV software is made available packaged in a .zip file. This file can be easily downloaded from: https://www.loc.gov/ndnp/tools/ .

Identify the version you desire and save it to your local computer's hard drive by right-clicking the link, and selecting "Save as…" or "Save Link as…". In the Windows dialog box, change the destination to a local drive such as the C:\ drive. (NDNP participants should use the most recently released version of the DVV software.)

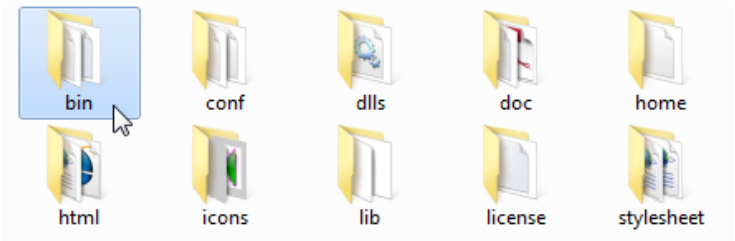
Once the file has been downloaded, Unzip or "extract" the zip file using a utility such as "WinZip".

For best results, ensure that the DVV files are extracted to a location on your computer's C:\ drive. Running the DVV from a networked directory or location has caused known issues, and therefore should be installed and run locally.

For example, C:\dvv2 would be a good location to run the software.

4. Starting the DVV

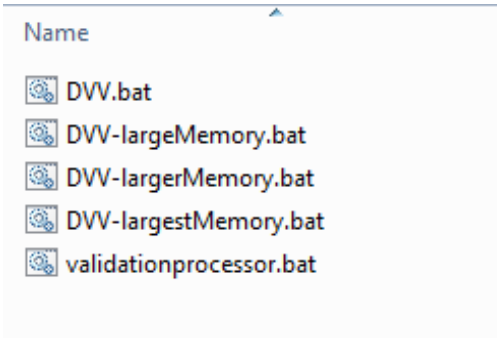
Inside the dvv2 directory, you will find several sub-directories (pictured below). Open or double-click the **\bin** directory to find the **DVV.bat** file. This is the command used to launch the graphical user interface (GUI) of the DVV software. For quicker launching, consider creating a shortcut to this icon and placing it on your Desktop.



Contents of the \bin directory

Click DVV.bat to launch the DVV

Right-click DVV.bat to "Send to… Desktop (create shortcut)"



Setting the JAVA_HOME Variable

When setting up the DVV for the first time (or following Java updates), it is essential to set the JAVA_HOME variable correctly. When this variable is not set, the following message will display in the CMD prompt window:

"ERROR: JAVA_HOME is not set and no 'java' command could be found in your PATH,"

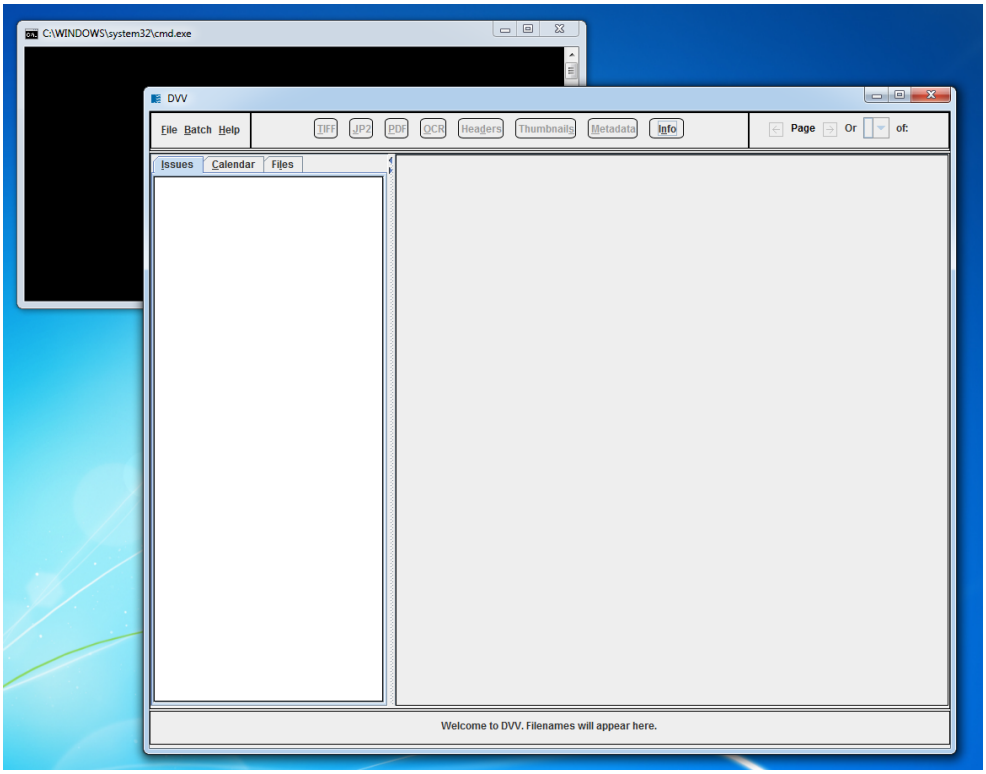
To set the variable using a script, follow the steps below:

1. Inside the dvv2 directory, open or double-click the \helpScripts directory
2. Inside the \helpScripts directory, double-click **setJavaHome.bat**
3. If prompted, press [R] to run once (omit the brackets)
4. Go back into \dvv2\bin and double click **DVV.bat**
5. This should only need to be done once. For later sessions, you would only need to click **DVV.bat** to launch.

If these steps fail to set the JAVA_HOME variable, see additional start-up troubleshooting below for a manual solution to setting the variable.

The DVV GUI and Command Prompt Window

When launched, both the DVV GUI and a Command Prompt window will open (pictured below-- GUI on right, Command Prompt on left). Users should keep the Command Prompt window open alongside the GUI. Messages may display in the Command Prompt window when using the DVV GUI.



Start-up Troubleshooting

No Java Installed

If you receive errors similar to the following, you may not have Java or the correct version installed. Contact your system administrator for help installing Java. See "Section 2: System Requirements" for more details.
Possible error messages that may indicate this problem:

```
java.lang.UnsupportedClassVersionError

 Cannot find path … because it does not exist.
```

Setting the JAVA_HOME Variable Script

If you are sure that Java is installed, and you encounter this error, please do the following:

"ERROR: JAVA_HOME is not set and no 'java' command could be found in your PATH,"

1. Inside the dvv2 directory, open or double-click the \helpScripts directory
2. Inside the \helpScripts directory, double-click **setJavaHome.bat**
3. If prompted, press [R] to run once (omit the brackets)
4. Go back into \dvv2\bin and double click **DVV.bat**
5. This should only need to be done once. For later sessions, you would only need to click **DVV.bat** to launch.

Setting the JAVA_HOME Variable Manually

If setting the JAVA_HOME variable using the setJavaHome.bat script (above) does not work, consider setting the variable manually using the command line. **These steps may also be followed if Java has been updated on your systems since you first installed the DVV**.

1. Locate exactly where Java is installed on your system. (For this example, we will use this location: C:\Program Files (x86)\Java\jre1.8.0_51.) Your location will be unique to your computer.

2. Open the command prompt and type this command substituting your specific Java location. Note the use of quotes and case (**JAVA_HOME must be all caps**):

```
setx JAVA_HOME "C:\Program Files (x86)\Java\jre1.8.0_51"
```

3. Execute the command by pressing "Enter"

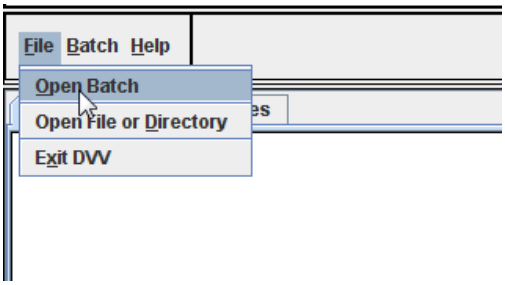
4. Close the command prompt window and re-try launching the DVV

Java Object Heap Space Error

If you encounter the message below, close down other applications you may have open and try again. If that doesn't work, try re-starting your computer.

```
"Error occurred during initialization of VM
Could not reserve enough space for object heap
Error: Could not create the Java Virtual Machine.
Error: A fatal exception has occurred. Program will exit.
Press any key to continue . . ."
```

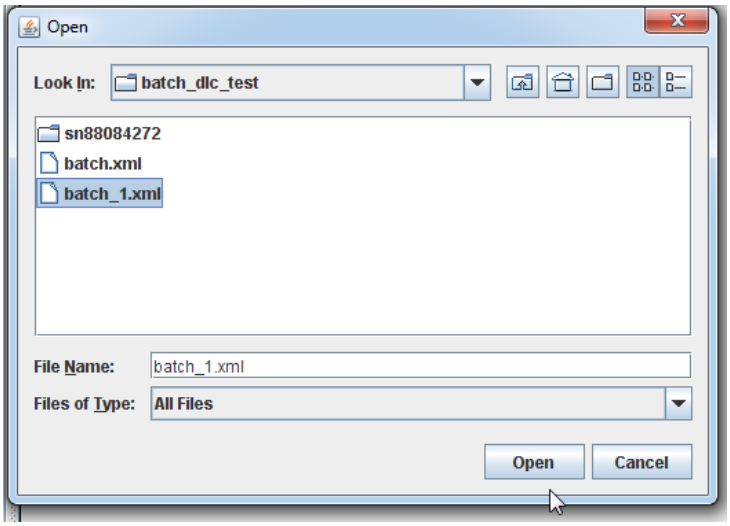
5. Opening a Batch



To open a batch, click on 'File' and then select 'Open Batch'. This action will cause a file manager dialog to open. Navigate to the batch file for the batch to be viewed/validated/verified, click the batch file name to select it and then click on 'Open' to load the batch file.

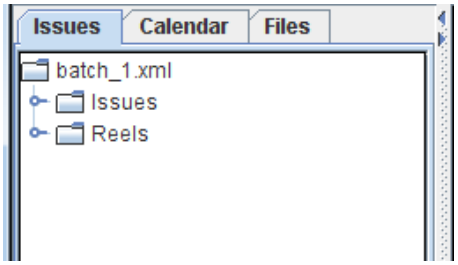
Note – if accessing a batch located on a remote drive, ensure that the remote drive is mapped locally (using Tools, Map Network Drive from the Windows File Manager interface) and use the facilities of the DVV file manager interface to navigate to the remote batch file.

The following illustration shows the dialog box being used to open a batch_1.xml file.



A progress bar will display the batch loading progress (shown below). After a batch has been opened, images and metadata can be viewed as described below in **Viewing Images and Metadata**.

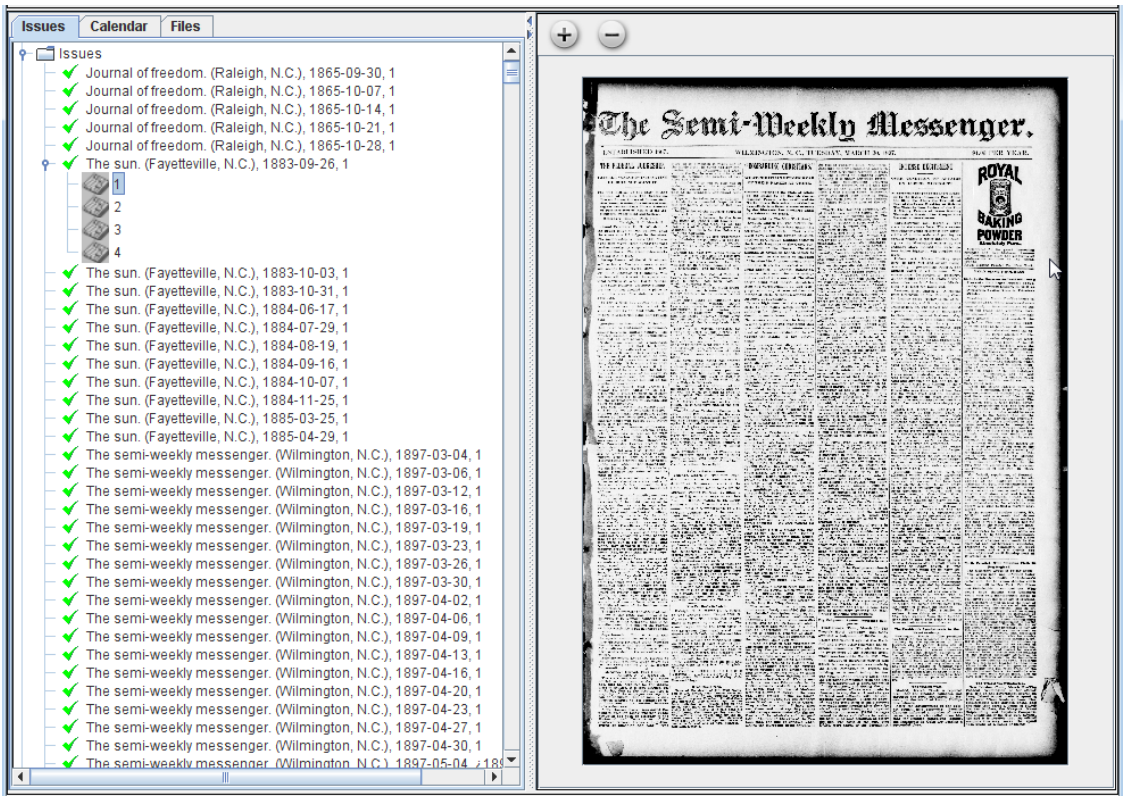
6. Viewer



Images and metadata can be organized for viewing in 3 ways: Issues View, Calendar View, and Files View. Users can toggle between Issue View and Calendar View to review objects within a batch structure. Files View enables users to view images and metadata (including image header metadata) as either part of a batch structure or as not part of a batch structure. Clicking on the appropriate tab in the left-hand navigation pane will take a user to the desired view. Each view will be described below.

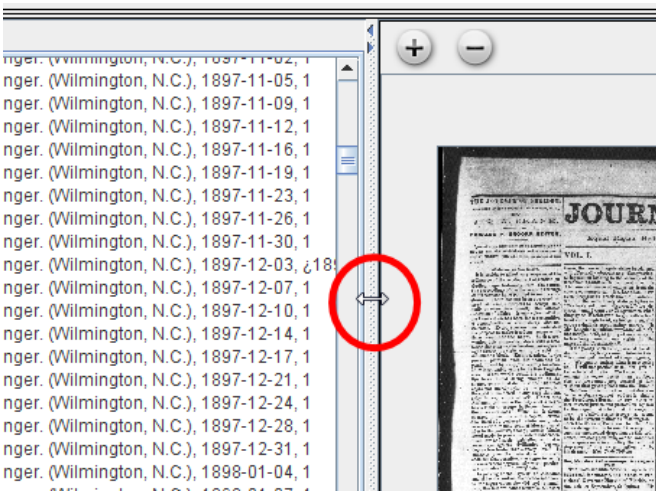
Issues View

By default, the Issues View mode is displayed upon loading a batch. Issues View can be activated by clicking on the Issues tab. Double-clicking the Issues folder will open the issues of the batch. Clicking on a specific issue title will open the pages of that specific issue.



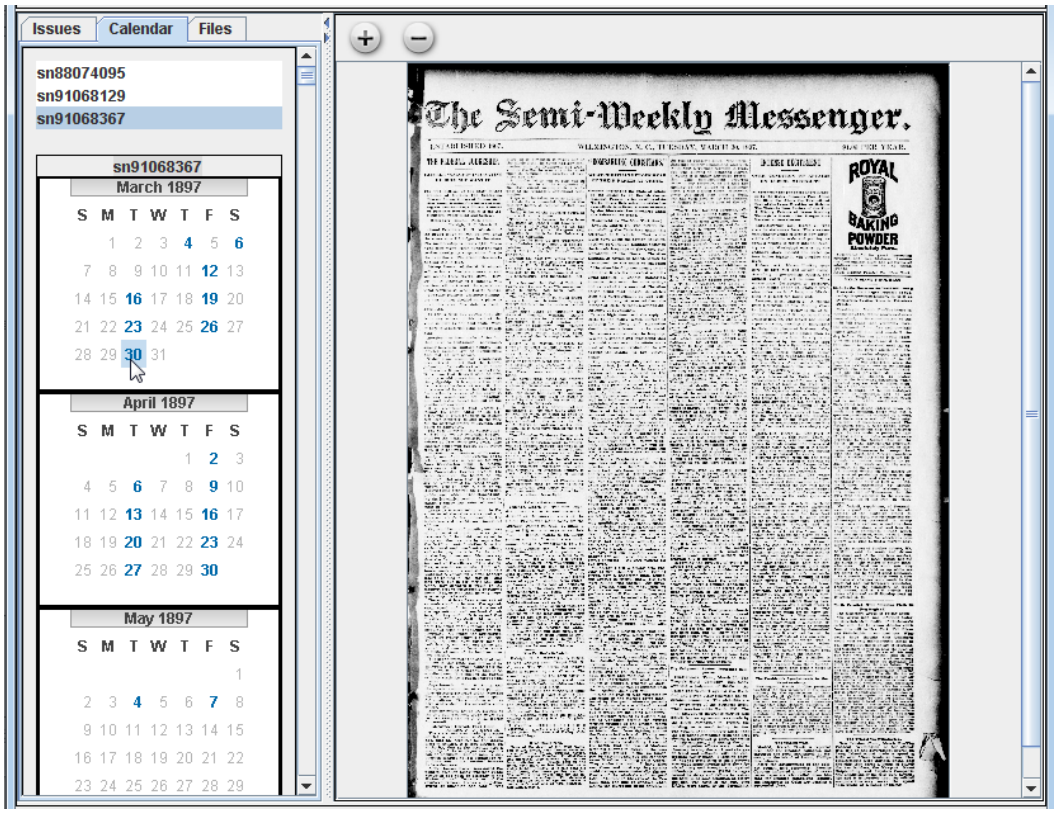
Resizing the Issues View Panes

If parts of the issue title or date are obscured, users may re-size the left pane by hovering over the bar that divides the left and right panes until a double ended arrow appears. Users can then drag the bar left or right to the desired size.



Calendar View

To view a batch in the Calendar View, click on the "Calendar" tab. Under the tabs, each title, listed by LCCN, will be available to view. When an LCCN is selected, monthly calendars will display for months in which digitized content exists on the batch. Highlighted dates indicate that digital content exists for that specific date. Click on a highlighted date to view images and metadata for that issue. If multiple editions exist on a specific date, a pop-up window will open prompting the user to select an available edition for that date.

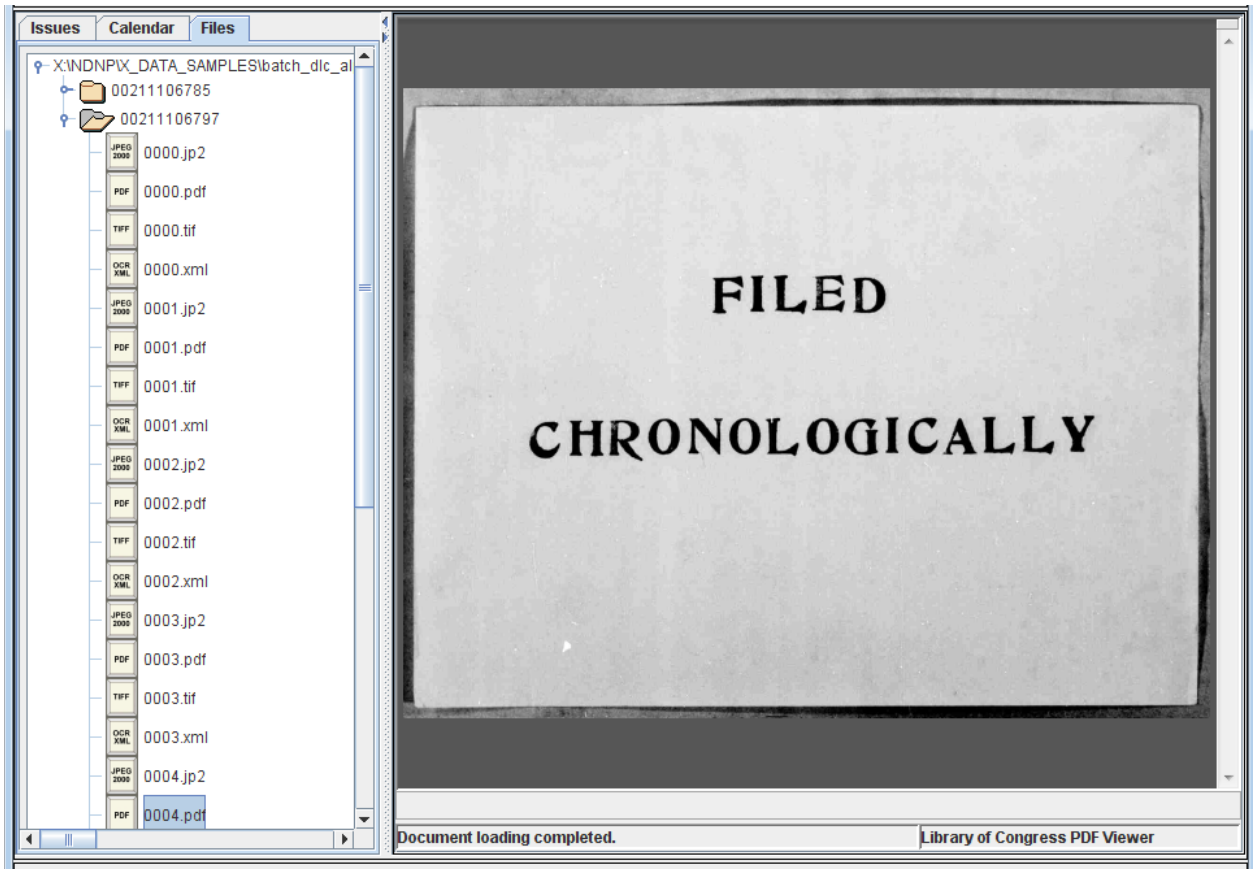


Files View

The Files View allows a user to view images and metadata, including embedded header information in images. This view does not require nor rely on the NDNP batch structure or NDNP validation library software. Files View can be accessed in two ways:

- File > Open File or Directory >Navigate to desired directory
- Files Tab > Browse to Directory > Navigate to desired directory

Once the contents of the desired directory are displayed in the left pane of the DVV, users can double-click images and metadata to view in the right pane. Below is a screenshot of a PDF opened in the Files View.



Viewing Images and Metadata

Whether in Issues, Calendar, or Files View, there are several ways to examine images and metadata. The various buttons and controls are pictured and described below.

Image and Metadata Buttons:

Page Navigation Controls:

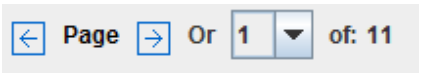


Image Zoom Buttons:

TIFF: Selecting the **TIFF Button** allows a user to view TIFF images. The **TIFF** view is the default view when opening issues in the Issues and Calendar view. When viewing a TIFF image, a user can use the **Image Zoom Buttons** to zoom in and out of the TIFF. When active, users can also use the **Page Navigation Controls** (arrows and page drop-down) to page through an issue's TIFF files.

JP2: Selecting the **JP2 Button** allows a user to view a JPEG rendering of the JPEG 2000 images. When using the JP2 viewer, a user can use the **Image Zoom Buttons** to zoom in and out of the images. When active, users can also use the **Page Navigation Controls** (arrows and page drop-down) to page through an issue's JP2 files.

PDF: The **PDF Button** allows a user to view PDF images. When activated, Adobe Acrobat Reader (free software available at: https://get.adobe.com/reader/) will launch inside the DVV's right image viewing-panel. Zooming and other actions can be achieved by using Adobe Reader's built-in tools. When active, users can also use the DVV's **Page Navigation Controls** to proceed forward or back through an issue's PDF files.

OCR: The **OCR Button** allows a user to view HTML renderings of ALTO XML files. When active, users can also use the DVV's **Page Navigation Controls** to proceed forward or backwards through an issue's OCR files.

Headers: Header data embedded in TIFF, JP2, and PDF images can be viewed first by clicking on the **Headers Button**, followed by selecting the appropriate image header for viewing. The header metadata is converted via XML Stylesheet Language Transformation (XSLT) to a formatted HTML page. Key metadata values have been formatted in other colors to assist in the quality review process.

Thumbnails: Thumbnail-sized images, derived from a selected issue's TIFF files, are displayed by clicking the **Thumbnails Button**. Clicking on a thumbnail image will cause the TIFF image of the same page to be loaded.

Metadata: Users can examine the metadata for both issue and reels METS files by selecting the **Metadata Button.** The METS XML is converted via XML Stylesheet Language Transformation (XSLT) to a formatted HTML page and displayed in a browser embedded in the DVV. Key metadata values such as volume number, edition, date, and page number have been formatted in other colors to assist in the quality review process.

7. Info Screen 

When viewing a batch in the DVV GUI, the Info screen will provide information about the batch. To show the Info screen, users should click on the Info button (shown above). The Info button turns red when new messages are available.

As shown below, following the loading of a batch, general messages, a batch load status, and NDNP batch identifiers will display in the Info screen. When an Internet connection is available, the DVV will also connect to Chronicling America to determine if bibliographic title records for the titles in the batch are currently ingested in the Chronicling America directory (https://chroniclingamerica.loc.gov/search/titles/).

When bibliographic title records for the LCCNs on a batch are found in the Chronicling America directory, the LCCN numbers and titles will be displayed in the Info screen. LCCN numbers will be hyperlinked to the Chronicling America MARC record view of said title.

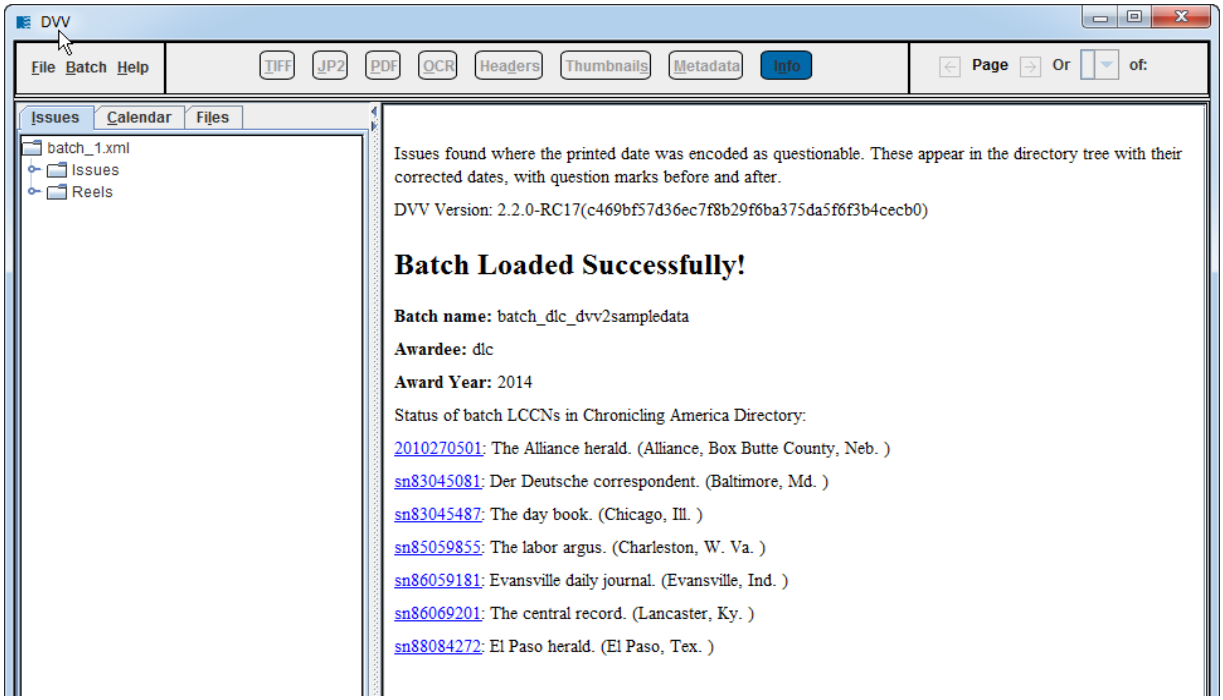
When bibliographic title records for the LCCNs on a batch are **NOT** found in the Chronicling America directory, an appropriate message will be displayed in red notifying the user of the error.

Error messages as well as validation and verification summary messages will also appear in the Info screen (explained below in section 8).

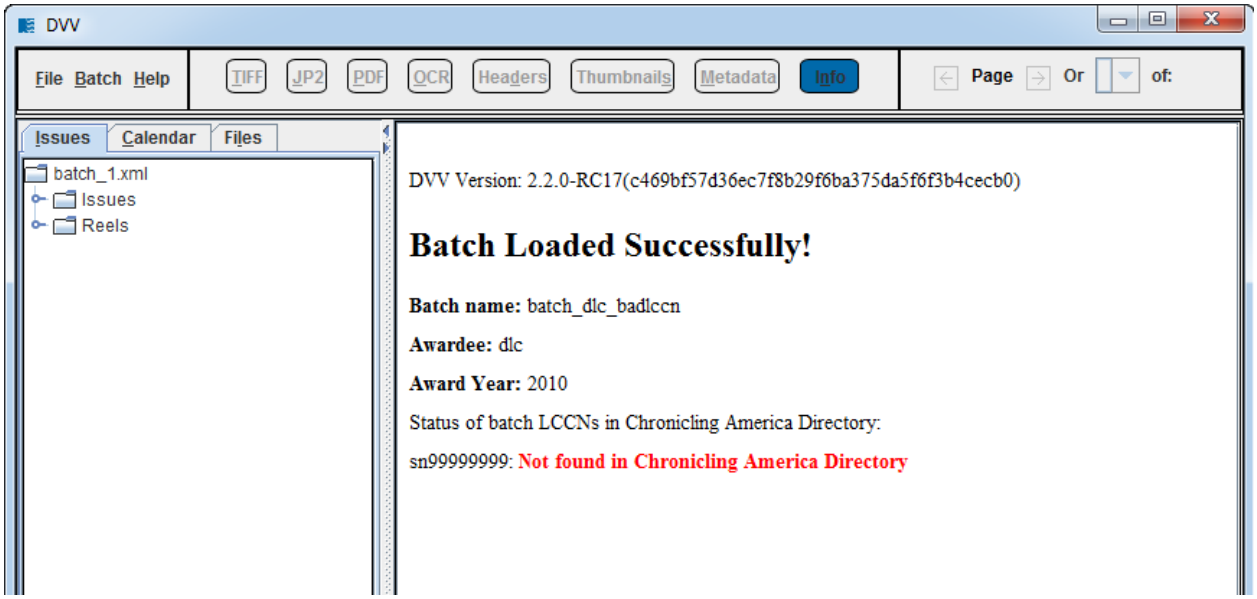
(NDNP participants should ensure that correct LCCNs and appropriate bibliographic title records are used in the metadata for their project. This topic is out of scope for this document, but information is available from LC staff.)

Info Screen Messages

Example message for LCCNs found in Chronicling America Directory:



Example message for an LCCN **NOT** found in Chronicling America Directory:

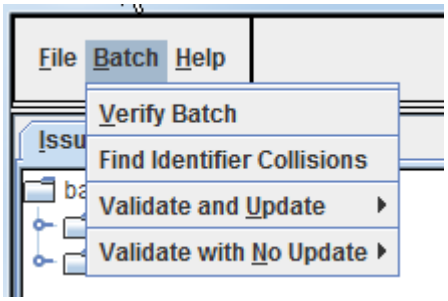


8. Batch Processing (GUI)

This section explains the validation and verification processes available via the DVV's Graphical User Interface (GUI). Section 1 of this manual provides basic information about the DVV validation library and validation processor. The GUI offers four categories of processing tasks: **Verify Batch**, **Find Identifier Collisions**, **Validate and Update**, and **Validate with No Update**. See **Appendix A** for instructions on how to perform batch processing tasks using the Windows Command Line. See **Appendix B** for details about the **Find Identifier Collisions** feature.

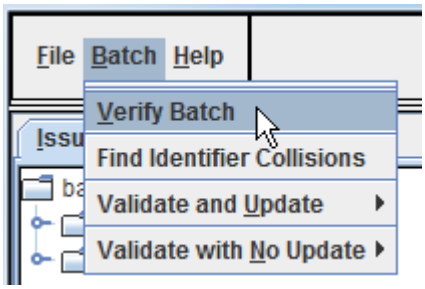
To initiate any of these batch processes:

1.) Start the DVV (section 4),
2.) Open a batch (section 5) in the GUI, and
3.) Click the Batch menu (pictured below).



Verify Batch

Batch verification is only completed on validated batches. Users should ensure that when opening a batch (section 5), that the batch_1.xml file is selected.



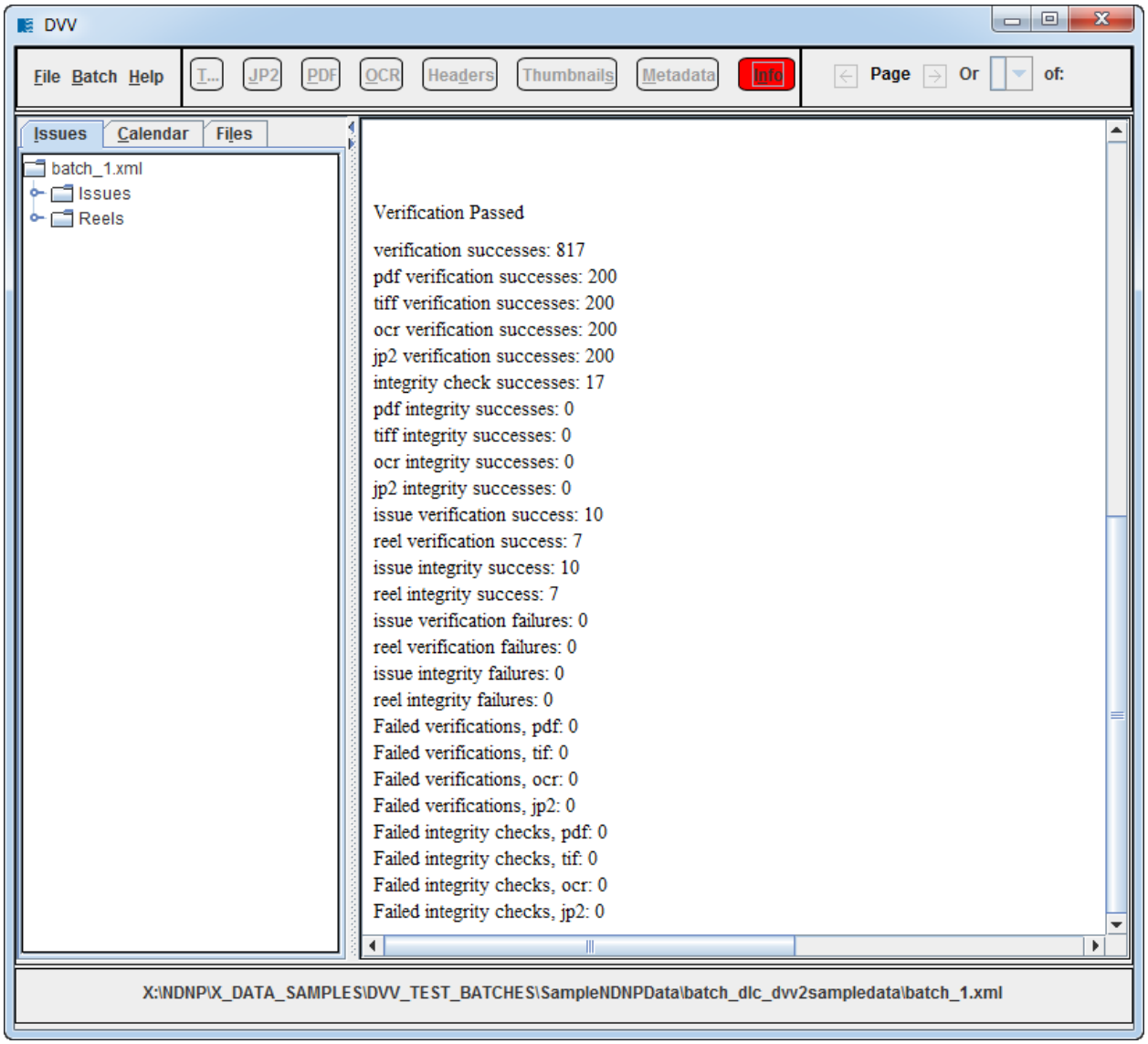
To verify a batch:

1.) Start the DVV (section 4),
2.) Open a batch (section 5) in the GUI,
3.) Click the Batch menu,
4.) Select "Verify Batch" (pictured left).

A message will appear in the Info Screen to indicate that the verification command has been invoked. Real time file level status messages of the verification process will also display in the CMD Prompt window that opens alongside the DVV GUI.

When verification is complete, a summary of the verification results will display in both the CMD Prompt window and in the GUI Info Screen. See screen shot of the verification summary below.

Verification Summary in GUI:



Validate with No Update

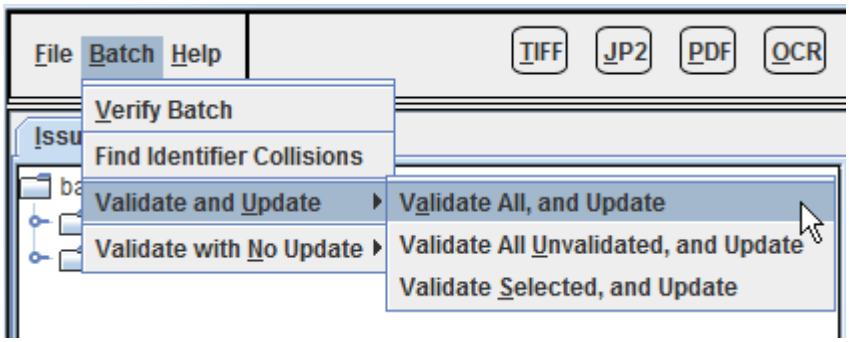
Validate with No Update runs the data through the checks of the validation library but does not create the digital signatures. This can be used as a trial validation before committing to creating signatures. Follow the steps below for "Validate with Update", but chose the corresponding "No Update" options instead. When validation processing is complete, a summary of the validation results will display in both the CMD Prompt window and in the GUI Info Screen. The validation summary is similar to the screen shot shown above for verification summary in the GUI.

Validate and Update

Validation can be performed on an entire batch or on select images or metadata. Note that the procedures are the same for brand new validations and re-validating of data that was once validated.

To **validate and update** (write digital signatures for) an entire batch:

1.) Start the DVV (section 4),
2.) Open a batch (section 5) in the GUI,
3.) Click the Batch menu,
4.) Select "Validate and Update",
5.) Select "Validate All, and Update" (pictured below),
6.) Consult Info screen for processing and validation summary messages



For a batch that has been partially validated, a user may wish to validate only the files that have not been validated, while skipping files that have already been validated.

To **validate and update** (write digital signatures for) any unvalidated files within a batch:

1.) Start the DVV (section 4),
2.) Open a batch (section 5) in the GUI,
3.) Click the Batch menu,
4.) Select "Validate and Update",
5.) Select "Validate All Unvalidated, and Update" (pictured above),
6.) Consult Info screen for processing and validation summary messages.

It is also possible for a user to validate single issues and reels individually.

To **validate and update** (write digital signatures for) any issue or reel within a batch:

1.) Start the DVV (section 4),
2.) Open a batch (section 5) in the GUI,
3.) Select the issue or reel for which validation is desired (select in DVV left pane),
3.) Click the Batch menu,
4.) Select "Validate and Update",
5.) Select Validate Selected, and Update,
6.) Consult Info screen for processing and validation summary messages.

Appendix A: Common DVV Command Line Commands

Command Line Usage of the Validation Processor

The Validation Processor is the command line component of the DVV and can be used to validate the contents of an entire batch, the contents of a single issue or reel within a batch, or a single image file. To run the Validation Processor, open a command (DOS) window and change to the directory in which the DVV was installed.

To open the command (DOS) window for DVV command line:

- Click on 'Start' in the Windows menu bar
- Enter the value 'cmd' (without the quotes) in the "Search Programs" box
- Press "Enter" or select program cmd.exe
- Navigate to the **\bin** directory of where the DVV is stored. In most cases this can be achieved by typing: **cd C:\dvv2\bin**

- You are ready to go when your prompt looks like this: **C:\dvv2\bin>**

TIP: You can also open the command prompt at a specific location by holding the SHIFT key while right-clicking in any Windows Explorer window. Select "Open command window here".

Command Line Commands

Below are the most commonly used commands to determine whether data technically meets the NDNP specifications.

Commands below assume the DVV is installed on the C:\ and use examples file paths using a batch on G:\batch_dlc_demo. You will need to substitute the correct path to your batch.

Display a List of Commands

To show a list of all Validation Processor commands, simply type validationprocessor after the prompt.

```
validationprocessor
```

Validate Batch No Update

Validate an entire batch with the no_update option (checks everything but leaves no signatures). This is useful if you don't want to alter data, or as a dry run before validating with update.

```
validationprocessor BATCH G:\batch_dlc_demo\batch.xml no_update
```

Validate Batch Update

Validate a batch with the update option (checks everything and creates_1.xml files with validation signatures.)

```
validationprocessor BATCH G:\batch_dlc_demo\batch.xml update
```

Verify Batch

Verify a batch with the verify option. This looks for and compares validation signatures and only works when pointed to batch_1.xml type files. This will not work on batch.xml type files.

```
validationprocessor BATCH G:\batch_dlc_demo\batch_1.xml verify
```

Validate Individual Images

Validate an individual JPEG2000, TIFF, or PDF file.

`validationprocessor JPEG2000` [enter file path]

Often when you receive a replacement file, after you validate the individual file, you will also have to validate the issue as well (see validating an issuemetsrecord below).

Validate a Single Issue No Update

This is done when needing to test the validity of the xml file and all images of a single issue.

`validationprocessor ISSUEMETSRECORD` [enter file path to ISSUE XML file] `no_update`

Validate a Single Issue Update

This is done when needing to validate and create signatures of all contents of an issue (the xml file and all images). Dry run command is above.

`validationprocessor ISSUEMETSRECORD` [enter file path to ISSUE XML file] `update`

Tips:

Any action that uses the "update" option alters the data by adding the _1 files. It may overwrite previous _1 files, which masks information from the previous validation. Use the "update" option with some consideration.

You can return a validated batch to its unvalidated state by searching and deleting "*_1.xml" -- all the _1 files --from the file structure, using Windows Explorer.

Appendix B: Find Identifier Collisions

Background

Participants of the National Digital Newspaper Program (NDNP) primarily use the DVV to review data that will eventually be made accessible in the Library of Congress (LC) web site, Chronicling America (https://chroniclingamerica.loc.gov/). Using the Chronicling America application programming interface (API), the DVV offers a feature, **Find Identifier Collisions**, which enables users to check if content currently being reviewed in the DVV is already available in Chronicling America.

What is an Identifier Collision?

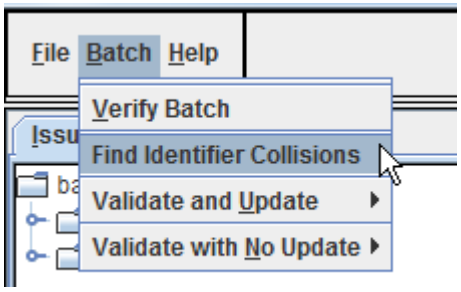
Newspaper issues digitized for NDNP are represented in Chronicling America by unique identifiers comprised of the LCCN, the date and edition number (i.e, sn99999999/01-01-1901/ed-1). If there are multiple issues containing the same LCCN-date-edition sequence, this is called an "identifier collision". When an identifier collision occurs in Chronicling America, only the latest issue ingested will be displayed. Some identifier collisions are intentional (i.e. replacement issues submitted to overwrite incorrect data), while many are not. NDNP participants should consult LC staff for more information regarding the use of identifier collisions for replacing data.

Finding Identifier Collisions

Using the DVV, identifier collisions can be identified before sending data to LC.
An internet connection is required to run this report.

To find potential **identifier collisions** in a batch:

1.) Start the DVV (section 4),
2.) Open a batch (section 5) in the GUI,
3.) Click the Batch menu,
4.) Select "Find Identifier Collisions",
5.) Consult the Info Screen for results.



If identifier collisions are found in the batch, issues will display in a report as pictured below. Clicking the issue hyperlinks in the report will take a user to relevant digitized newspaper issues in Chronicling America.

Example Identifier Collisions Report from Info Screen

Identifier Collisions Report

2010270501: Digitized content for this LCCN exists in Chronicling America

Issue 1902-10-31 edition: 1 already exists in Chronicling America

Completed identifying collisions in batch.