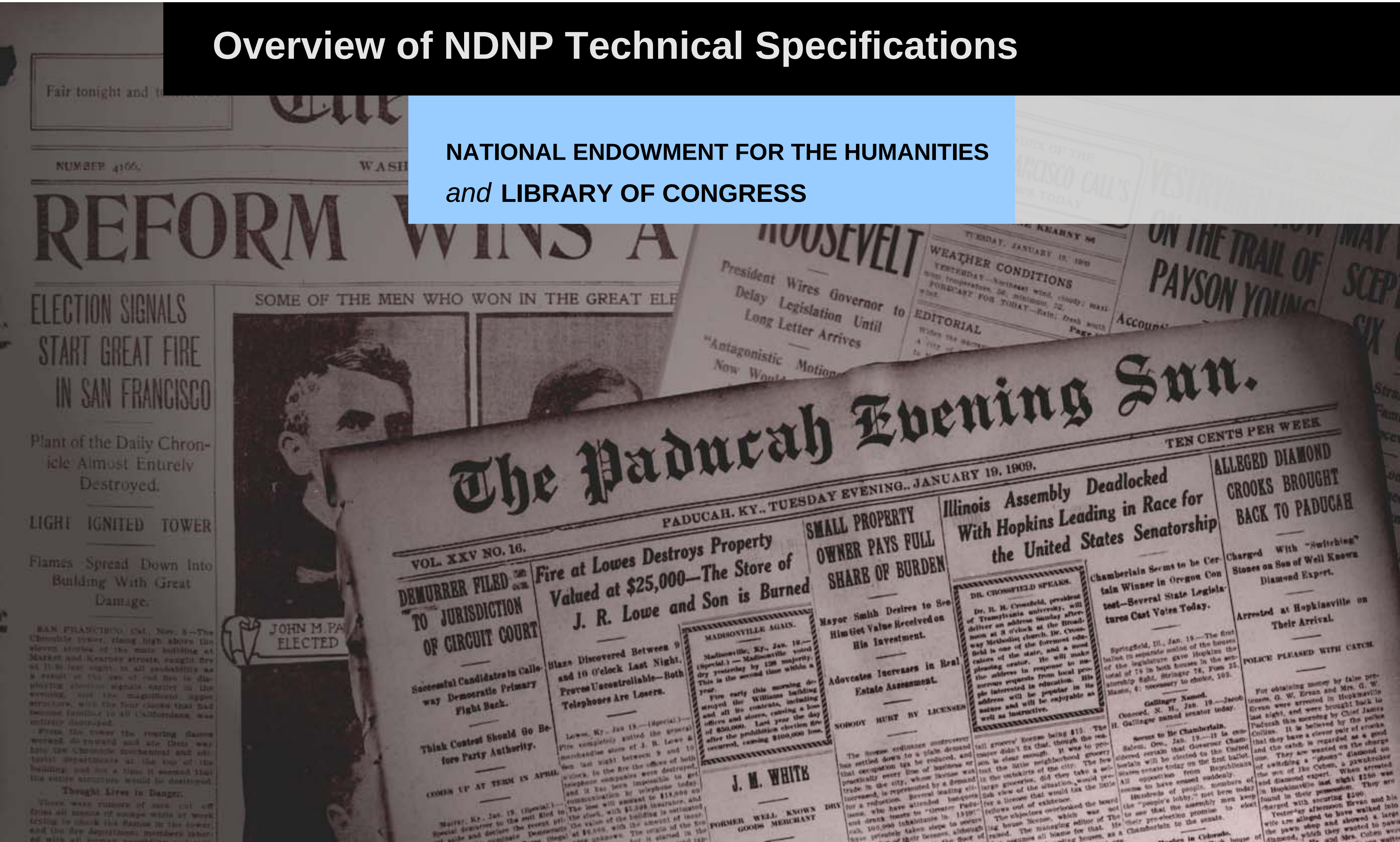
NATIONAL ENDOWMENT FOR THE Humanities

LIBRARY OF CONGRESS

Overview of NDNP Technical Specifications


REFORM WINS A
ELECTION SIGNALS START GREAT FIRE IN SAN FRANCISCO
SOME OF THE MEN WHO WON IN THE GREAT ELE
ROOSEVELT
WEATHER CONDITIONS
EDITORIAL
ON THE TRAIL OF PAYSON YOUNG
The Paducah Evening Sun.
PADUCAH, KY., TUESDAY EVENING, JANUARY 19, 1909.
TEN CENTS PER WEEK
VOL. XXV NO. 16.
DEMURRER FILED TO JURISDICTION OF CIRCUIT COURT
Fire at Lowes Destroys Property Valued at $25,000—The Store of J. R. Lowe and Son is Burned
SMALL PROPERTY OWNER PAYS FULL SHARE OF BURDEN
Illinois Assembly Deadlocked With Hopkins Leading in Race for the United States Senatorship
ALLEGED DIAMOND CROOKS BROUGHT BACK TO PADUCAH
J. M. WHITE

NATIONAL ENDOWMENT FOR THE HUMANITIES

and LIBRARY OF CONGRESS

Philosophy

- Digitization from preservation microfilm print negatives (2n) provides the most cost-efficient approach for large-scale digitization
- Distributed digitization model requires "rich" technical description
  - Structured enough to implement consistently
  - Flexible enough to represent range of intellectual organization
- Minimize opportunities for divergence from technical requirements
- Avoid "garbage in, garbage out"
  - Inspect for conformance to intellectual and technical intent
  - Validate against existing standards and profiles
  - "Trust, but verify"
- Automate processes where possible

The Journey from Analog to Digital

- Assessing master negative reels
- Technical considerations during conversion
- Awardee and vendor responsibilities

Technical Inspection

- Quality of original document
- Quality of microfilm capture
  - Questionable does not mean bad, it means questionable
- Reduction ratio
- Resolution test patterns
  - Their existence indicates a standards-based microfilm process
  - Do you examine them; how do they look?
- Density variations within & between exposures
- OCR test of sample page images

Imaging Microfilm Using Targets to Monitor Quality

- Required for NDNP scanning – targets imaged with every reel
- Why?
  - Supports imaging objectives
  - Measurable record is created
  - Analyze imaging performance with software
  - Vendor evaluation, before/during/after
- Preservation Microfilm Scanner Target PMT-1
- Imaging specifications
- Target analysis software
  - http://www.imagescienceassociates.com/softwaretools/mscan/mscan.php

Creating and Validating NDNP Data Objects

- Images
- OCR
- Metadata
- Data validation and inspection

Archival Image: TIFF

- Conforms with TIFF 6.0
- 8-bit grayscale
- 400 dpi preferred
- Uncompressed
- Only deskewing should be applied
- Cropped to page edge
- TIFF tags required for preservation
  - Matches 2009 Federal Agencies Digitization Guidelines

Production Image: JPEG 2000

- Conforms with JPEG 2000, 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
- Profile prepared with assistance of Rob Buckley, Xerox Labs

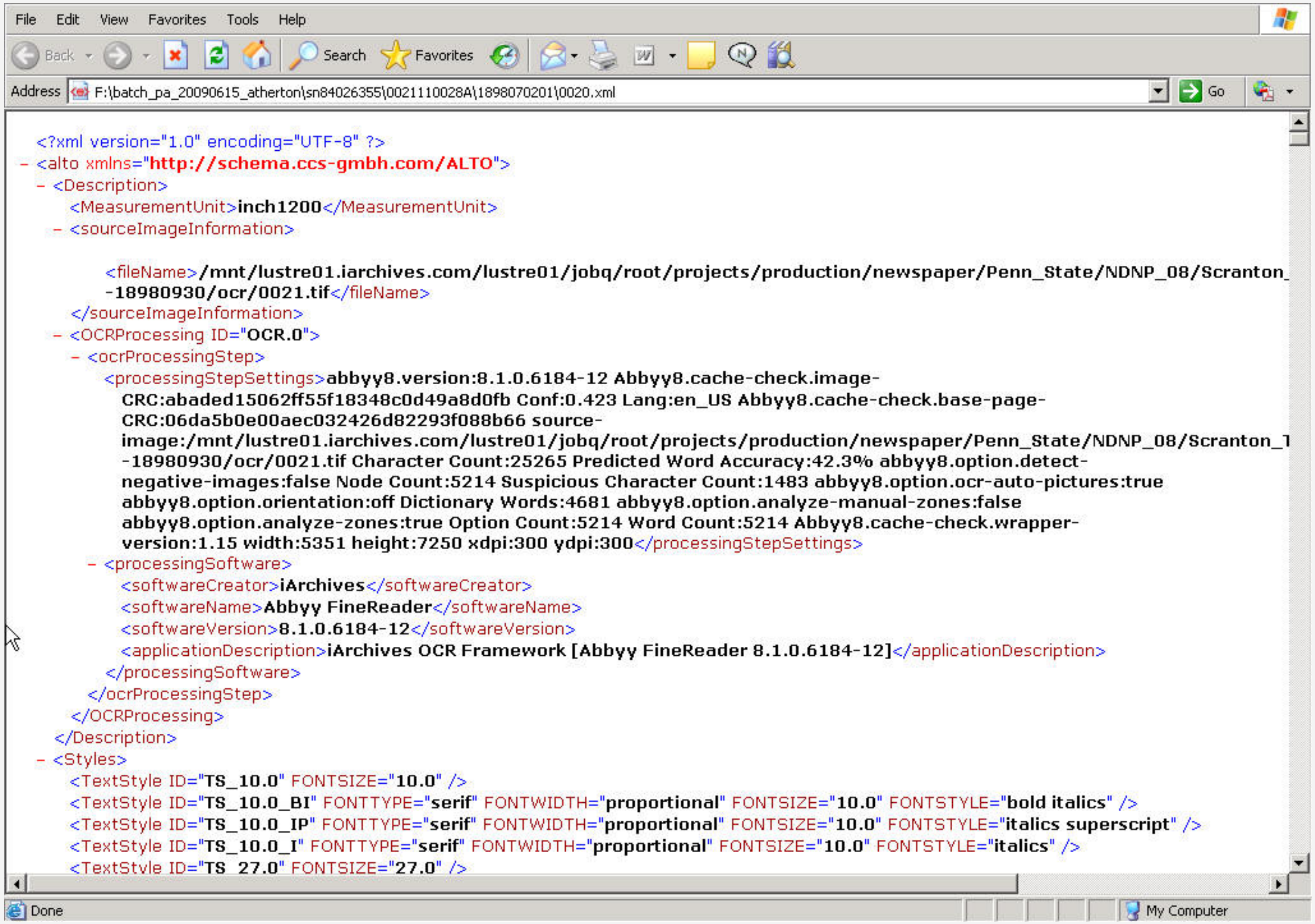
Printable Image: PDF

- 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

Searchable OCR text: NDNP-ALTO

- Conforms with ALTO (Analyzed Layout and Text Object) schema
- NDNP-ALTO is a simplified version of ALTO
- ALTO is product of EU-funded METAe project
- Mapping of OCRed text to image coordinates

OCR Format: NDNP-ALTO



Address F:\batch_pa_20090615_atherton\sn84026355\0021110028A\1898070201\0020.xml

```
  <?xml version="1.0" encoding="UTF-8" ?>
- <alto xmlns="http://schema.ccs-gmbh.com/ALTO">
  - <Description>
      <MeasurementUnit>inch1200</MeasurementUnit>
    - <sourceImageInformation>

          <fileName>/mnt/lustre01.iarchives.com/lustre01/jobq/root/projects/production/newspaper/Penn_State/NDNP_08/Scranton_
          -18980930/ocr/0021.tif</fileName>
      </sourceImageInformation>
    - <OCRProcessing ID="OCR.0">
      - <ocrProcessingStep>
          <processingStepSettings>abbyy8.version:8.1.0.6184-12 Abbyy8.cache-check.image-
            CRC:abaded15062ff55f18348c0d49a8d0fb Conf:0.423 Lang:en_US Abbyy8.cache-check.base-page-
            CRC:06da5b0e00aec032426d82293f088b66 source-
            image:/mnt/lustre01.iarchives.com/lustre01/jobq/root/projects/production/newspaper/Penn_State/NDNP_08/Scranton_T
            -18980930/ocr/0021.tif Character Count:25265 Predicted Word Accuracy:42.3% abbyy8.option.detect-
            negative-images:false Node Count:5214 Suspicious Character Count:1483 abbyy8.option.ocr-auto-pictures:true
            abbyy8.option.orientation:off Dictionary Words:4681 abbyy8.option.analyze-manual-zones:false
            abbyy8.option.analyze-zones:true Option Count:5214 Word Count:5214 Abbyy8.cache-check.wrapper-
            version:1.15 width:5351 height:7250 xdpi:300 ydpi:300</processingStepSettings>
        - <processingSoftware>
            <softwareCreator>iArchives</softwareCreator>
            <softwareName>Abbyy FineReader</softwareName>
            <softwareVersion>8.1.0.6184-12</softwareVersion>
            <applicationDescription>iArchives OCR Framework [Abbyy FineReader 8.1.0.6184-12]</applicationDescription>
          </processingSoftware>
        </ocrProcessingStep>
      </OCRProcessing>
    </Description>
  - <Styles>
      <TextStyle ID="TS_10.0" FONTSIZE="10.0" />
      <TextStyle ID="TS_10.0_BI" FONTTYPE="serif" FONTWIDTH="proportional" FONTSIZE="10.0" FONTSTYLE="bold italics" />
      <TextStyle ID="TS_10.0_IP" FONTTYPE="serif" FONTWIDTH="proportional" FONTSIZE="10.0" FONTSTYLE="italics superscript" />
      <TextStyle ID="TS_10.0_I" FONTTYPE="serif" FONTWIDTH="proportional" FONTSIZE="10.0" FONTSTYLE="italics" />
      <TextStyle ID="TS_27.0" FONTSIZE="27.0" />
```

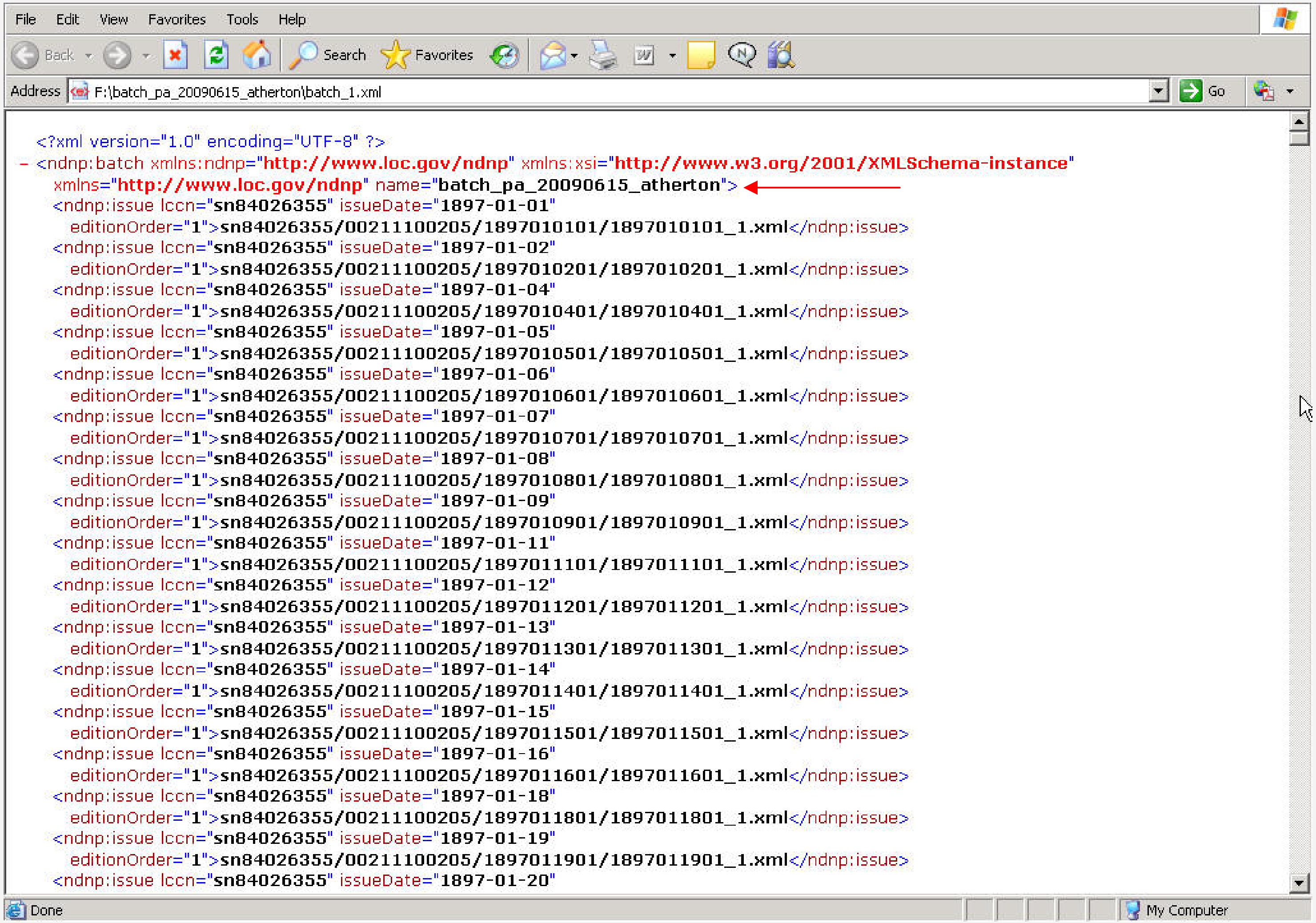
Structural Metadata

- Metadata Encoding and Transmission Standard (METS)
  - Developed at Library of Congress
  - XML standard
  - Many profiles for different object types
- NDNP data management – manifest XML
- Title, Issue, Reel, Essay Objects

Delivery: batch XML File

- Simple manifest
- Lists batch information – issues/reels
- Used for identification, validation, ingestion into digital repository system
- Example

Sample Batch XML File



Address F:\batch_pa_20090615_atherton\batch_1.xml

```
  <?xml version="1.0" encoding="UTF-8" ?>
- <ndnp:batch xmlns:ndnp="http://www.loc.gov/ndnp" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xmlns="http://www.loc.gov/ndnp" name="batch_pa_20090615_atherton">
    <ndnp:issue lccn="sn84026355" issueDate="1897-01-01"
      editionOrder="1">sn84026355/00211100205/1897010101/1897010101_1.xml</ndnp:issue>
    <ndnp:issue lccn="sn84026355" issueDate="1897-01-02"
      editionOrder="1">sn84026355/00211100205/1897010201/1897010201_1.xml</ndnp:issue>
    <ndnp:issue lccn="sn84026355" issueDate="1897-01-04"
      editionOrder="1">sn84026355/00211100205/1897010401/1897010401_1.xml</ndnp:issue>
    <ndnp:issue lccn="sn84026355" issueDate="1897-01-05"
      editionOrder="1">sn84026355/00211100205/1897010501/1897010501_1.xml</ndnp:issue>
    <ndnp:issue lccn="sn84026355" issueDate="1897-01-06"
      editionOrder="1">sn84026355/00211100205/1897010601/1897010601_1.xml</ndnp:issue>
    <ndnp:issue lccn="sn84026355" issueDate="1897-01-07"
      editionOrder="1">sn84026355/00211100205/1897010701/1897010701_1.xml</ndnp:issue>
    <ndnp:issue lccn="sn84026355" issueDate="1897-01-08"
      editionOrder="1">sn84026355/00211100205/1897010801/1897010801_1.xml</ndnp:issue>
    <ndnp:issue lccn="sn84026355" issueDate="1897-01-09"
      editionOrder="1">sn84026355/00211100205/1897010901/1897010901_1.xml</ndnp:issue>
    <ndnp:issue lccn="sn84026355" issueDate="1897-01-11"
      editionOrder="1">sn84026355/00211100205/1897011101/1897011101_1.xml</ndnp:issue>
    <ndnp:issue lccn="sn84026355" issueDate="1897-01-12"
      editionOrder="1">sn84026355/00211100205/1897011201/1897011201_1.xml</ndnp:issue>
    <ndnp:issue lccn="sn84026355" issueDate="1897-01-13"
      editionOrder="1">sn84026355/00211100205/1897011301/1897011301_1.xml</ndnp:issue>
    <ndnp:issue lccn="sn84026355" issueDate="1897-01-14"
      editionOrder="1">sn84026355/00211100205/1897011401/1897011401_1.xml</ndnp:issue>
    <ndnp:issue lccn="sn84026355" issueDate="1897-01-15"
      editionOrder="1">sn84026355/00211100205/1897011501/1897011501_1.xml</ndnp:issue>
    <ndnp:issue lccn="sn84026355" issueDate="1897-01-16"
      editionOrder="1">sn84026355/00211100205/1897011601/1897011601_1.xml</ndnp:issue>
    <ndnp:issue lccn="sn84026355" issueDate="1897-01-18"
      editionOrder="1">sn84026355/00211100205/1897011801/1897011801_1.xml</ndnp:issue>
    <ndnp:issue lccn="sn84026355" issueDate="1897-01-19"
      editionOrder="1">sn84026355/00211100205/1897011901/1897011901_1.xml</ndnp:issue>
    <ndnp:issue lccn="sn84026355" issueDate="1897-01-20"
```

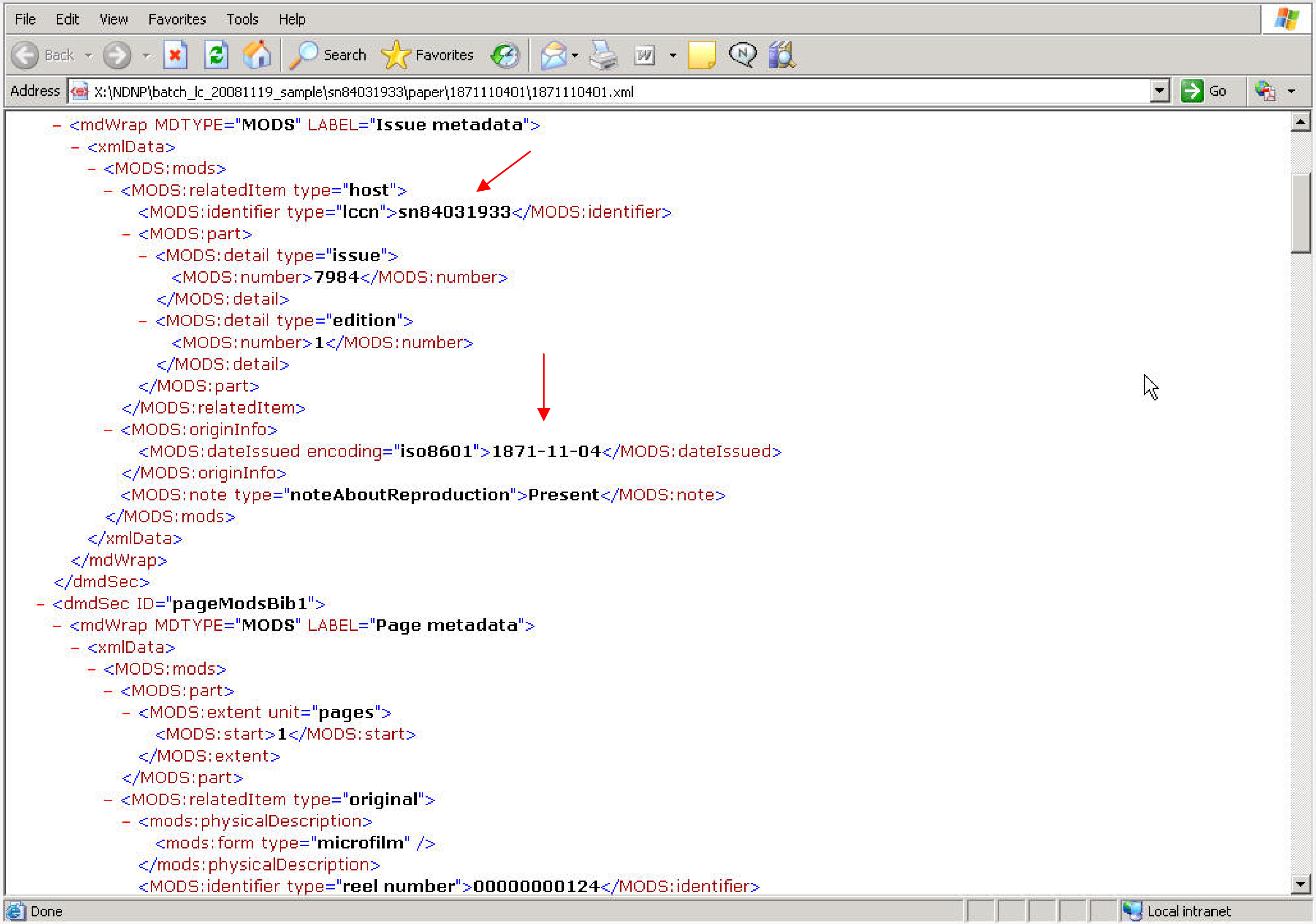
Title METS Object (Produced by LC)

- Produced and managed by LC from CONSER
- Typically CONSER-created, retrieved from OCLC
- Includes holdings records
- MARC to MARC XML transformation
- All objects have an LCCN
- LCCN is the unique identifier for each title

Issue METS Object

- Issue data
- Producer data
- Source data
- Individual page data rolled up into Issue METS
- Example
- During “validation” – PREMIS, MIX, and digital signatures are added with data derived from other files

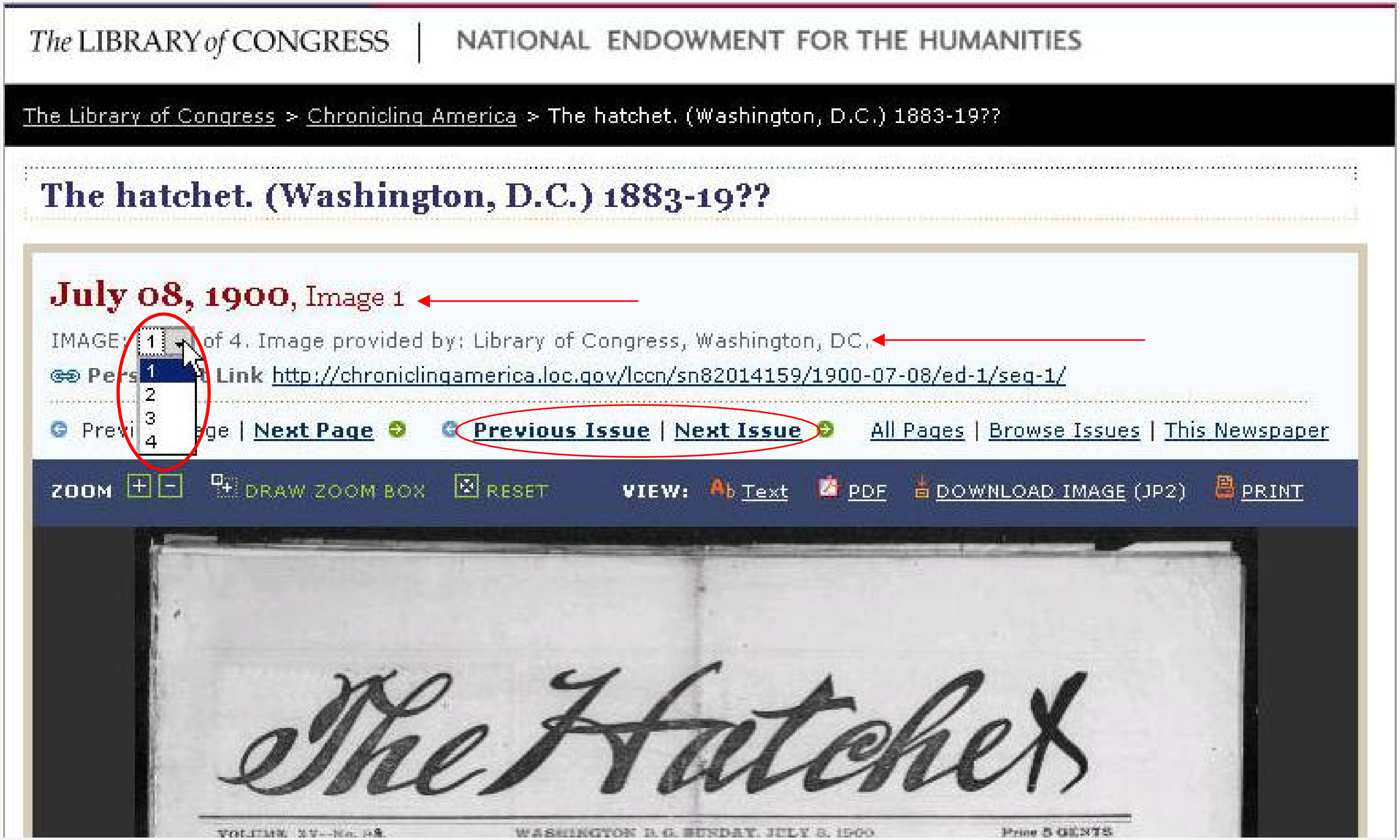
Issue XML



Address X:\NDNP\batch_lc_20081119_sample\sn84031933\paper\1871110401\1871110401.xml

```
  - <mdWrap MDTYPE="MODS" LABEL="Issue metadata">
    - <xmlData>
      - <MODS:mods>
        - <MODS:relatedItem type="host">
              <MODS:identifier type="lccn">sn84031933</MODS:identifier>
          - <MODS:part>
            - <MODS:detail type="issue">
                  <MODS:number>7984</MODS:number>
              </MODS:detail>
            - <MODS:detail type="edition">
                  <MODS:number>1</MODS:number>
              </MODS:detail>
            </MODS:part>
          </MODS:relatedItem>
        - <MODS:originInfo>
            <MODS:dateIssued encoding="iso8601">1871-11-04</MODS:dateIssued>
          </MODS:originInfo>
          <MODS:note type="noteAboutReproduction">Present</MODS:note>
        </MODS:mods>
      </xmlData>
    </mdWrap>
  </dmdSec>
- <dmdSec ID="pageModsBib1">
  - <mdWrap MDTYPE="MODS" LABEL="Page metadata">
    - <xmlData>
      - <MODS:mods>
        - <MODS:part>
          - <MODS:extent unit="pages">
              <MODS:start>1</MODS:start>
            </MODS:extent>
          </MODS:part>
        - <MODS:relatedItem type="original">
          - <mods:physicalDescription>
              <mods:form type="microfilm" />
            </mods:physicalDescription>
            <MODS:identifier type="reel number">00000000124</MODS:identifier>
```

Issue Data as Shown in Chronicling America



The LIBRARY of CONGRESS | NATIONAL ENDOWMENT FOR THE HUMANITIES

The Library of Congress > Chronicling America > The hatchet. (Washington, D.C.) 1883-19??

The hatchet. (Washington, D.C.) 1883-19??

July 08, 1900, Image 1

IMAGE: 1 of 4. Image provided by: Library of Congress, Washington, DC.

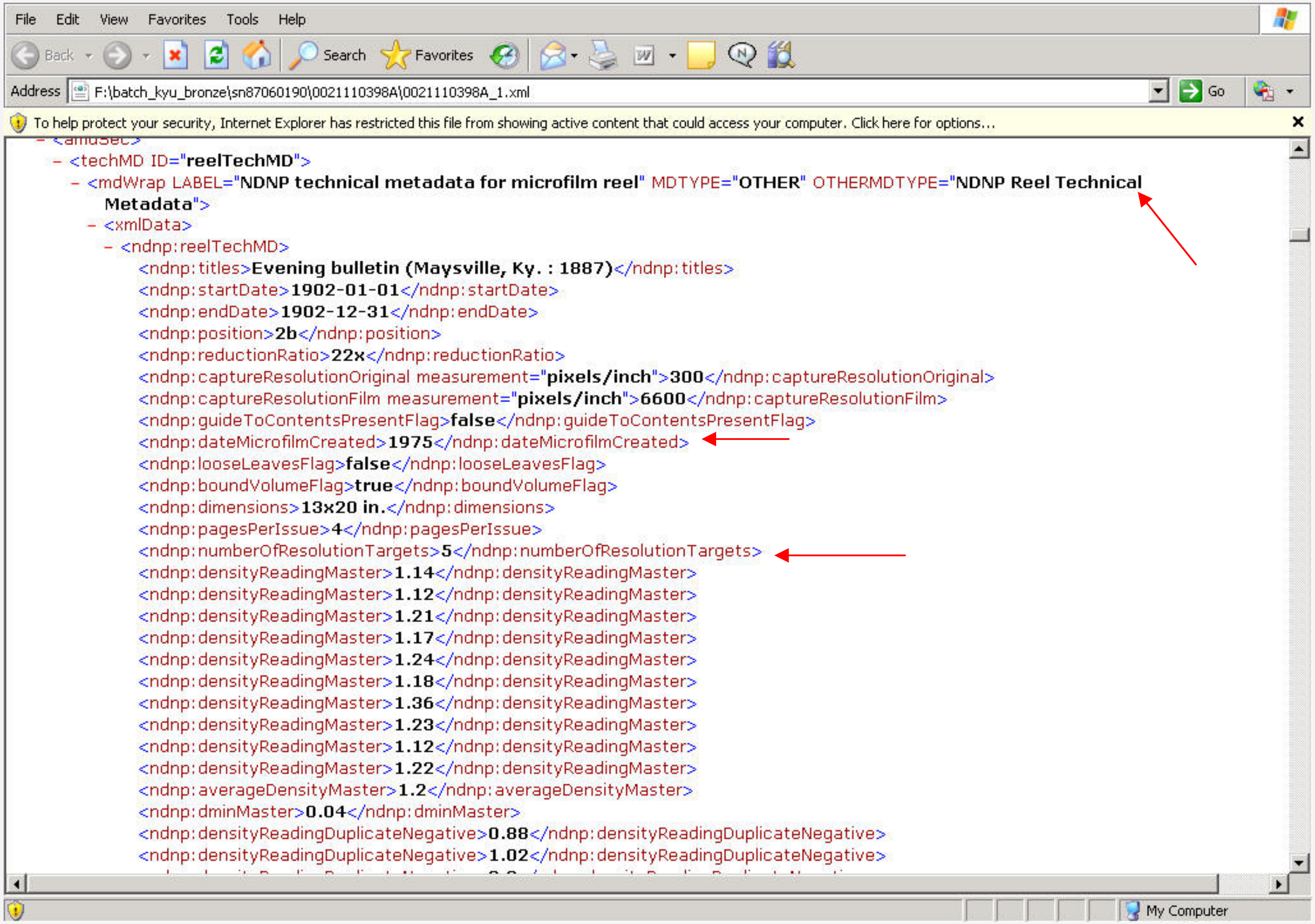
Pers Link http://chroniclingamerica.loc.gov/lccn/sn82014159/1900-07-08/ed-1/seq-1/

Previ ge | Next Page | Previous Issue | Next Issue | All Pages | Browse Issues | This Newspaper

Reel METS Object

- Reel data
- Records measured emulsion densities
- Measured resolution of original
- Technical target images
- Example

Reel XML

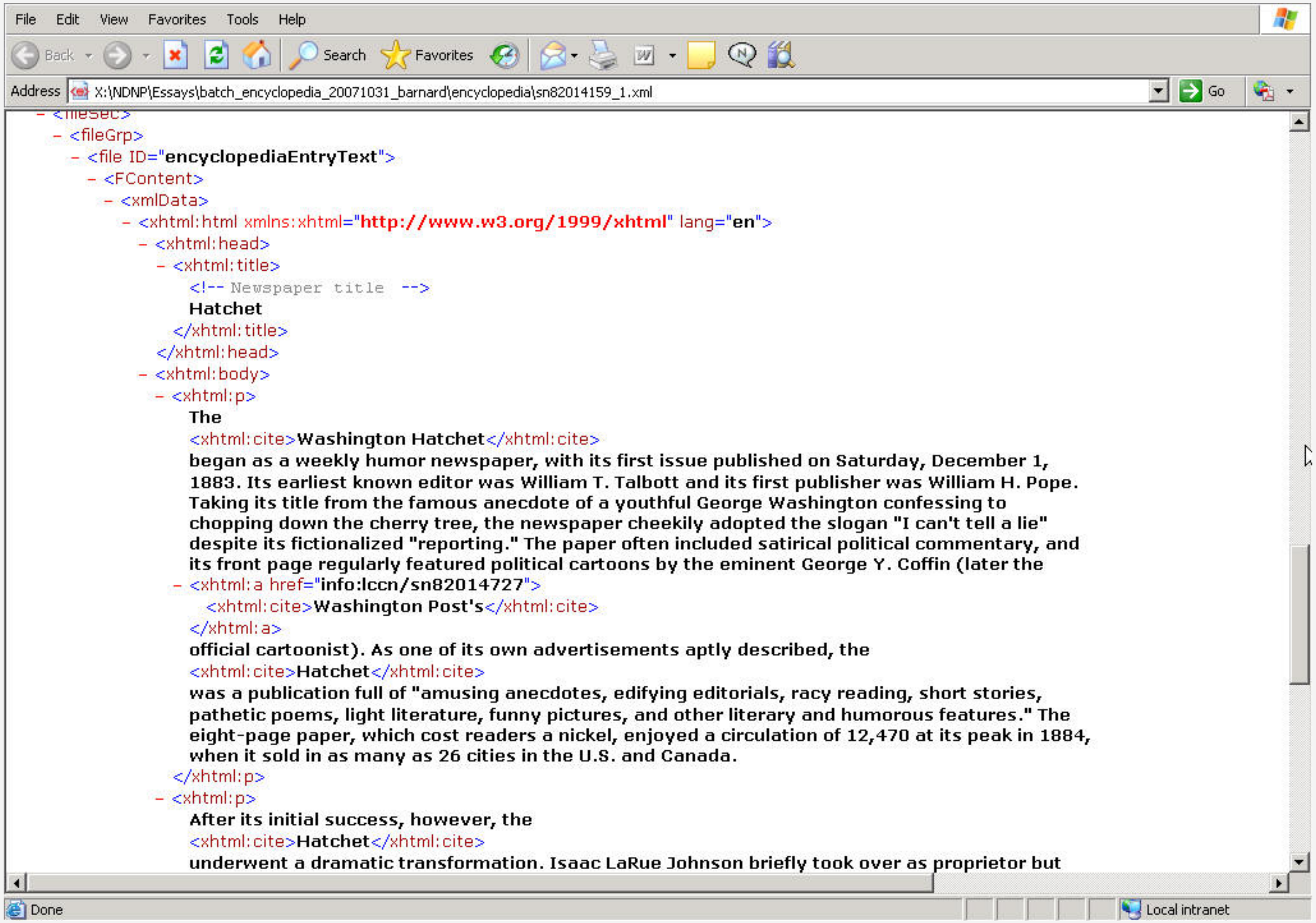

File Edit View Favorites Tools Help
Back
Search
Favorites
Address
F:\batch_kyu_bronze\sn87060190\0021110398A\0021110398A_1.xml
Go
To help protect your security, Internet Explorer has restricted this file from showing active content that could access your computer. Click here for options...
My Computer

```
- <amdSec>
  - <techMD ID="reelTechMD">
    - <mdWrap LABEL="NDNP technical metadata for microfilm reel" MDTYPE="OTHER" OTHERMDTYPE="NDNP Reel Technical
        Metadata">
      - <xmlData>
        - <ndnp:reelTechMD>
            <ndnp:titles>Evening bulletin (Maysville, Ky. : 1887)</ndnp:titles>
            <ndnp:startDate>1902-01-01</ndnp:startDate>
            <ndnp:endDate>1902-12-31</ndnp:endDate>
            <ndnp:position>2b</ndnp:position>
            <ndnp:reductionRatio>22x</ndnp:reductionRatio>
            <ndnp:captureResolutionOriginal measurement="pixels/inch">300</ndnp:captureResolutionOriginal>
            <ndnp:captureResolutionFilm measurement="pixels/inch">6600</ndnp:captureResolutionFilm>
            <ndnp:guideToContentsPresentFlag>false</ndnp:guideToContentsPresentFlag>
            <ndnp:dateMicrofilmCreated>1975</ndnp:dateMicrofilmCreated>
            <ndnp:looseLeavesFlag>false</ndnp:looseLeavesFlag>
            <ndnp:boundVolumeFlag>true</ndnp:boundVolumeFlag>
            <ndnp:dimensions>13x20 in.</ndnp:dimensions>
            <ndnp:pagesPerIssue>4</ndnp:pagesPerIssue>
            <ndnp:numberOfResolutionTargets>5</ndnp:numberOfResolutionTargets>
            <ndnp:densityReadingMaster>1.14</ndnp:densityReadingMaster>
            <ndnp:densityReadingMaster>1.12</ndnp:densityReadingMaster>
            <ndnp:densityReadingMaster>1.21</ndnp:densityReadingMaster>
            <ndnp:densityReadingMaster>1.17</ndnp:densityReadingMaster>
            <ndnp:densityReadingMaster>1.24</ndnp:densityReadingMaster>
            <ndnp:densityReadingMaster>1.18</ndnp:densityReadingMaster>
            <ndnp:densityReadingMaster>1.36</ndnp:densityReadingMaster>
            <ndnp:densityReadingMaster>1.23</ndnp:densityReadingMaster>
            <ndnp:densityReadingMaster>1.12</ndnp:densityReadingMaster>
            <ndnp:densityReadingMaster>1.22</ndnp:densityReadingMaster>
            <ndnp:averageDensityMaster>1.2</ndnp:averageDensityMaster>
            <ndnp:dminMaster>0.04</ndnp:dminMaster>
            <ndnp:densityReadingDuplicateNegative>0.88</ndnp:densityReadingDuplicateNegative>
            <ndnp:densityReadingDuplicateNegative>1.02</ndnp:densityReadingDuplicateNegative>
```

Newspaper History Essays

- Associates with Title METS object
- < 500 words
- History and significance of title
- Embedded links to other titles, as needed

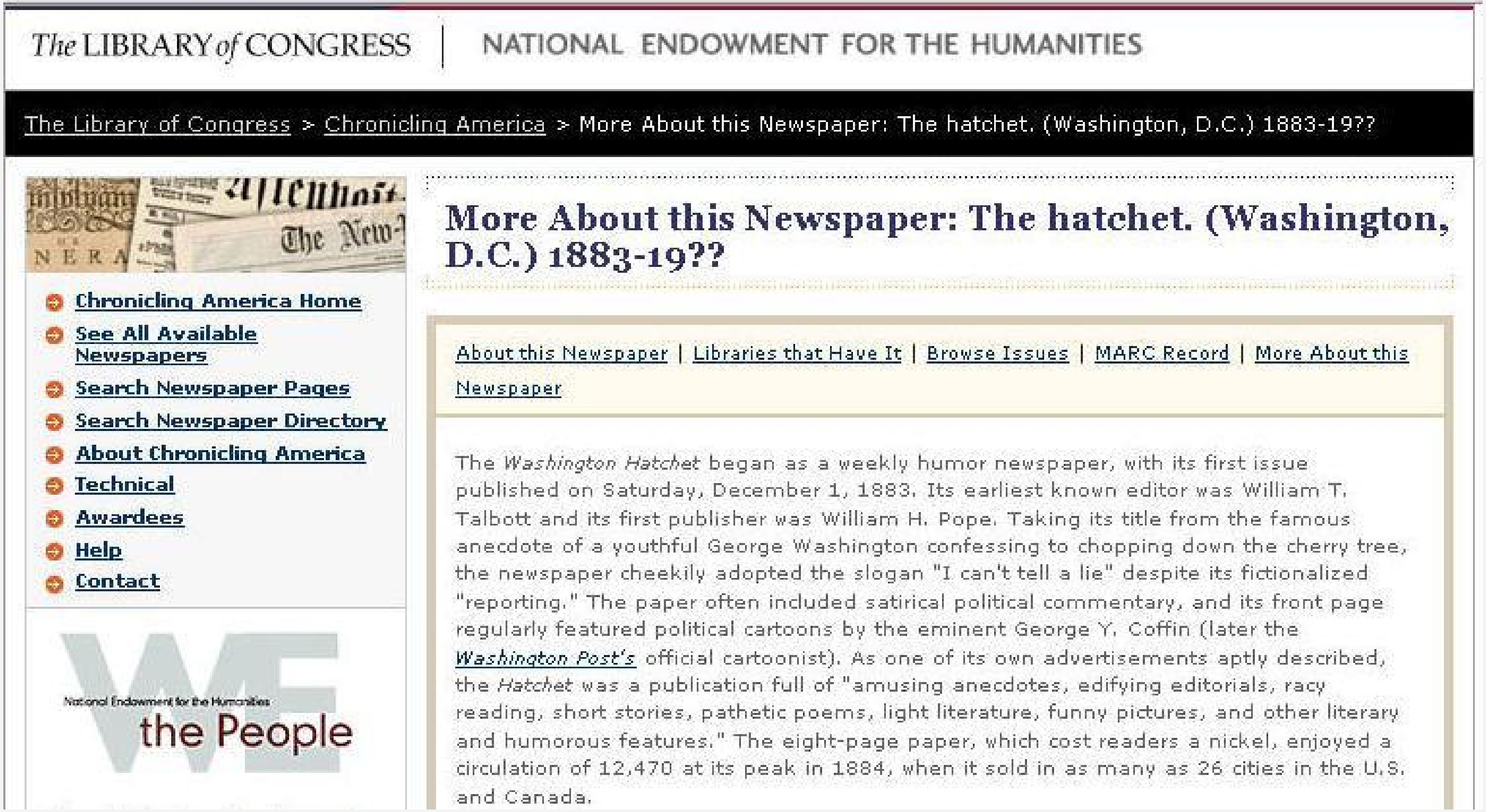
Essay METS XML (produced by LC)



Address X:\NDNP\Essays\batch_encyclopedia_20071031_barnard\encyclopedia\sn82014159_1.xml

```
- <fileSec>
  - <fileGrp>
    - <file ID="encyclopediaEntryText">
      - <FContent>
        - <xmlData>
          - <xhtml:html xmlns:xhtml="http://www.w3.org/1999/xhtml" lang="en">
            - <xhtml:head>
              - <xhtml:title>
                  <!-- Newspaper title  -->
                  Hatchet
                </xhtml:title>
              </xhtml:head>
            - <xhtml:body>
              - <xhtml:p>
                  The
                  <xhtml:cite>Washington Hatchet</xhtml:cite>
                  began as a weekly humor newspaper, with its first issue published on Saturday, December 1,
                  1883. Its earliest known editor was William T. Talbott and its first publisher was William H. Pope.
                  Taking its title from the famous anecdote of a youthful George Washington confessing to
                  chopping down the cherry tree, the newspaper cheekily adopted the slogan "I can't tell a lie"
                  despite its fictionalized "reporting." The paper often included satirical political commentary, and
                  its front page regularly featured political cartoons by the eminent George Y. Coffin (later the
                - <xhtml:a href="info:lccn/sn82014727">
                    <xhtml:cite>Washington Post's</xhtml:cite>
                  </xhtml:a>
                  official cartoonist). As one of its own advertisements aptly described, the
                  <xhtml:cite>Hatchet</xhtml:cite>
                  was a publication full of "amusing anecdotes, edifying editorials, racy reading, short stories,
                  pathetic poems, light literature, funny pictures, and other literary and humorous features." The
                  eight-page paper, which cost readers a nickel, enjoyed a circulation of 12,470 at its peak in 1884,
                  when it sold in as many as 26 cities in the U.S. and Canada.
                </xhtml:p>
              - <xhtml:p>
                  After its initial success, however, the
                  <xhtml:cite>Hatchet</xhtml:cite>
                  underwent a dramatic transformation. Isaac LaRue Johnson briefly took over as proprietor but
```

Essays in Chronicling America


The LIBRARY of CONGRESS
NATIONAL ENDOWMENT FOR THE HUMANITIES
The Library of Congress > Chronicling America > More About this Newspaper: The hatchet. (Washington, D.C.) 1883-19??
Chronicling America Home
See All Available Newspapers
Search Newspaper Pages
Search Newspaper Directory
About Chronicling America
Technical
Awardees
Help
Contact
National Endowment for the Humanities
We the People
About this Newspaper | Libraries that Have It | Browse Issues | MARC Record | More About this Newspaper

More About this Newspaper: The hatchet. (Washington, D.C.) 1883-19??

The *Washington Hatchet* began as a weekly humor newspaper, with its first issue published on Saturday, December 1, 1883. Its earliest known editor was William T. Talbott and its first publisher was William H. Pope. Taking its title from the famous anecdote of a youthful George Washington confessing to chopping down the cherry tree, the newspaper cheekily adopted the slogan "I can't tell a lie" despite its fictionalized "reporting." The paper often included satirical political commentary, and its front page regularly featured political cartoons by the eminent George Y. Coffin (later the *Washington Post's* official cartoonist). As one of its own advertisements aptly described, the *Hatchet* was a publication full of "amusing anecdotes, edifying editorials, racy reading, short stories, pathetic poems, light literature, funny pictures, and other literary and humorous features." The eight-page paper, which cost readers a nickel, enjoyed a circulation of 12,470 at its peak in 1884, when it sold in as many as 26 cities in the U.S. and Canada.

Validation and Assessment

NDNP Validation Library

- Java library for validating batches, issues, reels, TIFFs, PDFs, JPEG2000, and ALTO
- Extends validation capabilities of JHOVE1
- Digitally signs files as having passed validation
- Adds technical metadata to METS
- Can be run from command line or embedded in other applications

NDNP Digital Viewer and Validation Toolkit

- Integrates Validation Library with Graphic Interface for subjective quality assessment of content
- Embedded viewers for all file formats/metadata across objects.
- Visual display of file relationships as objects (titles, issues, reels)
- Distributed to all program participants

NDNP Deliverables for Each Award

- Summary of all Deliverables delivered to LC per award*
  - Validated digital objects per specification – approx.100,000 pages
  - Associated newspaper history essays for each title digitized
  - Updated MARC records for each title digitized
  - Duplicate print negatives (2n) microfilm used for digitization

*Refer to NDNP Program Web site (http://www.loc.gov/ndnp/) for updates.

Resources

- http://www.loc.gov/ndnp/
- http://www.digitizationguidelines.gov