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
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 OCR red text to image coordinates
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
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
XML code representing issue metadata:

```xml
<mdWrap MDTYPE="MODS" LABEL="Issue metadata">
  <xmlData>
    <MODS:mods>
      <MODS:relatedItem type="host">
        <MODS:identifier type="lccn">sn84031933</MODS:identifier>
      </MODS:relatedItem>
      <MODS:part>
        <MODS:detail type="issue">
          <MODS:number>7984</MODS:number>
        </MODS:detail>
      </MODS:part>
      <MODS:part>
        <MODS:detail type="edition">
          <MODS:number>1</MODS:number>
        </MODS:detail>
      </MODS:part>
      <MODS:part>
        <MODS:relatedItem>
          <MODS:originInfo>
            <MODS:dateIssued encoding="iso8601">1071-11-04</MODS:dateIssued>
          </MODS:originInfo>
          <MODS:note type="noteAboutReproduction">Present</MODS:note>
        </MODS:relatedItem>
      </MODS:part>
    </xmlData>
  </mdWrap>
</mdWrap>
```

XML code representing page metadata:

```xml
<mdSec ID="pageModsbib1">
  <xmlData>
    <MODS:mods>
      <MODS:part>
        <MODS:extent unit="pagos">
          <MODS:start>1</MODS:start>
        </MODS:extent>
      </MODS:part>
      <MODS:part>
        <MODS:relatedItem type="original">
          <mods:physicalDescription>
            <mods:form type="microfilm" />
          </mods:physicalDescription>
        </MODS:relatedItem>
      </MODS:part>
      <MODS:identifier type="real number">00000000124</MODS:identifier>
    </xmlData>
  </mdSec>
```

Reel METS Object

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

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

After its initial success, however, the Hatchet underwent a dramatic transformation. Isaac Lauer Johnson briefly took over as proprietor but...
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