Technical Implementation Specifications

Justin Littman, MLIS

Office of Strategic Initiatives

Dave Woodward,

Information Technology Services/Office of
Strategic Initiatives

Philosophy

- Avoid "garbage in, garbage out"
- "Trust, but verify"
  - Validate early
  - Validate once
- Use existing standards

Archival Master: TIFF

- Conforms with TIFF 6.0
- 8-bit greyscale
- ~400 dpi
- Uncompressed
- Only deskewing should be applied
- Cropped to page edge
- Additional TIFF tags required

JPEG 2000

- JPEG 2000 is a new image coding system that uses state-of-the-art compression techniques based on wavelet technology.

- JPEG 2000 was designed to provide high image quality at low bit rates.

- JPEG 2000 employs progressive rendering and transmission through a layered image file structure. This allows many views/resolutions (thumbnail to full-size) to be delivered from a single compressed file.

JPEG 2000, cont.

- JPEG2000 is a result of collaboration between:
  - The International Organization for Standardization (ISO),
  - The International Telecommunications Union (ITU-T, formerly CCITT), and
  - industry experts from a variety of companies.

JPEG 2000 and Aware

- Image delivery in the initial NDNP system will utilize JPEG 2000 (.jp2) files and viewing services created with a software development kit (SDK) from Aware, Inc.

- Aware, a provider of wavelet-based codecs since 1987, has been actively involved in the development of JPEG 2000.

JPEG 2000 and Aware

- Benchmark testing found the performance and functionality of software created with the Aware SDK to be superior to software from the open-source initiative, *The JasPer Project*.

- JPEG 2000 and Aware SDK were recently adopted as a compression and delivery mechanism for large-format/high-resolution images at LC, e.g.,
  - In American Memory: *The Rochambeau Maps Collection*, and
  - in Global Gateway: *Polish Declarations of Admiration and Friendship for the United States.*

Production Master: 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

Derivative: PDF

- Compatible with Acrobat 5.0 (PDF 1.4)
- Image with text behind
- Image will be a greyscale, 150dpi JPEG, using a medium (or 40) quality setting
- XMP/RDF/Dublin Core metadata

OCR text: ALTO

- Conforms with ALTO (Analyzed Layout and Text Object) schema
- ALTO is product of EU-funded METAe project
- Mapping of OCRed text to image coordinates
- Tools available for conversion from ABBYY XML and PrimeRecognition PRO

Tools

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

Tools cont.

- NDNP Desktop Viewer and Validator
  - GUI for performing validation and examining files
  - Under development
- ALTO converters
  - ABBYY XML to ALTO
  - PrimeRecognition PRO to ALTO
  - Require further testing