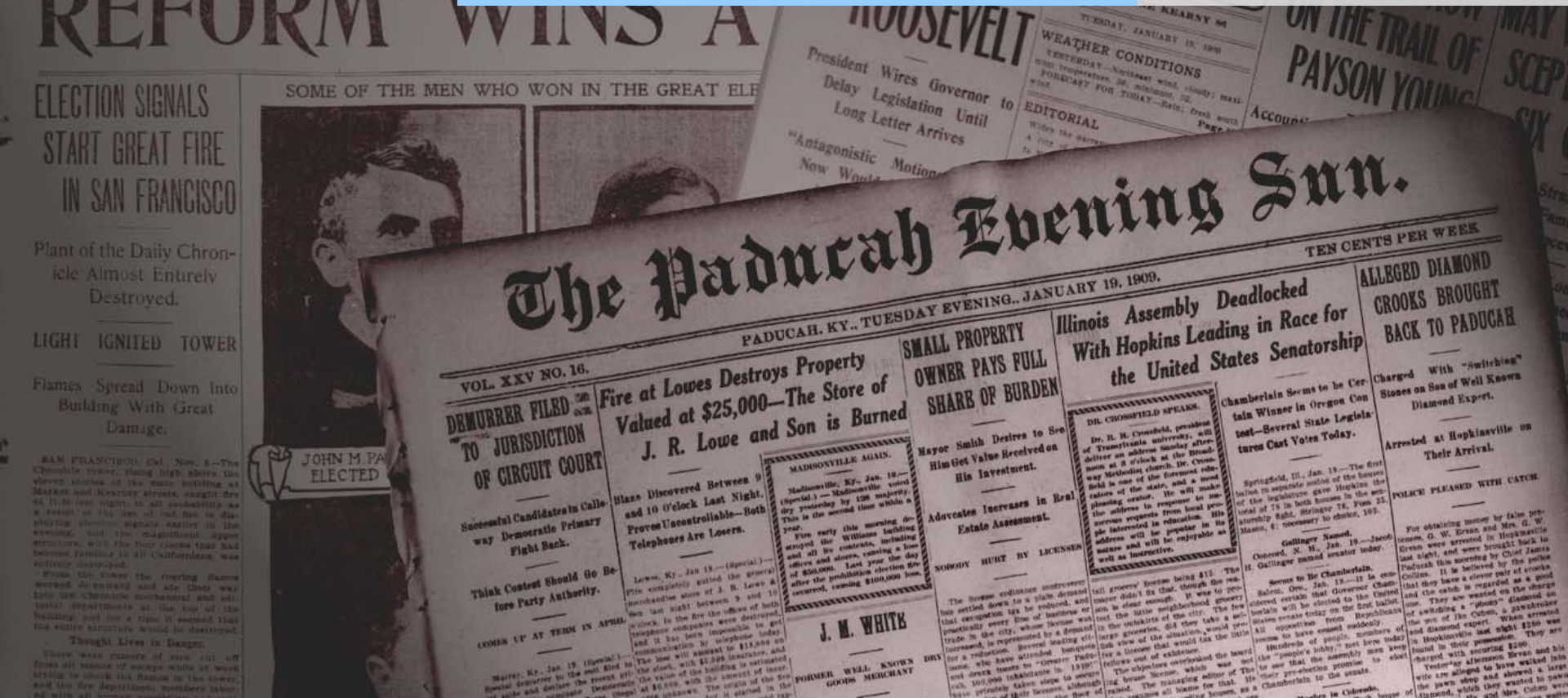
NATIONAL ENDOWMENT FOR THE Humanities

LIBRARY OF CONGRESS

JPEG2000 and the National Digital Newspaper Program

NATIONAL ENDOWMENT FOR THE HUMANITIES
and **LIBRARY OF CONGRESS**

Deborah Thomas
US Library of Congress



The National Digital Newspaper Program

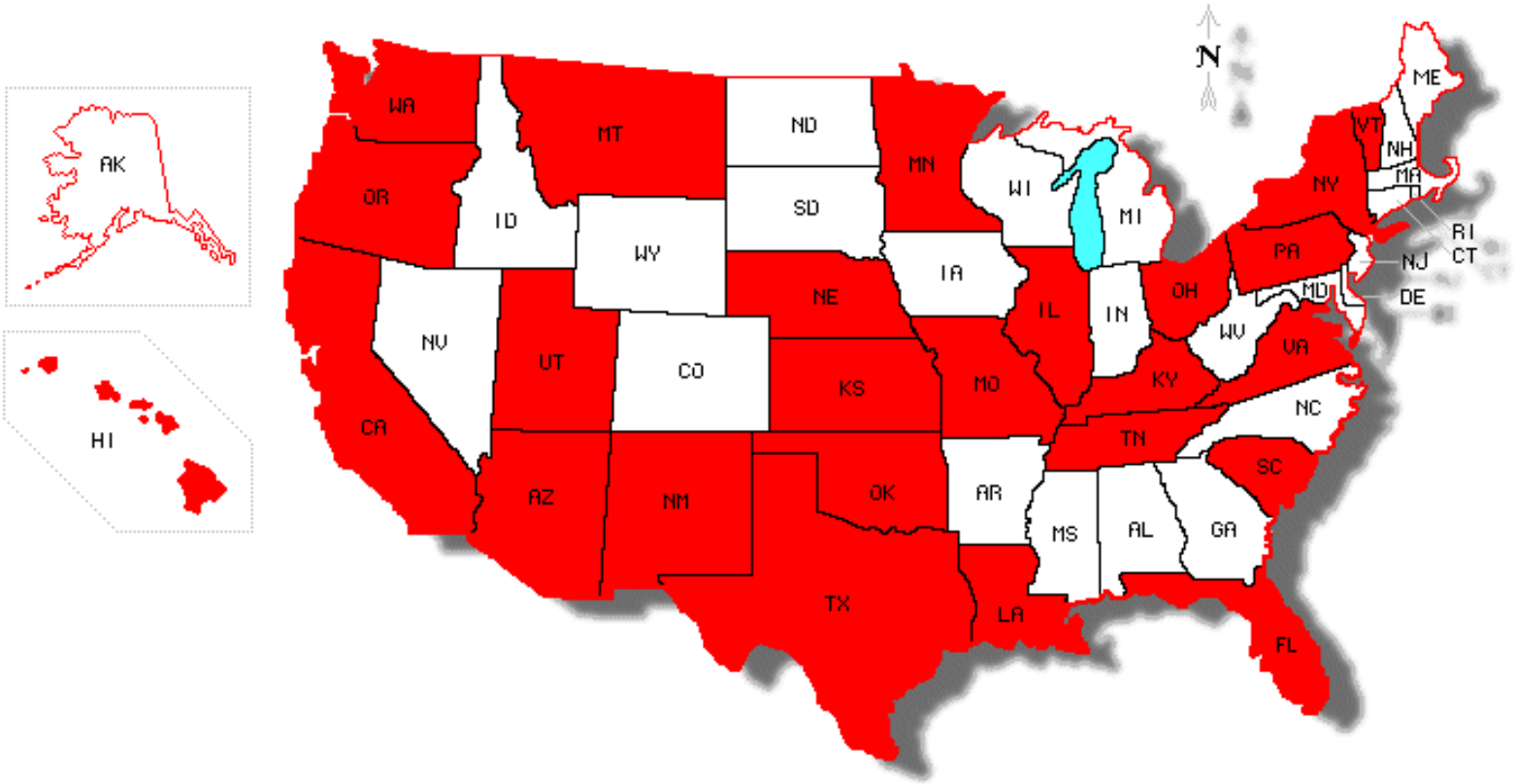
GOALS:

- To enhance access to historic American newspapers
- To develop best practices for the digitization of historic newspapers
- To apply emerging technologies to the products of USNP (United States Newspaper Program, 1984-2010)
  - 140,000 titles cataloged,
  - 900,000 holding records created,
  - more than 75 million pages filmed

The National Digital Newspaper Program

- NEH grants ***2-year awards (up to $400k) to state projects,*** to select and digitize historic newspapers for full-text access (100,000 pages per award).
- LC creates and ***hosts Chronicling America Web site*** to provide freely accessible search and discovery for digitized papers and descriptive newspaper records.
- State projects ***repurpose NDNP contributions for local purposes***, as desired.

PARTNERS:
24 institutions | >4 million pages by 2012 | 1836-1922



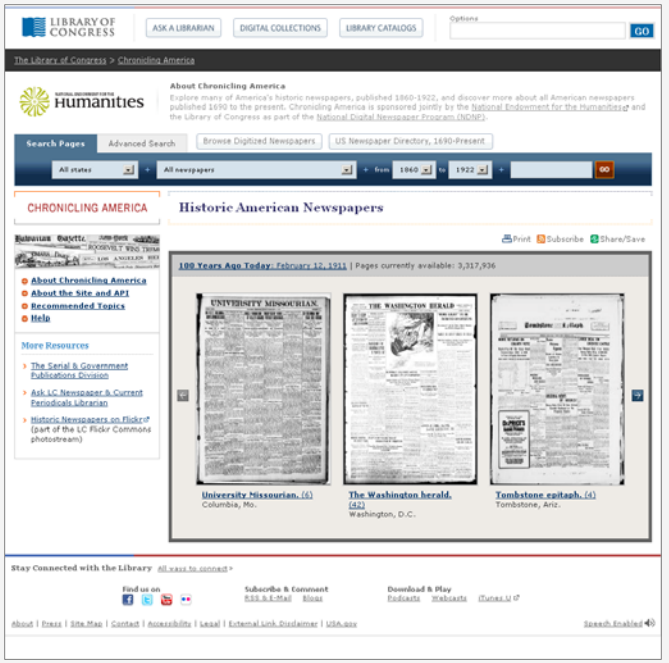
NATIONAL ENDOWMENT FOR THE Humanities

LIBRARY OF CONGRESS

Chronicling America: Historic American Newspapers

- >3.7 million pages
- 1859-1922
- >500 titles from 22 states and DC
- http://chroniclingamerica.loc.gov/

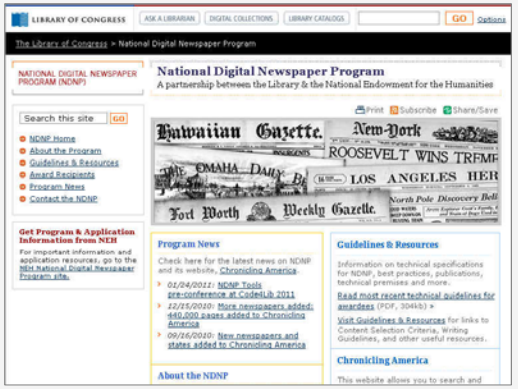
- Awards 2005-2010
  - *2005 awards* - CA, FL, KY, NY, UT, VA (1900-1910)
  - *2007 awards* - CA, KY, MN, NE, NY, TX, UT, VA (1880-1910)
  - *2008 awards* - AZ, HI, MO, OH, PA, WA (1880-1922)
  - *2009 awards* - IL, KS, LA, MT, OK, OR, SC (1860-1922)
  - *2010 awards* - AZ, HI, MO, NM, OH, PA, TN, VT, WA (1836-1922)
  - and onward! (next awards announced July 2011)



- Coming Soon:
  - Content from states added in 2010 (New Mexico, Tennessee, Vermont)
  - Newspapers from 1836-1859

Beyond NDNP

- Data specifications in use beyond NDNP
  - NDNP Guidelines - http://www.loc.gov/ndnp/guidelines/
  - Federal Agencies Digitization Guidelines Initiatives - http://www.digitizationguidelines.gov/
  - National Libraries - METS/ALTO used in
    - UK, France, Australia, New Zealand, Austria, Norway, Slovenia, Slovak Republic …
- Open-Source Software Development –
  - LC Newspaper Viewer, available on Sourceforge.net - http://sourceforge.net/projects/loc-ndnp/
  - LC Newspaper Viewer in Action
    - e.g., Oregon Historical Newspapers - http://oregonnews.uoregon.edu/
    - Other awardees and interested parties working on software development collaboration through SourceForge



Working with Historic Newspapers – Image Characteristics

- Scanned from microfilm 2n negatives
- Large format, little tiny type/varying type quality
- Changes in print technology over time – type, illustrations
- Varying quality: print (paper) and film (lighting, focus, process)
- Damage (acid paper, exposure, handling, etc.)
- Color space is grayscale rather than truly “high contrast” (bitonal)

NDNP Data Specifications

- Should be as simple as is practical, producible with current technology
- Data to be created by multiple producers/vendors, and be aggregated into LC infrastructure
- Support desired research functions of the system
- Support enduring access

DIGITAL OBJECT - Issue

- Archival Image: TIFF
- Production Image: JPEG 2000
- Printable Image: PDF
- ALTO XML for OCR
- METS with MODS/PREMIS/MIX metadata objects (issue/reel)

JPEG2000 in NDNP

- Specification derived from "JPEG 2000 Profile for the National Digital Newspaper Program" Report, April 2006 (Prepared by: Robert Buckley and Roger Sam)

- 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
- Identifying RDF/Dublin Core metadata in XML box
- See NDNP JPEG2000 v2.7 profile - http://www.loc.gov/ndnp/guidelines/archive/JPEG2kSpecs09.pdf

Benefits and Challenges of working with JPEG2000

- BENEFITS
  - Format is free to use
  - Efficient compression (limited)
  - Data transfer efficiency for access
  - Supports tiling and efficient transformation supporting pan/zoom Web functions
  - Used for production, reduces amount of storage needed on access servers

- CHALLENGES
  - Complex format, little forgiveness
  - Complex specification, not available to the public
  - Patent encumbered specification
  - Commercial tool support – expensive and inconsistent
  - Open-source tool support – limited in both conformance and performance

Uses and Alternatives

- How JPEG2000 are used in NDNP:
  - Used in “production” role: used to export JPEG files to Web browser, supports “pan/zoom” behavior; available for download (compact file size)
  - Aware Imaging Library from Python (wrote code)

- Alternatives in use by NDNP Awardees:
  - Direct delivery of JPEG (browser native)
  - Pre-tiled single file in any format (PNG, JPEG, GIF)
  - Lossless compressed TIF (LZW)
  - Dynamic, cached delivery of derivatives (PNG, JPEG, GIF)

Thank you!

- NDNP Public Web http://www.loc.gov/ndnp/
- NDNP Web Service
  Chronicling America: Historic American Newspapers http://chroniclingamerica.loc.gov
- Contact us at ndnptech@loc.gov
- Technical contact: dbrun@loc.gov