The National Digital Newspaper Program (NDNP)

An NEH/LC Collaborative Program

Enhancing access to American historical newspapers

Release: September 2006
• Enhance access to all American newspapers
• Improve products of United States Newspaper Program (USNP) using current technologies
• Establish standards and “best practices” for newspaper digital reformatting and access
• Develop geographically-diverse program that benefits all US communities
• Use multi-phased approach for research and scaled development
Why Newspapers?

• Newspapers = fundamentals of history
• No single U.S. collection
• Enormous corpus = archival and access challenge
• Text-intensive layout
• Digitization of microfilmed corpus economically feasible for otherwise difficult physical material
What We Know

- Broad user-base, high demand for access
- Newspaper format challenges
- 100 million + frames of newspaper microfilm created
- Historical newspapers do not follow contemporary “layout” rules
  - Automated analysis tools are prone to error
What We Don’t Know

• Economic models for developing and maintaining digital access
• Microfilm analysis - what’s worth digitizing (research value, quality, OCR)
• Best practices and standards (digital) for newspaper description and text
  – Several commercial approaches - proprietary, not interoperable
• OCR quality and search access needed for millions of pages - what’s necessary?
• Technical issues and costs of repository development and migration
What will NDNP Produce?

• Web access to
  – National directory of US newspaper holdings (what, when, where) – based on USNP legacy data
  – Millions of page images of historical newspapers digitized primarily from microfilm, with full-text
  – Historical context of newspapers

• Depository of duplicate digitized microfilm at LC
How?

• Multi-partner, phased program
  – NEH: Funds the program ("We the People" initiative)
  – LC: Aggregates, preserves and serves
  – Awardees: Select and convert

• Phase I – FY04-FY07 (Pilot)
  • 2005 NEH awardees (6 for $1.9 million) with existing digital collections infrastructure and master microfilm negatives
  • 100,000 pages ea + 100,000 LC pages by 2007 (from 1900-1910)
  • Microfilm reel data for future analysis
  • Incorporate use of open standards and software dev
  • Determine resource needs for production system and future phases
Newspaper Title Directory

• Re-use of CONSER and Newspaper Union List, created under USNP (maintained by OCLC)
  • 147,000 newspaper titles
  • 900,000 holdings records

• Searchable, Web access to all USNP-collected data, linked to digitized issues when available (and newspaper Web sites when linked through CONSER)
Historical Context

Curatorial Input to Repository

• Brief essays for each title digitized
  – Publisher, geography, significant events covered, audience/community, politics

• History of newspaper publishing in each state, effect on state or national events
Full Text with Page-level Access

- Basic level of access served
- Preserves integrity of primary historical content, text in context
- Minimal metadata and human interaction required
- Economics of large-scale, large-format digitization
- Move ahead with creation of substantial content-base for R&D in search and presentation
Digital Asset Specifications

- Page Image - grayscale, 400 dpi, from microfilm
  - TIFF 6.0; JPEG 2000 (.jp2); PDF with Hidden Text

- OCR
  - XML – NDNP/ALTO Schema
  - Page-level, uncorrected, column zones with “bounding box” mapping coordinates

- Metadata
  - XML in METS/MODS for digital objects
• Web access - *American Chronicle*
  – Newspaper Title Directory, 1690-present
  – Full-text of content w/in visual newspaper layout (page-level access)
  – Contextual historical material

• Converted content from all awardees
  – Initial time period covered: 1900-1910
  – 250,000 pages at launch
For information, see NDNP Overview at http://www.neh.gov/projects/ndnp.html

For technical information, see http://www.loc.gov/ndnp/ or contact ndnptech@loc.gov.