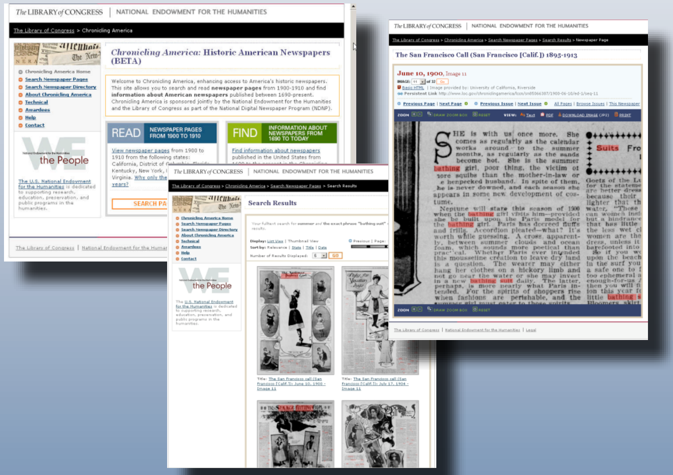
Mitigating Preservation Threats:

Standards and Practices in the National Digital Newspaper Program

The Colored American
NATION'S LAST TRIBUTES OF LOVE AND GRIEF.
THURSDAY TO BE MEMORIAL DAY.
GREAT CONCERN FOR ROOSEVELT.
The San Francisco Call.
HALF MILLION MOURNING AMERICANS VIEW CASKET CONTAINING DEAD CHIEF EXECUTIVE DURING THE JOURNEY TO WASHINGTON
REMAINS OF THE MARTYRED PRESIDENT BORNE PAST MANY THOUSAND MOURNERS


The LIBRARY of CONGRESS | NATIONAL ENDOWMENT FOR THE HUMANITIES

The Library of Congress > Chronicling America

Chronicling America: Historic American Newspapers (BETA)

- Chronicling America Home
- Search Newspaper Pages
- Search Newspaper Directory
- About Chronicling America
- Technical
- Awardees
- Help
- Contact

Welcome to Chronicling America, enhancing access to America's historic newspapers. This site allows you to search and read **newspaper pages** from 1900-1910 and find **information about American newspapers** published between 1690-present. Chronicling America is sponsored jointly by the National Endowment for the Humanities and the Library of Congress as part of the National Digital Newspaper Program (NDNP).

READ NEWSPAPER PAGES FROM 1900 TO 1910

View newspaper pages from 1900 to 1910 from the following states: California, District of Columbia, Florida, Kentucky, New York, Utah, Virginia. Why only these years?

FIND INFORMATION ABOUT NEWSPAPERS FROM 1690 TO TODAY

Find information about newspapers published in the United States from

The U.S. National Endowment for the Humanities is dedicated to supporting research, education, preservation, and public programs in the humanities.

The Library of Congress > Chronicling America > Search Newspaper Pages > Search Results

Search Results

Your fulltext search for summer and the exact phrase "bathing suit"

Display: List View | Thumbnail View

Sort by: Relevance | State | Title | Date

Number of Results Displayed: 6 GO

Title: The San Francisco call (San Francisco [Calif.]): June 10, 1900 - Image 11

Title: The San Francisco call (San Francisco [Calif.]): July 17, 1904 - Image 11

The Library of Congress > Chronicling America > Search Newspaper Pages > Search Results > Newspaper Page

The San Francisco Call (San Francisco [Calif.]) 1895-1913

June 10, 1900, Image 11

IMAGE: 11 of 32 Go

Basic HTML | Image provided by: University of California, Riverside

Persistent Link http://www.loc.gov/chroniclingamerica/lccn/sn85066387/1900-06-10/ed-1/seq-11

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

ZOOM DRAW ZOOM BOX RESET VIEW: Text PDF DOWNLOAD IMAGE (JP2) PRINT

The Library of Congress | National Endowment for the Humanities | Legal

Deborah Thomas and David Brunton, Library of Congress

Digital Library Federation - Spring Forum

Minneapolis, MN – April 29, 2008

Preservation Threats

- Organizational threats
- Errors (creation, conversion)
- Obsolescence (formats, hardware)
- Failure (hardware, humans)

- Rosenthal, David S. H., Thomas Robertson, Tom Lipkis, Vicky Reich, Seth Morabito. Requirements for Digital Preservation Systems: A Bottom-Up Approach. *D-Lib Magazine*. <doi:10.1045/november2005-rosenthal>
- Littman, Justin. "Actualized Preservation Threats: Practical Lessons from Chronicling America." D-Lib Magazine <doi:10.1045/july2007-littman>

OVERVIEW

THE SALT LAKE HERALD.

STATE CHANGES FRONT IN THE BENBROOK CASE

WHITEHEAD HAD

PORTO RICAN BILL WINS

CAMPBELL WENT

LADYSMITH RELIEVED

THE BEE

WASHINGTON

The Call Prints More News Than Any Other Paper Published in San Francisco

THE CALL

TERRIFIC HURRICANE SWEEPS ISLANDS OF SUNNY SOUTH SEAS AND TAHITI SUFFERS SEVERELY.

ONLY TWO LIVES ARE LOST IN VISITATION

RAIN, SNOW AND HAIL FALL AND WIND BLOWS

TORNADO AT MERIDIAN CAUSES HEAVY DAMAGE

PAPEETE IS BADLY WRECKED

Islands Are Stripped by Sea to Coral Foundations.

Cocoanut Palm Trees Uprooted by the Wholesale.

Suffering May Follow Later From Lack of Food.

French Government Is Heaviest Loser at Tahiti.

NINETEEN DEAD TAKEN FROM RUINS

Belief That More Victims Are in Wreckage.

All Buildings in Path of the Storm Are Demolished.

SANTA CLARA VALLEY IN PATH OF STORM.

Chilly Blasts Sweep Through Orchards and Petting Showers Hit Blossoms.

NORSE FISHING FLEET LOST IN FIERCE STORM

Little Boats Overturned by Mountainous Seas During a Snow Squall

- The Program
- Format Challenges
- Data Specifications
- Repository Development
- Mitigating Threats
- Enduring Access

The Many Facets of the NDNP

- Enhancing access to historic American newspapers
- Cooperative national digitization effort (LC/NEH/state-level institutions)
- Digital preservation in support of enduring access
- National resource free to use and reuse

Historic Newspapers: Old Problems

- challenges to using historic newspapers
  - enormous corpus of content
  - Characteristics: hierarchical structure, multiple information elements, layouts
  - very little indexing
  - limitations of physical formats
  - far-flung collections (distributed holders)
  - competition for use

Historic Newspapers: New Opportunities

- Large scale collection in digital form
  - Chronologically, geographically diverse
- Reuse of USNP Products
  - bibliographic info, location info, film
- Cross-search content
  - By dates, keywords
- High-quality images
  - Grayscale, resolution
- Available to multiple users, worldwide, 24/7

Program Infrastructure

Mitigate organizational threats to preservation

- NEH/LC successful partnership for USNP
- LC experience with digitization
- Roles: NEH, LC, awardee institutions
- Phased development

NDNP: Technical Challenges

- How do we ensure good, consistent content produced over time? (*Scope*)
- What do we have to do to be able to support 54 producers and aggregate the content over the course of 20 years? (*Scale*)
- How do we provide enduring access? (*Sustain*)

Good Digital Objects for NDNP

Mitigate preservation threats thru standards and validation

- Should be as simple as is practical, producible with current technology (multi-producer)
- “Known” before and after aggregated into LC infrastructure (apples and apples)
- Specifications should support desired research functions of the system
  - Find it, Browse it, View it
- Provide enduring access

View (and Preserve): Images

- Archival Image: TIFF
  - Conforms with TIFF 6.0, metadata
  - 8-bit grayscale, 400 dpi uncompressed
- Production Image: JPEG 2000
  - JPEG 2000, Part 1 (.jp2), lossy 8:1
  - RDF/Dublin Core metadata in XML box
- Printable Image: PDF
  - Compatible with Acrobat 5.0 (PDF 1.4)
  - XMP/RDF/Dublin Core metadata
- OCR – ALTO
  - Modified ALTO (Analyzed Layout and Text Object) XML schema with word coordinates

Browse, Search, Find: METS Objects

METS (Metadata Encoding & Transmission Standard) – XML Schema-based specification for wrapping varying metadata types, filenames and locations into information packets.

- Title METS Object
  - CONSER records, from OCLC (biblio + holdings)
- Issue METS Object
  - Date, title, LCCN, producer, source, etc.
  - Individual page data folded into Issue METS
  - PREMIS/MIX metadata added at validation
- Reel METS Object
  - Technical film measurements
  - Technical target images

Our Goal: Mitigate Threats

Our goal is to mitigate threats to the content, by creating, using, and enforcing standards from the birth of digital content through the eventual viewing of the content by visitors to the website.

1. Deb covered the ways in which we create and select standards to use.
2. I will talk about technical ways we enable and enforce standards.

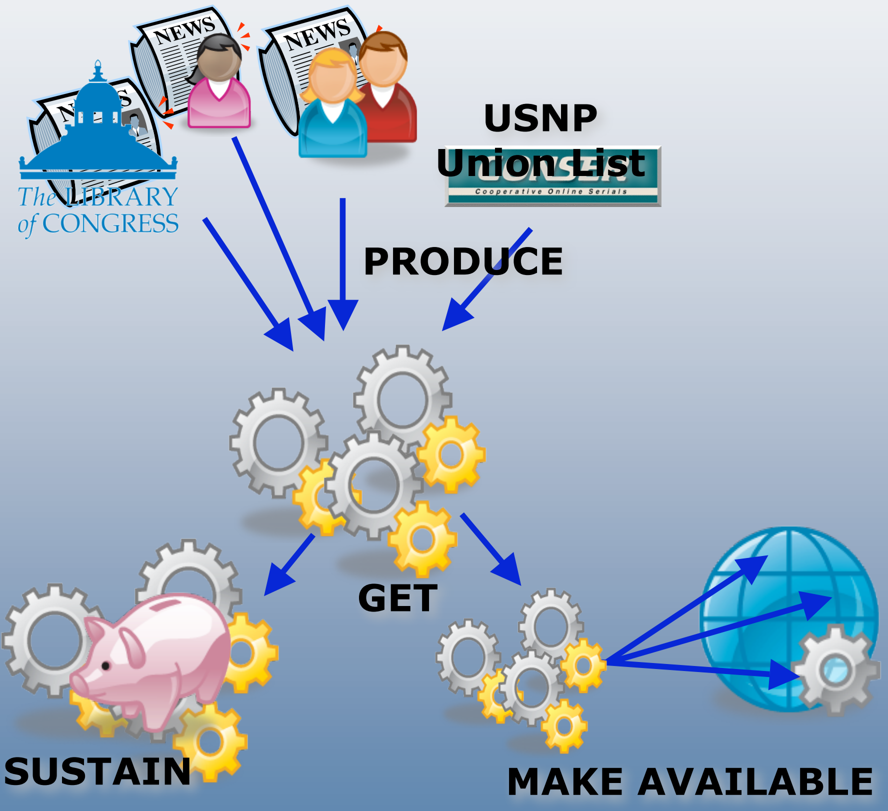
Repository Development Center (RDC)

Our group at the Library of Congress builds processes, environments, and tools to help manage collections:

- Historic Newspapers
- Electronic Journals
- Multilingual Content
- Library Directories
- Inventory Management
- Content Transfer

Newspapers Come to LC in Batches

- Created by publishers
- Digitized by awardees
- Validated in the field
- Transferred to The Library
- Reviewed by digital conversion specialists
- Archived for posterity
- Processed into our access systems
- Made available online



Enforcing Validity Early and Often

- Digital Viewer and Validator
  - The Validator
    - Structure
    - Formats
    - Sanity
    - Fixity
  - The Viewer
    - Content
    - Form
    - Metadata

Tracking Changes

Upon receipt of batches, The Library assumes custodianship, keeping track of all intentional and unintentional changes:

1. Verify fixities
2. Quality review
3. Archive a true copy
4. Transfer to processing
5. Verify fixities
6. Process
7. Create Fixities
8. Transfer to access systems
9. Verify Fixities

Preparing for Data Loss

Data loss will happen.

- Data archived to tape
- Tapes stored off-site
- Batches periodically retrieved
- Fixities checked
- Decay repaired

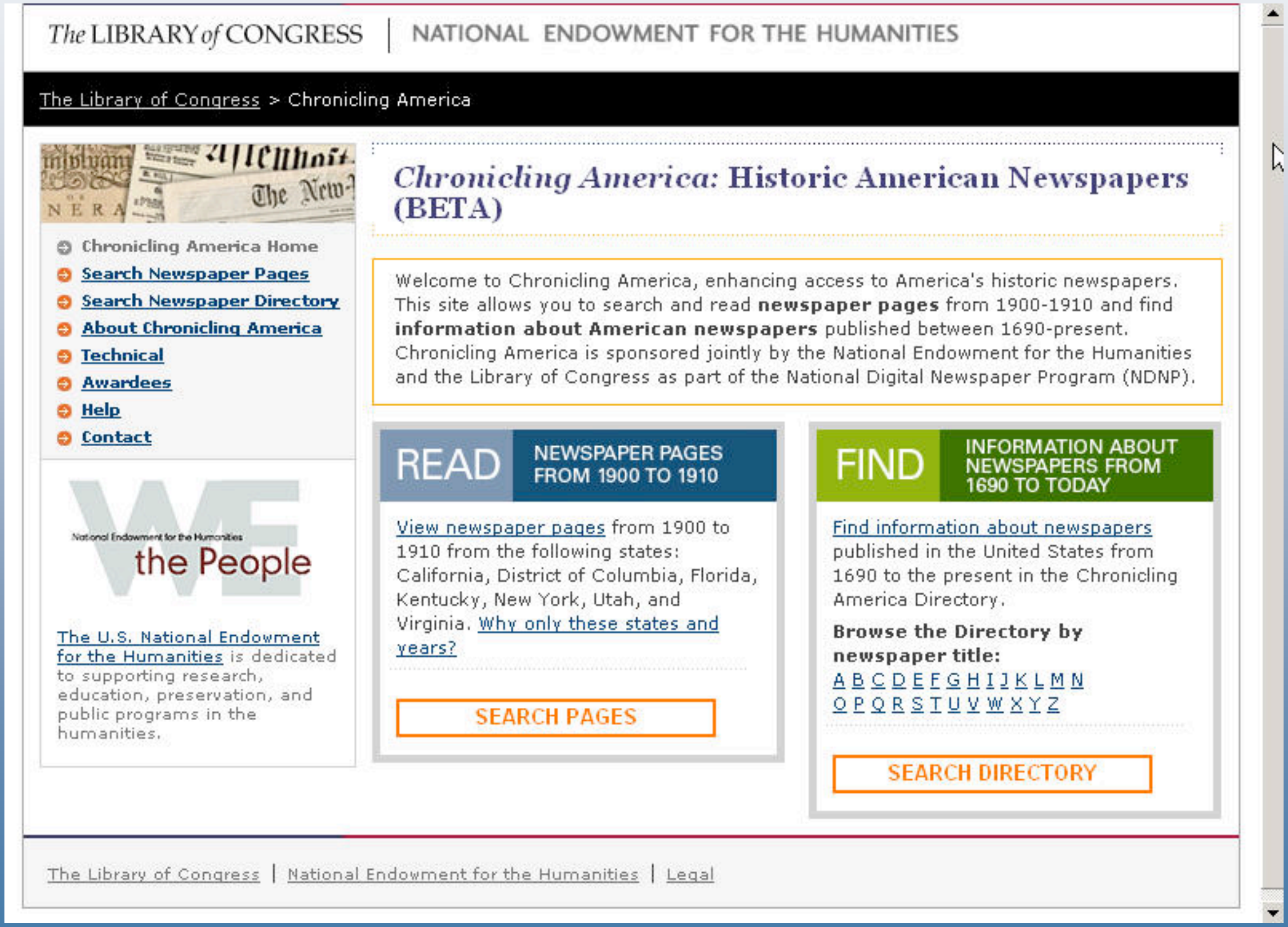
Managing Transfers

As files are moved throughout our system, an appreciable threat is that one of our staff may miss an important step or cause an error, intentionally or unintentionally. Managed transfer reduces this risk:

- Automated receipt
- Automated movement from machine to machine
- Tracking inventory
- Auditing live systems
- Reporting progress through the system

Providing Enduring Access

http://www.loc.gov/chroniclingamerica


The LIBRARY of CONGRESS | NATIONAL ENDOWMENT FOR THE HUMANITIES
The Library of Congress > Chronicling America
Chronicling America Home
Search Newspaper Pages
Search Newspaper Directory
About Chronicling America
Technical
Awardees
Help
Contact
National Endowment for the Humanities
We the People
READ
NEWSPAPER PAGES FROM 1900 TO 1910
SEARCH PAGES
FIND
INFORMATION ABOUT NEWSPAPERS FROM 1690 TO TODAY
SEARCH DIRECTORY
The Library of Congress | National Endowment for the Humanities | Legal

Chronicling America: Historic American Newspapers (BETA)

Welcome to Chronicling America, enhancing access to America's historic newspapers. This site allows you to search and read **newspaper pages** from 1900-1910 and find **information about American newspapers** published between 1690-present. Chronicling America is sponsored jointly by the National Endowment for the Humanities and the Library of Congress as part of the National Digital Newspaper Program (NDNP).

View newspaper pages from 1900 to 1910 from the following states: California, District of Columbia, Florida, Kentucky, New York, Utah, and Virginia. Why only these states and years?

Find information about newspapers published in the United States from 1690 to the present in the Chronicling America Directory.

Browse the Directory by newspaper title:
A B C D E F G H I J K L M N
O P Q R S T U V W X Y Z

The U.S. National Endowment for the Humanities is dedicated to supporting research, education, preservation, and public programs in the humanities.

Access Evolving

- Interface modifications since March 07 launch
  - Flash-based zoom/pan page to AJAX-based viewer
  - Browse Issues (calendar view)
  - Historical context provided for each digitized title
  - Persistent links to return/cite page, record, calendar
  - See All Available Newspapers table and download
- Steady growth of use as content grows
- Over 2.5 million page views so far
- Some specifics:
  - 75 % of users come from other sites – government, libraries, genealogy portals, blogs
  - 20% of users search Newspaper Directory
  - File downloads: JPEG2000 to PDF = 50/50
  - Public uses – genealogy, linguistics, history (blogs, Flickr)

What's Available Now?

- 138,000 title records (updated annually) and associated holdings
- 568,000 pages in 61 titles, 1897-1910 (updated quarterly)
- California, DC, Florida, Kentucky, New York, Utah, Virginia

What’s Next?

COMING SOON – more, more, more

- California, DC, Kentucky, Minnesota, Nebraska, New York, Texas, Utah, Virginia…!!!
- Phased chronological coverage:1880-1910 (2007 awards), 1880-1922 (2008), 1860-1922 (2009), 1835-1922 (2010)…
- NEH announces 2008 awards in June;
- NEH announces 2009 Program Guidelines in August 2008 (1860-1922)

Thank You

- NDNP Technical Information

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

- NDNP Web Portal

Chronicling America: Historic American Newspapers

http://www.loc.gov/chroniclingamerica

Questions?