

Digitization: Let's Turn Our Attention to Newspapers



Leigh Grinstead, Projects Coordinator
Collaborative Digitization Program

Sarah Friedmann, CHNC Project
Technician
Collaborative Digitization Program

December 21, 2005



Collaborative Digitization Program



CDP Partners 1998

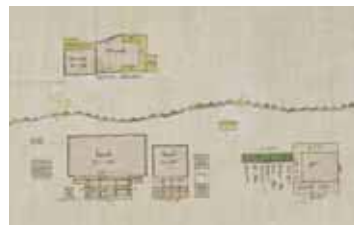


CDP Partners 2004-05

In the beginning...1998



...2002...



...2003...



...2004...



Digital Audio

...2006...

Digitized
motion picture film



COLORADO'S
NEWSPAPER



HISTORIC
COLLECTION

Colorado State Library & Colorado Historical Society

Colorado's Historic Newspaper Collection (CHNC)

A partnership between:

- Colorado Historical Society (CHS)
- Colorado State Library (CSL)
- Collaborative Digitization Program (CDP)



Advisory Committee Libraries/Museums/Historical Societies

- Colorado State Library
- Longmont Public Library
- Southern Peaks Public Library
- Pueblo City-County Libraries
- Jefferson County Libraries
- Telluride Public Library
- City of Littleton Historical Museum
- Colorado State University
- Boulder Public Library Carnegie Branch for Local History
- Denver Public Library
- Colorado Historical Society
- Telluride Public Library
- Collaborative Digitization Program

Program Staff / Tasks

- Program Director
- Project Technician
- Duplication Vendor
- Digitization Vendor
- OCR Vendor
- Server Support
- Educator
- Marketing
- Web Development
- Metadata
- Added Value Content creation
- Grant writer

Funding

- started with LSTA funding: \$120,000
- expanded with IMLS funding: \$250,000
- 2nd LSTA funding: \$47,000
- Future Funding
 - NEH and Library of Congress
 - National Digital Newspaper Program
 - 20 years
 - Create a national digital resource of historically significant newspapers

CHNC Goals

- To offer free, online access to all newspapers published in Colorado up to 1923.
- All titles and issues up to 1923
- 1.6 million pages

So far...on current funds

- 86 titles
 - First paper: Rocky Mountain News Weekly (April 23, 1859)
 - Most recent addition: New Castle News (April 15, 1893 – April 29, 1898)
- 200,000+ pages currently available

It Sells Itself

- Enthusiastically received
- “Life is so much easier now that these newspapers are available online.”

Private funding...

- CHNC has raised over \$200,000

CHNC Usage Stats

2005

- Total Views (all papers): 3,461,235
- Total Articles Viewed (all papers): 3,242,883

2004

- Total Views (all papers): 1,272,144
- Total Articles Viewed (all papers): 1,196,118

November 2005

- Total Views: 683,558
- Total Successful Hits (Entire Site): 880,730
 - Average Visit Length: 23 minutes

What is included in the cost?

- Microfilm Creation
- Microfilm Duplication
- Digitization
- OCR & Article Segmentation (distillation)
- Software License (delivery)
- Hardware Infrastructure
- Storage
- Backup

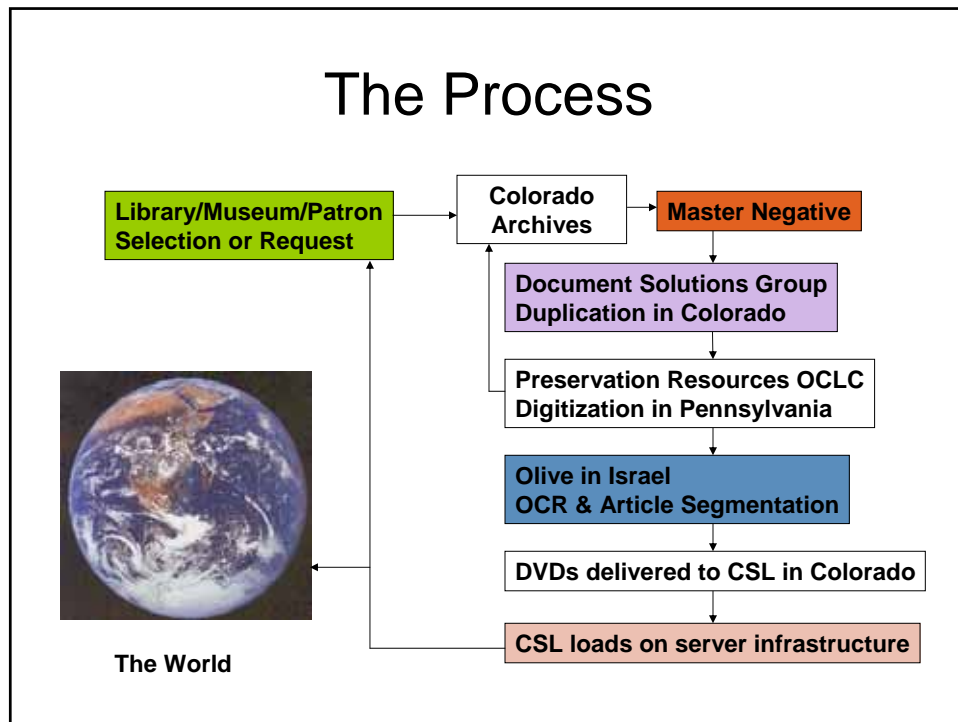
CHNC Costs

- \$400 set up fee per title layout
- \$1.25 page
 - \$.29 page for duplication
 - \$.32 page for digitization
 - \$.14 page for online storage
 - \$.10 page for backup storage
 - \$.40 page overhead (staff, server support, storage reserve funds, etc.)

Private Funding Contributions

- Publisher--\$25,000
- Editor-in-Chief--\$10,000
- News Editor--\$5,000
- Sports Page Editor--\$2,500
- Society Page Editor--\$2,000
- Editorial Writer--\$1,000
- Columnist--\$500
- Circulation Supervisor--\$100

The Process



Digitization

- Preservation Resources Equipment
 - SunRise and NextScan microfilm scanners
 - Zeutschel, PowerPhase, Fujitsu, Microtek, and BetterLite direct scanners for image capture and processing

OCR and Article Segmentation



Microfilm

- Master Negative
 - It can be good!
 - It can be bad!
- Quality of film
- Condition of paper
- Ink Bleed



Original Papers

- Condition of paper
 - Touch it, and it crumbles
- Bound?
- Will you film, too?
- Up to 20% higher OCR accuracy
- Option of RGB digitization (\$\$\$\$\$)

Microfilm vs. Originals

- Missing Issues
- Bi-tonal display or grayscale?
- Master file resolution?
- Costs

CHNC Delivery Infrastructure

- Active Paper from Olive Software
 - Full text searching
 - View entire page or individual articles
 - Browse by title
 - Search by date, keyword, newspaper title
 - Email an article to a friend
 - Print an article or full page
- www.olivesoftware.com



Accuracy of Search

- Accuracy of spelling in historic newspapers

"It's a damn poor mind that can think of only one way to spell a word." ~ Andrew Jackson
- Accuracy of OCR
 - Condition of Master negative
 - Condition of the newspaper when filmed

OCR is never perfect

- If you can't read the film or the paper, the OCR won't be able to either
- If the OCR doesn't work, there's no point
- What accuracy rate is high enough?
- OCR vs. OWR (www.iarchives.com)

Copyright...you can't escape



Keep it simple

- Digitize issues published prior to 1923
- Form a relationship with state press associations
- Form a relationship with local newspaper owners

Dealing with Digital Data

- Master tiff files
- XML files
- Server Infrastructure
- Ongoing need for additional storage
- Adequate backups (what is adequate?)

Sustainable Funding

- On-going digitization costs
- Server upgrades
- Software upgrades
- Software obsolescence
- More and more and more storage
- And more and more and more storage
- Digital Preservation

How to Contact Us

Jill Koelling, Executive Director

jill.koelling@du.edu

Leigh Grinstead, Projects Coordinator

leigh.grinstead@du.edu

Sarah Friedmann, CHNC Project Technician

sarah.friedmann@du.edu

Collaborative Digitization Program

www.cdpheritage.org

