Newspaper Digitization: Issues and Opportunities

Phil Sager
OHS
MISSING YOUTHS FOUND BY HERDER

Boys, Seeking Body Of Lion, Become Lost in Hills.

After straying from a Moonheart picnic at the head of Indian canyon on the Duchesne highway at noon Sunday, two Helper boys, Jack Will Hams, 9, and Thomas Pogano, 5, became lost and were not found until 7:15 that evening. The youths were found by L. Tatton, a sheepherder, in Avintaquin canyon, eight miles from the picnic grounds.

News Advocate (Price, UT)
September 4, 1930
John, I saw your article in the Tribune the other day and put it to the test. This morning I found the article I thought might exist on me and my uncle getting lost near Duchesne 74 years ago. My name then was Thomas Pagano and now is Thomas Billis. My family is ecstatic over the find. My daughter who is a genealogist went absolutely bonkers and said it was actually almost unbelievable to find such information. She said "but Dad you were only 5 years old then and are 78 now". Thank you for sharing the information that you guys exist and how we can use it.

Thomas W. Billis (2/10/04)
Why Digitize Historic Newspapers?

- Fosters interest and pride in a community’s history
- Birth, death, marriage announcements
- New technologies to experience newspapers online
- Wealth of historical information
- 24/7 access, worldwide
Why more attention lately?

- National Digital Newspaper Project (NDNP)
  - “Ultimately, over a period of approximately 20 years, NDNP will create a national, digital resource of historically significant newspapers from all the states and U.S. territories published between 1836 and 1922”
NDNP

- 2005 awardees:
  - University of California, Riverside
  - University of Florida Libraries, Gainesville
  - University of Kentucky Libraries, Lexington
  - New York Public Library, New York City
  - University of Utah, Salt Lake City
  - Library of Virginia, Richmond
Why more attention lately?

- The technology has improved
- Though still relatively few, projects digitizing historic newspapers are increasing in number
- Public is finding out through news, announcements, google, etc.
So What’s So Hard About Putting Historic Newspapers Online?

- Readability issues (human and OCR)
- Conservation issues
- Digitization / image processing
- OCR processing
- Recording metadata
  - structural, descriptive
- Storage
Examples

- Utah Digital Newspapers
  - http://www.lib.utah.edu/digital/unews/
- Brooklyn Daily Eagle
  - http://www.brooklynpubliclibrary.org/eagle/
High-Level Process

- Select titles UofU w/ Adv Bd
- Obtain source materials UofU w/
  - Originals owner
  - Film BYU
- Repair originals (if needed) UofU
- Scan eprep
  - Originals iArchives
  - Film iArchives
- Zone / metadata / OCR iArchives
- Database index DiMeMa
- Database load UofU
- Web development
Flow of Original Papers

1. Eprep sends hard drive of raw scans
2. iArchives sends dvd’s of completed digital files
3. iArchives sends hard drive of prep’ed, high res archival images
4. DiMeMa sends dvd’s of completed files ready for database loading
5. UofU loads into CDM
Image processing workflow

- Text Images
  - Image Acquisition
  - Image Processing
  - Page Metadata
  - Article Metadata
  - OWR Framework
  - Post Process
  - Customer Data

Optional Metadata

- Shared Storage (NAS)
- Workflow Manager
- DB

- Data
- Automatic process [image processing, OCR, …]
- Manual process [image + article metadata]
- Quality Control
- Metadata entry Delhi and Coimbatore, India
Image acquisition

The quality of the original image has direct effect on every other aspect of the newspaper digitization process.

- Photo realism
- Text readability
- Metadata collection
- Searchable word accuracy
Image Acquisition

Selecting the **image type** is the first decision to be made when considering the viewing experience.

- Color, grey, or black-and-white (Bit-depth)
- **Image resolution** (ppi)
- Format (TIF, JPG, JP2, GIF, PDF)
- Compression type and level
Image Acquisition

Several factors come into play when selecting the image type to be used for an online newspaper.

- End-user bandwidth
- Original media type (microfilm, microfiche, paper)
- Original material format (page size, font size, photos, etc)
Image Processing

- Cropping
- Deskewing
- Despeckling
  - removal of “salt and pepper” noise that prevents good OCR
- Binarization (for OCR)
  - converts greyscale and color images into black and white images
There are Two levels of metadata in newspapers

- **Page-level**
  - NDNP at present
  - Structural, Text Layout

- **Article-level**
  - Structural, Text Layout
  - Title, subject, etc...
  - Others including UofU
Metadata

- Page-level metadata
  - Inexpensive to capture
  - OK for browsing
  - OK for searching
  - Relatively small amount of data
    - Article metadata and zoning not present
Metadata

- Article-level metadata
  - More expensive to capture
  - Good for browsing
  - Good for searching
  - Relatively large amount of data
    - Article metadata and zoning present
Metadata

- Page level view
  - For example, NDNP
  - http://www.loc.gov/ndnp/xml/issueTemplate.xml
Article level view

- For example, University of Utah
- XML includes all page-level data, plus:
  - Article headline
  - Article category
    - birth announcement name
    - death announcement name
    - marriage announcement name
  - County
Tool Set - HeaderMan

HeaderMan is a very efficient means of entering metadata to the page level. It requires 2 entries and reconcile to maximize accuracy.
Tool Set - ZoneMan

ZoneMan is a flexible, efficient means to collect customized article-level meta data.
OCR and OWR

- OWR = Optical Word Recognition
- First, it’s important to realize that OCR/OWR does not yield article “transcriptions”
- The text OCR’d from the images of historic newspapers is used for searching purposes
OCR and OWR

Text Images → Image Acquisition → Image Processing → Page Metadata → Article Metadata → OWR Framework

Text image

OCR Engine 1
- boat (73%)

OCR Engine 2
- beat (90%)

OCR Engine 3
- beet (86%)

Index
- apple
- beat
- cat
- dog
- east
- frog
- grape
- [etc…]
OCR and OWR

What if there are OWR ties?
OCR and OWR

Can put two (or more) words in the searchable index

Index
apple
beat
boat
cat
dog
east
frog
grape
[etc…]
OCR and OWR

Text Images → Image Acquisition → Image Processing → Page Metadata → Article Metadata → OWR Framework

Text image

Dictionary 1

Dictionary 2

Place name dictionary

Personal name dictionary

Searchable Index

apple
beat
cat
dog
east
frog
grape
hat
[etc…]
Digital Library (NDNP)
Digital Library (NDNP)
Digital Libraries

- ContentDM
- Olive (ActivePaper platform)
- Greenstone
- Fedora
- DSpace
- “Homegrown”
NDNP Digital Library (LoC)

- Gentoo Linux (operating system)
- Fedora (digital repository)
- Tomcat (servlet container)
- Cocoon (web application framework)
- Lucene (index/search engine)
- MySQL (database)
- Apache (web server)
- JDK 1.5 (java environment)
NDNP Digital Library
NDNP Digital Library
The End

Phil Sager
OHS