Getting Started with Scanning Newspapers

Upper Midwest CONTENTdm Users Conference 11-15-2011

Presenters:
Bill Wittenbreer, Augsburg College
Jill Strass, St. Olaf College

Photo by Ben Shahn for Farm Security Administration, New Orleans, 1935
http://commons.wikimedia.org/wiki/File:JacksonSquareNewspaper1935.jpg
St. Olaf’s Story: How we got our scanner:
Before we applied for a local grant

- Before we applied, we had a record of successful digital projects.
- We had buy-in from the Norwegian American Historical Association, as this was a long-standing project with a recognized institutional need.
- We had support from our Library Director.
- At the time of the application, awareness of digital projects was rising within the upper Administration.
St. Olaf’s Story: How we got our scanner: Our proposed digital content

• We had content that matched the intent of the grant perfectly.
• Odd Lovoll, professor Emeritus published a book called Norwegian Newspapers in America (2010).
• The Rolvaag Library is a repository for a number of Norwegian American Newspaper titles described in the book.
• A bit about Norwegian American Newspapers…. 
St. Olaf’s Story: How we got our scanner: Breaking down the project into Two steps

- We won a Nygaard grant, which was offered by St. Olaf College for projects that foster an appreciation for Norwegian Heritage.
- As part of the grant, we wrote the project into two segments:
  - Digitization (image capture)
  - Access
- We got funding for the Digitization part of the project.
- Funding covered the scanner, student hours and image storage for 25,000 pages.
Augsburg’s Story: How we got our scanner

- A coworker, Mike Bloomberg discovered the ATIZ scanners.
- The ATIZ scanners had the functionality we wanted, but at a fraction of the price of a traditional overhead scanner.
- Though more affordable than an overhead scanner, the price was still an obstacle.
- Mike and I decided to apply for legacy funding from the Minnesota Arts and Cultural Heritage Fund.
- The application process was relatively easy.
- We were turned down for funding.
- The Miracle.
Everything you want to know about ATIZ
Atiz SCANNERS: Differences

• Different lighting systems
  o BookDrive Pro has a tent over the cameras, BookDrive DYI is open to atmospheric light
  o BookDrive Pro has LED lights that give a constant level and intensity of light

• Different size scanning cradles
  o BookDrive Pro has capacity for larger documents

• Auto Capture Switch
  o BookDrive Pro has the Auto Capture Switch, which allows you to take pictures automatically when the platen is lowered

• Cost differences
  o BookDrive Pro was 24K
  o BookDrive DYI was 15K
Augsburg’s BookDrive DIY scanner plus Hannah
Atiz SCANNERS: Glare

• BookDrivePro has a cute tent
  o Even with the tent, we needed to make a few modest changes (cardboard taped as a flap behind the camera) as the lighting in the room is florescent and we had some glare.
  o Be aware of the room you will use to scan items, as atmospheric light will make scans appear inconsistent. The tent and LED lights make controlling lighting conditions a lot easier.

• Weird light blobs
  o We've had unexplained light blobs appear on scans which we cannot get to repeat.
  o Flash-related light blobs.
ATIZ Software: Capture

• Captures images from the cameras to the hard drive
• You have options for file formats for capture
  • We use RAW as our archival file and then save to TIFF, as our goal is to use OCR at some point.
• Option of using 1 computer per camera, which could increase production output
• We use one computer for 2 cameras
• Image captures are “pages” in a book, folders are digital representations of a book
• Unstructured documents
ATIZ Software: Editor Pro

- Similar functionality to other editing software, it rotates, deskews, crops, adjusts lighting and contrast.
- The more sophisticated features include despeckle, border and background removal, and sharpening.
- Deleting scans works much better in the editing mode.
- The system is set up so that the RAW file is the archived copy. This means that edits can be totally redone in the future if needed.
ATIZ Software: Editor Pro

• You can set up group of settings called profiles for different projects so you don’t have to redo your settings.
  – Profiles allow multiple users and logins to use preselected settings so that all the files have the same look and feel.
• It is possible to batch edit or edit individual scans.
• Image/file type is determined in the editing mode along with quality, resolution and color depth.
Workflow issues: Things to Consider Before you Start Scanning

• OCR (Optical Character Recognition) is a computer program that extracts characters of text from an image.
• Educate your Administrators that OCR is a separate process from scanning.
• We haven’t found software yet that can handle OCR for newspapers automatically.
• We have found that if we hand-select text column by column, some software works, but this process is purely manual and time-consuming.
Workflow issues: Things to Consider Before you Start Scanning

- Decide what the “unit” is for a project
  - an issue of a newspaper is a “unit” vs a volume of several issues of newspapers
- Consider the condition of what you are scanning:
  - Huge number of images possible in short time if clean, even pages.
  - Relatively slow for paper in poor condition
- When to scan in color, when in grayscale
  - This decision will impact how much storage space your project will need
Workflow issues: Things to Consider Before you Start Scanning

• Quality Assurance for images
  • Folder and File name proofing - we use a Python script
  • Image standards for text extracting
  • Image standards for aesthetics
• Version control
  • Initial RAW scans
  • Edited scans
• File Format choices
  • JPG, JP2, TIF, RAW, others?
Workflow issues: Things to Consider Before You Start Scanning

• File Management
• Making backup copies
• Storage issues for image files
• Cloud vs internal server vs external drives
• Guesstimating future disk space
Workflow issues: software and file formats.…

• Camera file formats
  • RAW
• Training scanner operators
  • Student scanners sat in on vendor training
  • Students train each other.
Workflow issues: Tinkerer’s dream

• Bound vs Unbound – shimming and propping up books and pages to get them flat under the glass
• Backer paper for newspaper fragments – plain brown paper behind a newspaper fragment
• Piece of cardboard taped to block florescent light
• What does the F-stop and shutter speed really mean? What happens with no film speed?
• Troubleshooting camera settings to get the consistent look and feel
Work Flow at Augsburg

• We had a student worker attend the vendor training
• This student trained his peer.
• Both Inter-library students scan, in the down times.
• We keep a running log of the scanning for the newspaper.
• Log includes documentation of scan settings, file names and other details.
• Raw and edited scans are stored in network folders.
Work Flow at St. Olaf

• We had a student worker attend the vendor training.
• This student trained two peers.
• We are using library student work hours and grant funding pay for scanning.
• We keep a running log of the scanning for the newspaper.
• I act as chief editor for the project, so I set standards for images, determine settings, offer guidance and troubleshooting for students.
• Scans and Scan edits are stored on a hard drive and external drives.
In Summary

• The more upfront planning, the easier it will go
  • Don’t let the perfect be the enemy of the good
• Have clear goals for the project
• Let administrators know this isn’t a fast process
• Be ready for compromises
  • Leave perfection behind; figure out what is reasonable to do
• Be flexible and creative; cardboard can solve a lot of problems!
• Face this like MacGyver
Face this like Macgyver....