Crossing State Lines for Collaborative Newspaper Digitization

The Gateway to Oklahoma History

Jennifer Day
Mallory Newell
Chad Williams
Oklahoma Historical Society

Sarah Lynn Fisher
University of North Texas Libraries

Society of Southwest Archivists Annual Meeting - May 24, 2012

Overview of the Project

Funding

- The project is funded in part by the Excellence and Ethics in Journalism Grant

Pre-1923

- The goal of the project is to digitize all of the pre-1923 newspapers in the Historical Society's collection
- The newspapers range from 1844 all the way through the end of 1922.
- Once completed the project will have digitized approximately 5,000,000 pages of newspapers
- The Gateway to Oklahoma History will allow easy access to newspapers for students, researchers and journalists.

Project participants

Staff and Volunteers

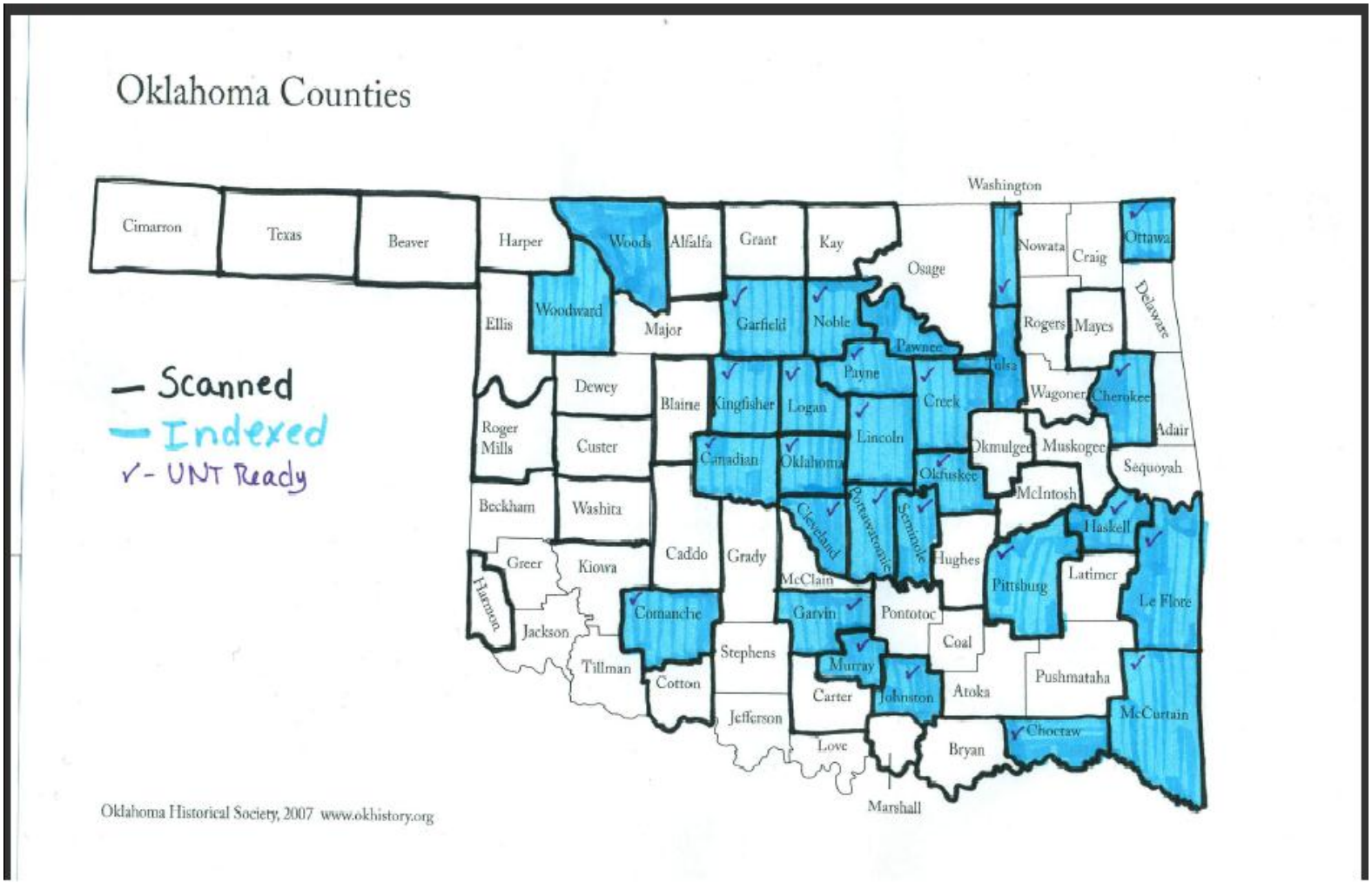
- Currently we have 2 full time staff and seven volunteers working on the project
- Four volunteers index, one scans, and two write essays

Workflow

Overview

- There are seven steps involved in processing a reel of microfilm
  - Scanning
  - Auditing
  - Indexing
  - Sort 1
  - Quality Control 1
  - Quality Control 2
  - Sort 2
- Each step of the processing workflow has its own designated folder with the last being University of North Texas Ready
- To keep track of all the reels in different stages of processing we have a master list in excel which is color coded.

Map of Progress



Scanning

- NextStar scanners are used to scan each reel of microfilm
- We have recruited one volunteer to help with this process
- We hope to have all the scanning done by the end of next year

Scanning Volunteer

Auditing

- Using the NextStar auditing software we can look at the images after they have been saved.
  - Images are checked for readability
    - Too dark or light
    - Focus
  - Make sure the images are actually there
  - Check the reel number against paper title
- This is where we split the images into individual pages

Indexing

- Each reel of microfilm is indexed according to six elements in an excel spreadsheet
  - Date
  - Filename
  - Edition
  - Volume
  - Issue
  - Note
- During indexing, we collect a lot of the metadata used later on
- Images are viewed with ACDSee Pro.

B228

	A	B	C	D	E	
1	date	filename	edition	volume	issue	note
2	1910/05/06	1	1	2	8	
3	1910/05/20	6	1	2	6	
4	1910/06/10	11	1	2	9	
5	1910/06/17	15	1	2	10	
6	1910/06/24	19	1	2	11	
7	1910/07/01	23	1	2	12	
8	1910/07/08	27	1	2	13	
9	1910/07/15	31	1	2	14	
10	1910/07/22	35	1	2	15	
11	1910/07/29	29	1	2	16	
12	1910/08/05	43	1	2	17	
13	1910/08/12	47	1	2	18	
14	1910/08/19	51	1	2	19	
15	1910/08/26	55	1	2	20	
16	1910/09/02	59	1	2	21	
17	1910/09/09	63	1	2	22	
18	1910/09/16	67	1	2	23	
19	1910/09/23	71	1	2	24	
20	1910/09/30	75	1	2	25	
21	1910/10/07	79	1	2	26	
22	1910/10/14	83	1	2	27	
23	1910/10/21	87	1	2	28	
24	1910/10/28	91	1	2	29	
25	1910/11/04	95	1	2	30	
26	1910/11/11	99	1	2	31	
27	1910/11/18	103	1	2	32	

Note Field Example

Technical Notes

F20

	A	B	C	D	E	F
109	1912/08/02	585	1	4	17	
110	1912/08/09	591	1	4	17	Reprint of 589
111	1912/08/16	595	1	4	18	
112	1912/08/23	599	1	4	19	
113	1912/08/30	603	1	4	20	
114	1912/09/06	607	1	4	21	Front page is split down the middle
115	1912/09/13	612	1	4	22	
116	1912/09/20	616	1	4	23	
117	1912/09/27	620	1	4	24	
118	1912/10/04	624	1	4	25	
119	1912/10/11	628	1	4	25	
120	1912/10/18	632	1	4	26	
121	1912/10/25	636	1	4	27	
122	1912/11/01	640	1	4	28	
123	1912/11/08	644	1	4	29	
124	1912/11/15	648	1	4	30	
125	1912/11/22	652	1	4	31	
126	1912/12/06	657	1	4	33	
127	1913/01/03	662	1	4	37	
128	1913/01/10	666	1	4	38	
129	1913/01/21	671	1	4	40	
130	1913/01/31	675	1	4	41	
131	1913/02/07	679	1	4	41	Last page is split down the middle
132	1913/02/21	686	1	4	43	First page is split down the middle

Important Headlines

F42

	A	B	C	D	E	F	G	H	I	J	K	L
43	1914/10/22	283	1	24	41							
44	1914/10/29	289	1	24	42							
45	1914/11/05	295	1	24	43							
46	1914/11/12	303	1	24	44							
47	1914/11/19	309	1	24	45							
48	1914/11/26	315	1	24	46							
49	1914/12/03	321	1	24	47							
50	1914/12/10	327	1	24	48							
51	1914/12/17	333	1	24	48	ENGLISH CITIES BOMBARDED BY GERMANS						
52	1914/12/24	343	1	24	49							
53	1915/01/07	351	1	24	52							
54	1915/01/14	359	1	24	54							
55	1915/01/21	365	1	25	3							
56	1915/01/28	371	1	25	4							
57	1915/02/04	377	1	25	5							
58	1915/02/11	383	1	25	6							
59	1915/02/18	389	1	25	7							
60	1915/02/25	395	1	25	8							
61	1915/03/04	401	1	25	9							
62	1915/03/11	407	1	25	10							
63	1915/03/18	413	1	25	12							
64	1915/03/25	419	1	25	13							
65	1915/04/01	425	1	25	14							
66	1915/04/08	431	1	25	15							
67	1915/04/15	439	1	25	16							
68	1915/04/22	445	1	25	17							
69	1915/04/29	451	1	25	18							
70	1915/05/06	457	1	25	19							
71	1915/05/13	463	1	25	20	Lusitania torpedoed by German Sub						

Sort # 1- A Two Step Process

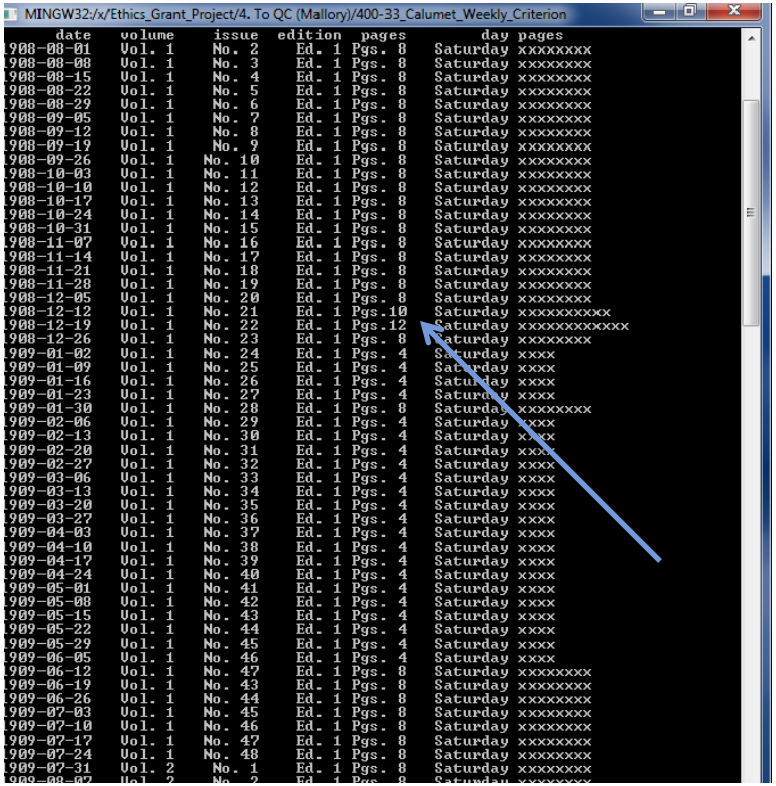
- Step one of the Sort is creating folders
  - A folder is made for each day
  - The folders are created using a python based script
- Step two is the actual sorting
  - Each excel is saved as a CSV file
  - The images are sorted into their proper folders using a python based script
- Before moving the folders to the next step, each reel is checked for accuracy

Quality Control 1&2

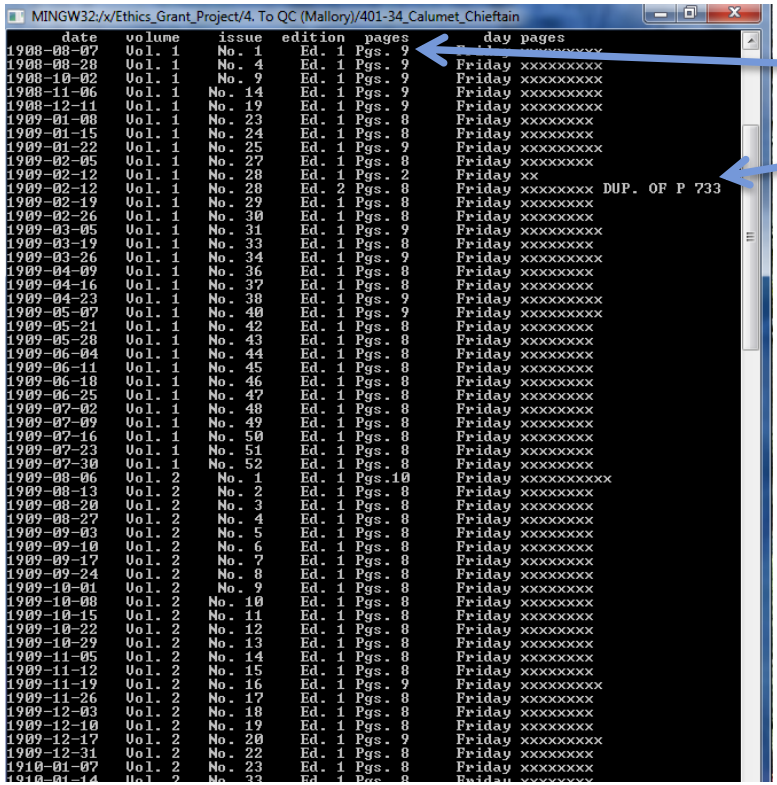
- We have two people complete this step using a python based script to show all the important elements
- The quality control step ensures that each issue has the correct amount of pages, the dates, volume numbers, edition, and issue numbers are correct
- In this step extraneous pages are also deleted ex. duplicates
- During this process Photoshop is used to recombine and split pages as needed

Quality Control

Extra Pages



Notes and Possible Missing issue images



Sort #2

- At this point the newspapers are reorganized by title and year
- Folders are created based on the title of the newspaper and the Library of Congress Control Number
- Reel number does not matter anymore at this point all issues with the same title are put in one folder based on year
- This is the last step before they are sent to the University of North Texas

Additional Material

Newspaper History Essays

- Brief historical sketches will be included with each title
- The essays include important information about the papers
  - City of publication
  - Start Date and End Date
  - Editors, Publishers, Managers, Owners
    - Any relevant information about these people- Where they worked before, where they went after (if available)
  - Paper size
    - Number of Columns
    - Paper measurements
  - Subscription Fees
  - Political Affiliations
  - Any other interesting info about the paper

Example Essay

Norman Transcript [LCCN: sn86064114]

The *Norman Transcript* was first published in July, 1889. Editor, publisher, and owner Ed P. Ingle claimed a business lot on present day West Main and Santa Fe. Ingle had originally come from Purcell, Oklahoma where he established the *Purcell Register*. The paper began as a weekly newspaper published on Saturdays, but moved to Thursdays because of its Republican affiliations. In his salutary editorial in the first issue, Ingle explained the newspapers mission as being dedicated to the progression of Norman as well as the prosperity of the residents.

The first issue appeared with four pages and seven columns. By the second issue, the paper had expanded into eight columns and used a larger type. By 1900, the paper consisted of eight pages and measured 15x22. From 1905 to1906 the papers' circulation expanded from 1,000 to 1,240. In 1912, J.J. Burke replaced Ingle as the editor of the paper. He had previously worked for the *Oklahoma Times-Journal* and the *Daily Oklahoman*. During Burke's tenure, he moved the operation to a new building on East Main Street and merged with the *Cleveland County Enterprise* in 1917. As a result of the merger the *Transcript* was converted into a tri-weekly paper until 1920, when the *Enterprise* was discontinued.

The paper absorbed several other publications including the *Cleveland County Democrat News,* the *Cleveland County Times*, and the *Cleveland County Record*. The *Norman Transcript* started publishing under the name the *Norman Daily Transcript* in 1920, originally issued three times a week. It changed to a daily paper in 1922. The *Norman Daily Transcript* is still in publication today.

End Product

Gateway to Oklahoma History

UNT Libraries
THE POWER OF IDEAS STARTS HERE

University of North Texas Libraries

Development of The Gateway to Oklahoma History

History of the Partnership

- OHS contributes items to The Portal to Texas History - texashistory.unt.edu

- 2009: OHS receives Chronicling America grant from Library of Congress and NEH to digitize 100,000 newspaper pages in two years (renewed in 2011)
  - UNT serves as technical coordinator

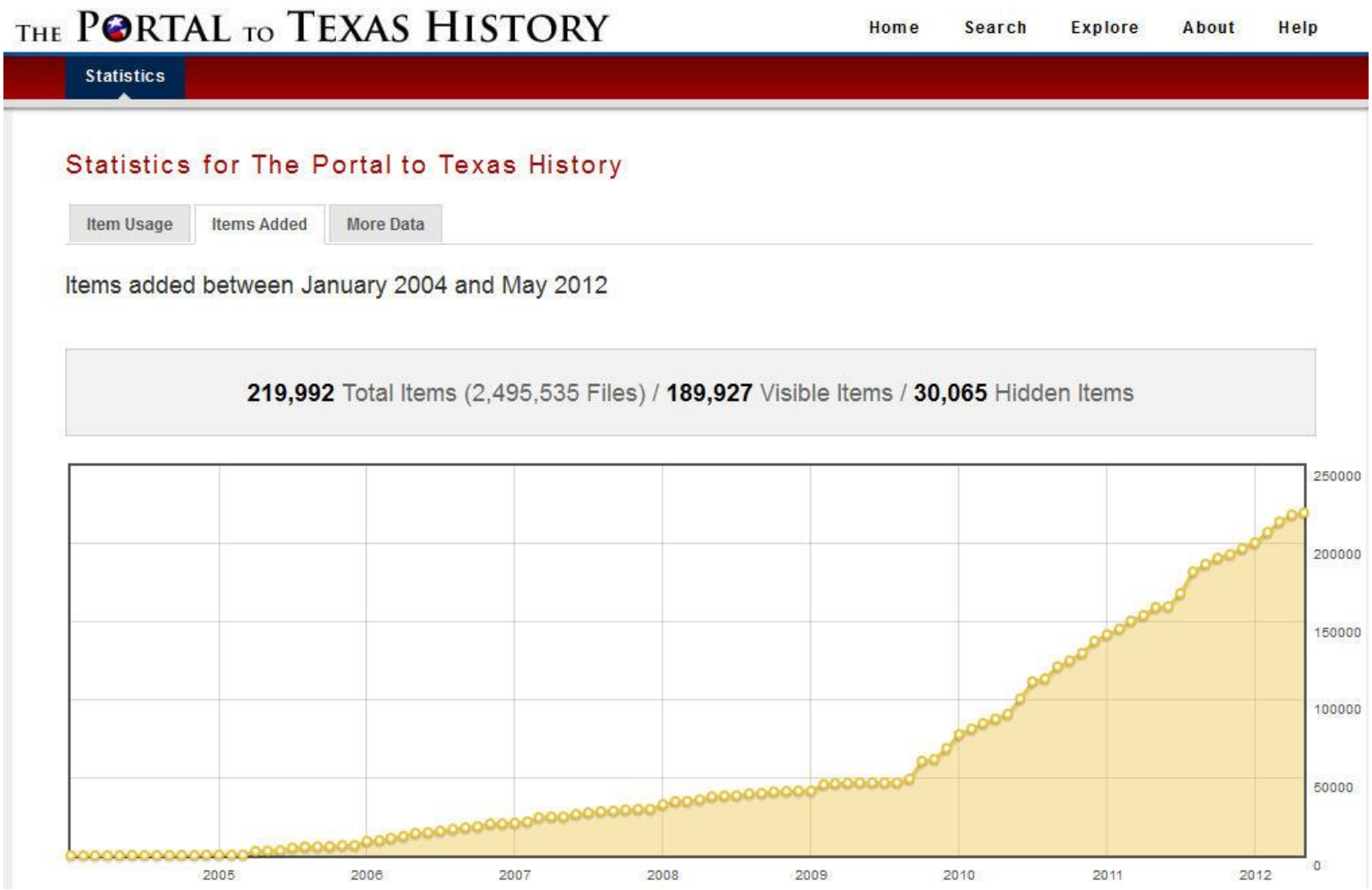
- Ethics and Excellence in Journalism Grant: 5 million pages online in 3 years

UNT Had in Place…

- Established workflow for digitizing newspapers to preservation standards and hosting them online in an open access format
- Digital curation utilities
  - Linux applications to manage data over time
- Content delivery system that can quickly scale

Advantages for UNT

- Expands our newspaper digitization operation
- Benefits our students and scholars
- Improves access to the collection



http://texashistory.unt.edu/stats/. Accessed 5/23/2012

Developing the Gateway

- Aubrey
  - UNT's public-facing content delivery and metadata management system
  - Funded by a grant from IMLS
  - Rapid development framework for interface development in digital libraries
  - One large collection of content from the same servers
  - Gateway is a new interface within Aubrey

Developing the Gateway, contd.

- User Interfaces Unit at UNT Libraries worked with OHS to design the Gateway
- Similar look to the Portal to Texas History and UNT Digital Library with OHS branding; matches their newly-redesigned website
- 2 month process from design to launch
  - Kickoff meeting at UNT in December 2011
  - UI created design mockups, OHS created icons and content descriptions
  - Communication with Basecamp (project management software)

Adding Content

- 14,000 issues of Oklahoma NDNP newspapers were added to the beta site
- By the fall, addition of 12 TB of data, estimated 1 million pages
- Full-text searching
- Collections: titles and subjects

Workflow

- OHS creates content using the naming conventions, scripts, and metadata developed for the Portal
- JPEG 2000 masters and metadata (text files) delivered to UNT via hard drive
- At UNT:
  - ABBY OCR software: running cluster of 30 cores (~ 7 computers)
  - Created a supplemental dictionary of Oklahoma place names and geography
  - Generate: Raw XML, Text file, PDF file for each newspaper page

Workflow, Contd.

- Derivatives are delivered to Oklahoma on drives for long-term storage and also moved to the network at UNT
- Additional metadata added at UNT
- Data is ingested and uploaded
- Collection descriptions and icons created by OHS and UNT

"Super-Metadata" at UNT

- Allows for standardization with variation in high-level records
- Quickly create records for individual newspaper issues
- Some fields are automated: creation date, volume and issue numbers (from metadata text file)
- Others are customized for groups of issues within a date range:
  - Title
  - Creator
  - Description
  - Subjects
  - Location
  - Language
  - Collection
  - Identifier (LCCN)

"Super-Metadata," Contd.

- New Record Creator: reduces errors, access to UNT subject dictionary (UNTL-BS)
- http://edit.texashistory.unt.edu/nrc/
- Create records from templates; import and edit XML records within the system
- Metadata can be accessed for each newspaper issue in the Gateway under "Full Record" tab

The Tulsa Star (Tulsa, Okla.), Vol. 1, No. 24, Ed. 1, Friday, April 11, 1913

Brief Record | Full Record | Statistics

Creator (Editor):	Smitherman, Andrew Jackson A. J. Smitherman, Editor & General Mgr.
Publisher Info:	Publisher Name: The Tulsa Star Printing and Publishing Co. and Publisher Info: 501 North Greenwood Street. Place of Publication: Tulsa, Oklahoma
Publisher Info:	Publisher Name: Andrew Jackson Smitherman Publisher Info: A. J. Smitherman Place of Publication: Tulsa, Oklahoma
Original Creation Date:	April 11, 1913
Coverage:	Place United States - Oklahoma - Tulsa County - Tulsa Era New South, Populism, Progressivism, and the Great Depression, 1877-1939
Description:	Weekly newspaper from Tulsa, Oklahoma. Coverage includes local, state, and national news, along with advertising.

Other Resources

The Oklahoma Historical Society has served as the centralized repository for records dealing with the history of Oklahoma since before statehood.

For more information about these rich collections learn more here.