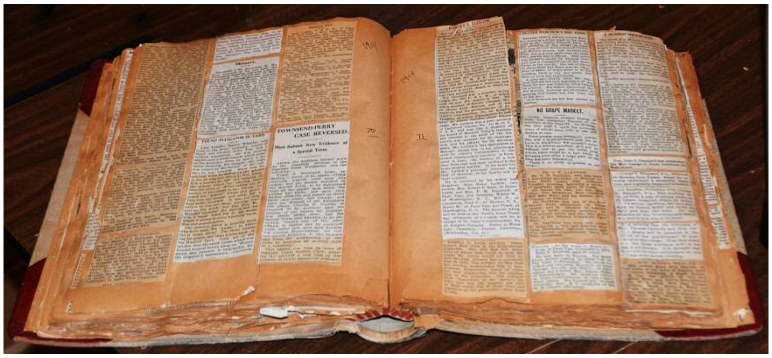
Old Newspaper Scrapbooks Made Searchable with Imaging and OCR
The Sheppard Scrapbooks

The Sheppard scrapbooks plus index are available for research at the Yates County Genealogical and Historical Society (History Center). The 15 books labelled A through M, “Scraps” and “No.48” are filled with mostly newspaper clippings from the very early 1900s through the 1940s. The books were created contemporaneously by George S. Sheppard; his son, Oliver, did work on two of the later volumes after George’s death. A handwritten index was prepared by the Sheppards and later photocopied and bound. Additional information about each of the scrapbooks is in Appendix A.

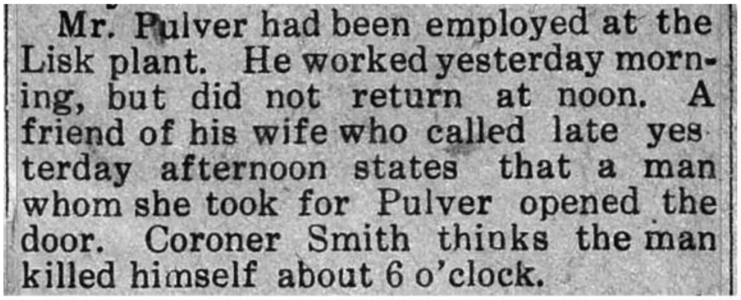
George S. Sheppard was prominent attorney in Penn Yan for more than 60 years. His obituary in Appendix B provides additional information about his life and service to the Yates County community.



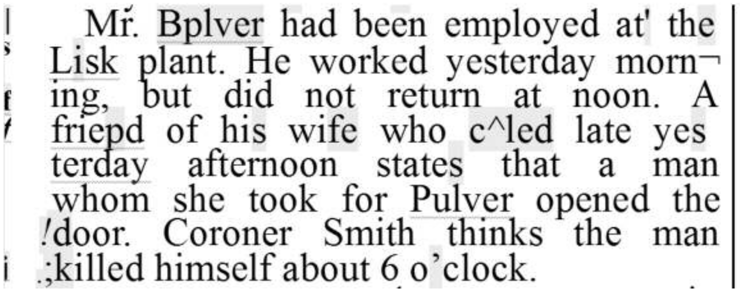
Time and handling have taken a toll on the scrapbooks. The newspaper clippings and the scrapbook pages themselves are in such condition that merely opening a volume and turning a page causes additional deterioration. A project was initiated to make a digital copy of each page in all 15 of the scrapbooks. It was decided to copy the pages using a digital camera instead of a scanner since that process would entail less handling of the delicate pages. Once the pages were copied they could then be viewed on a computer screen without further damaging the delicate old scrapbooks. The page could then be printed if desired. Further information on how the pages were copied using a camera is presented in Appendix C.

Once the more than 2200 digital images (approximately 150 pages per book in 15 books) became available the obvious next step was to make the pages searchable. If, for example, you wanted to know more about George S. Sheppard one could use the existing index and then leaf through the pages of the scrapbooks and find articles that contain his name. If the pages were made searchable on the computer, using Adobe Reader, you would enter the name “sheppard” and find all the occurrences in any and all 150 pages in a scrapbook in less time that it would take you to search for the name on just one page. The existing index is quite limited in that only the name or names of the principal people in an article are indexed. For example, for the article about the wedding of my uncle Joseph Bullock and Madeline Carlson in 1937 the principals are listed in the index. However, none of the other four names in the article appear in the index.

Making the pages searchable was accomplished by using available software to do OCR (optical character recognition) and convert the image files into searchable PDF (portable document format) files. Some specific details of this process are contained in Appendix D.



Mr. Pulver had been employed at the
Lisk plant. He worked yesterday morn-
ing, but did not return at noon. A
friend of his wife who called late yes
terday afternoon states that a man
whom she took for Pulver opened the
door. Coroner Smith thinks the man
killed himself about 6 o'clock.



Mr. Bplver had been employed at' the
Lisk plant. He worked yesterday morn¬
ing, but did not return at noon. A
friepd of his wife who c^led late yes
terday afternoon states that a man
whom she took for Pulver opened the
!door. Coroner Smith thinks the man
.;killed himself about 6 o'clock.

It's important to discuss here the limitations of OCR and PDF files when dealing with old newsprint. For purposes of this particular application, think of a PDF file as having two layers. The top layer is a view of the document as it appears. The second layer, which under normal circumstances can't be seen by the user, is the text of the document that results from the OCR process. The example to the left, from page 29 of scrapbook F, shows the top layer on top and the second layer on the bottom. The OCR process incorrectly determined that the "Pu" in "Pulver" on the first line was a "Bp." This was undoubtedly caused by the smudge between the "P" and the "u". If one had been searching for the family name "Pulver" that particular occurrence of "Pulver" would not be found. The one on the sixth line, however, would be found. Another smudge led to the "n" in "friend" on the fourth line being OCRed as a "p". Since smudges and other anomalies abound on old newspaper pages, these types of errors will occur when dealing with old newspapers.

Studies of OCRing old newspapers in Australia and Utah have resulted in a word "accuracy" of about 60%. That is, in searching a PDF file of an old newspaper the chances are only 60% of finding a word that is actually there. If, in our project of making these scrapbooks searchable, our hit rate is only 60% it may not be worth the effort to digitize such historical holdings.

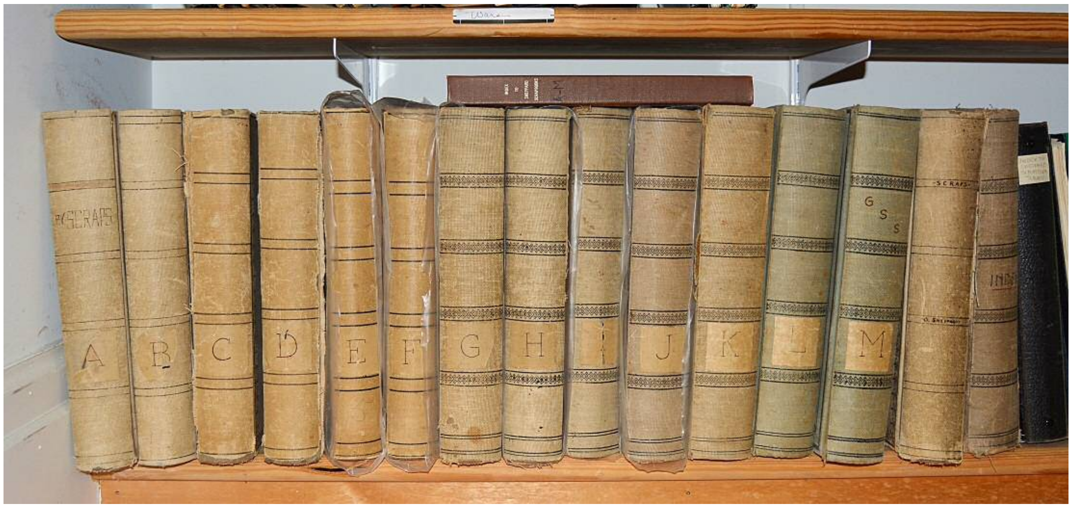
To address this question relative to this particular project, a random sample of words, mostly proper nouns, was selected from the scrapbooks themselves. 50 pages were randomly selected from each scrapbook and a word on that page was selected at random. The sample size was 750 (15 x 50). A search using Adobe Reader was made for each of the preselected 750 words and recorded as a hit or a miss. A hit was recorded only when the search zeroed in on the selected word in the selected place on the selected page. The results were 704 hits and 46 misses out of the 750 tries. There were 6.1% misses with a sampling error of + or - 2%. If the true miss rate is as high as 8.1%, would it be worth one's while to search for a name knowing that when the name appears in the scrapbook the search will find it only 92 times out of 100? More information about the sampling study is in Appendix E.

A hit rate of 9 out of 10 is remarkably good considering the source material. However, it does remain to be determined if these searchable pages will actually be used by researchers at the History Center. Before a decision is made to make other historical holdings searchable we should wait to see the extent to which these now searchable files are actually searched.

PDBullock - June 5, 2014
bullockpd11@verizon.net

APPENDIX A

Information Concerning the 15 Scrapbooks



Volume	# of Pages	Description
A	149	First 100 pages newspaper clippings from 1903, 1904, 1905 and 1906 The last 50 pages are mostly from 1800s
B	181	Newspaper clippings from the years 1906 and 1907
C	166	Newspaper clippings from the years 1907, 1908 and 1909
D	154	Newspaper clippings from the years 1910, 1911 and 1912
E	153	Newspaper clippings from the years 1913, 1914, 1915 and 1916
F	159	Newspaper clippings from the years 1916, 1917, 1918, 1919 and 1920
G	153	Newspaper clippings from the years 1920, 1921, 1922 and 1923
H	150	Newspaper clippings from the years 1923, 1924, 1925, 1926 and 1927
I	145	Newspaper clippings from the years 1927, 1928, 1929 and 1930
J	150	Newspaper clippings from the years 1930, 1931, 1932 and 1933
K	149	Newspaper clippings from the years 1933, 1934, 1935, 1936 and 1937
L	153	Newspaper clippings from the years 1937, 1938 and 1939 There are also many clippings from 1800s newspapers
M	150	Newspaper clippings from the years 1940 through 1953
Scraps	147	Newspaper clippings from 1898 through 1929 Programs for plays and other theater productions The last 30 pages contain mostly drawings and ads created by Oliver Sheppard.
No. 48	100	Newspaper clippings from the years 1934 through 1949 Put together by Oliver Sheppard

APPENDIX B

George S. Sheppard's Obituary

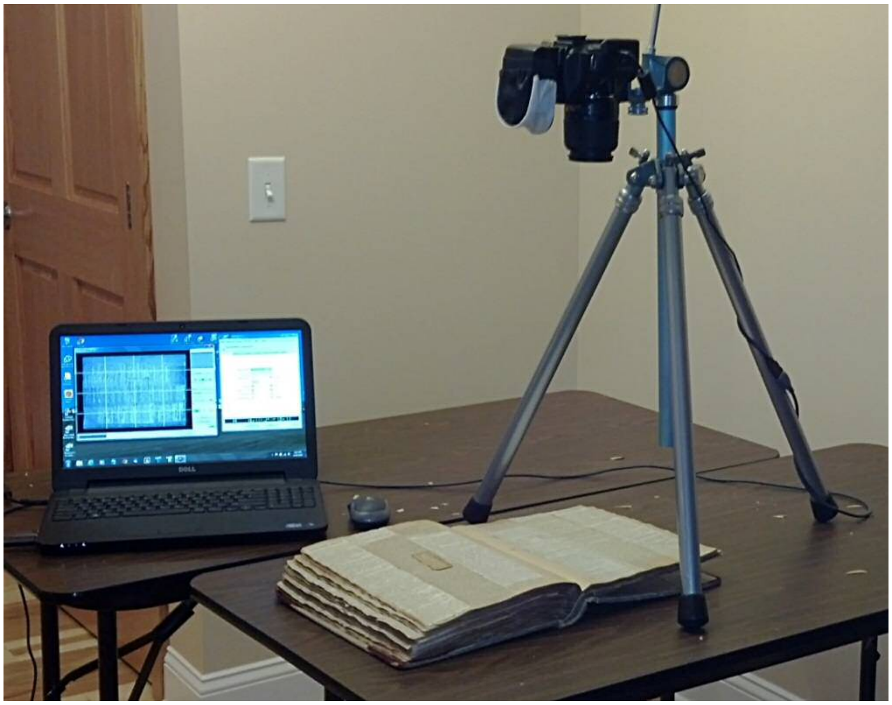
GEORGE S. SHEPPARD

"George S. Sheppard, 90, dean of Yates County lawyers, died in Soldiers and Sailors Memorial Hospital here about 3 p. m. (Sept. 25, 1945) after an Illness of several weeks. Born in Penn Yan Sept. 12, 1855, he was a graduate of Cornell University and of Columbia Law School. He was admitted to the New York State Bar without examination in 1877 and became a partner of William T. Morris in Penn Yan. In 1880 ill health compelled him to stop his office work, but he opened a law office in Penn Yan again in 1891. … In 1920 he was admitted to practice before the United States District Court in Canandaigua. In addition to his legal practice, Mr. Sheppard had been interested and active in community affairs, serving as secretary of the Soldiers and Sailors Memorial Hospital Association since its organization in 1919 and of the Penn Yan Hospital for three years before that. He was a member of St. Mark's Episcopal Church, being its senior warden for many years; a past master of Milo Lodge, F&AM, members of which will conduct the ritual of their organization at the grave; he was a past high priest of Penn Yan Chapter, RAM, and a past commander of Jerusalem Commandery, Knights Templar. He also was past district deputy of the 31st Masonic District of New York, as well as past grand of Keuka Lodge, IOOF. He is survived by a son, Oliver Sheppard of Penn Yan, and brother, Walter B. Sheppard of Denver, Colo. Funeral services will be in the Thayer Funeral Chapel in East Elm Street at 2:30 p. m Friday with the Rev. Hiram Rogers, rector of St. Mark's Episcopal Church, officiating."

A paragraph from a newspaper article published shortly before his death on his ninetieth birthday reads, "His collection of scrapbooks of historical data are famous throughout the state, and he is considered an authority on Yates county historical facts, having a remarkably clear and accurate memory of events over three-quarters of a century."

APPENDIX C

Capturing the Image Files



Equipment - A digital camera was mounted on a tripod pointing down, the tripod placed on a table and a page of the scrapbook placed within the view of the camera. An attached flash unit provided the lighting although flood lamps would have done as well if not better. See photo to the right. The camera was connected to a computer via a USB port and software on the computer enabled the camera live view to be shown on the monitor. The camera shutter release was controlled by the computer and the image files stored directly into computer memory.

Camera Settings

Camera: Nikon D5200
Monochrome (grayscale): Sharpness 9 and Contrast + 3
Aperture priority: f 11
Shutter speed: 1/60 sec.
ISO sensitivity: 200
Exposure compensation: + 1.0
Matrix metering
Auto focus area mode: 9 points
Image size: largest 4000 by 6000 pixels
Image quality for 14 of the scrapbooks: JPEG fine - compression ratio 1:4
Image quality for “No.48” scrapbook: RAW and later converted to TIFF image files

Does a camera have enough pixels to provide an image amenable to OCRing?

Size of document: 8 ½ by 11 ½ inches
Camera sensor size: 4000 by 6000 pixels
Limiting dimension of document: 8 ½ inches
4000 pixels / 8.5 inches = 471 pixels per inch (ppi or dpi)
The recommended ppi for viable OCRing is 300
Small typesets (8 points for example) often found in old newspapers may require a higher ppi

APPENDIX D

OCR of the Image File and Creating PDF files

ABBYY Finewriter 11 was used - highly recommended by independent software evaluators.

Settings within Finewriter for document handling:
Document type - Fax
Color mode - Full color

Settings within Finewriter for Opening the image file
General - Automatically read acquired page images
Image preprocessing - enable image preprocessing and detect page orientation

Settings within Finewriter for Reading
Reading mode - Thorough reading
Training - Use only built-in patterns

Settings within Finewriter for saving PDF files
Save mode - make sure Enable Tagged PDF is checked - VERY IMPORTANT!!
Exact Copy

Computer Memory Needed in this Project

Storage of Image Files
2200 image files at 14 MB each = 31000 MB = 31 GB

Storage of PDF Files
2200 PDF page files at .4 MB each = 880 MB = 0.9 GB
15 PDF volume files at 60 MB each = 900 MB = 0.9 GB

Storage of Finereader Document Files
15 Finereader Files at 6 GB each = 90 GB

Total
More than 120 GB

APPENDIX E
Details of Sampling Study to Estimate Word Error Rate

Scrapbook Volume	Sample Size	Misses	% Misses	95% Conf. Interval
A	50	3	6	1 - 18
B	50	3	6	1 - 18
C	50	2	4	1 - 13
D	50	2	4	1 - 13
E	50	4	8	3 - 19
F	50	5	10	4 - 21
G	50	1	2	1 - 10
H	50	2	4	1 - 13
I	50	3	6	1 - 18
J	50	5	10	4 - 21
K	50	4	8	3 - 19
L	50	3	6	1 - 18
M	50	2	4	1 - 13
Scraps	50	5	10	4 - 21
No. 48	50	2	4	1 - 13
Total	**750**	**46**	**6.1**	**4.6 - 8.1**

These results show that if the sample of 750 words selected truly represents the estimated more than 2 million words contained in the 15 scrapbooks, the word error rate for all 2 million is between 4.6 % and 8.1 % with 95% confidence. The error rate for all of the words could be as high as 8.1 %. Assuming the worst case that the word error rate is in fact 8.1% (hit rate of 91.9%) in our application, how does this compare with results from other studies.

There is an interesting and informative article in D-Lib Magazine, March/April 2009 titled "How Good Can It Get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs" by Rose Holley, Manager - Australian Newspaper Digitisation Program, National Library of Australia. This article about OCRing old newspapers contains a table that speaks to good, average and poor OCR "accuracy."

"*The question of what is acceptable* (character hit rate or accuracy) *has not been answered, but in speaking to other libraries and OCR contractors, it was generally agreed for historic newspapers that when we talk about good, average and bad OCR we mean:*

Good OCR accuracy	*= 98-99% accurate*	*(1-2% of OCR incorrect)*
Average OCR accuracy	*= 90-98% accurate*	*(2-10% of OCR incorrect)*
Poor OCR accuracy	*= below 90% accurate*	*(more than 10% of OCR incorrect)*

Our results showed, at worst, a 91.9 % hit rate relative to searching for a word. How does this compare to the above figures of character hit rate? The words we searched on were on average six characters long. To get a word of six characters correct 91.9 % of the time we need to have a character hit rate that is the 6th root of 0.919 or $\sqrt[6]{0.919}$ = .986 or 98.6%. That's true since to get the word correct we need to get the first character correct and the second character correct and the third character correct and so forth up through the sixth character. Since in our project our character hit rate was 98.6 % it falls into the article's Good OCR accuracy category.