Sound, to Print, to Cloud: saving the news with the Milo Ryan Phonoarchive Finding Aid

Tigh Bradley, MLIS
University of Washington

In short

This is a project to transform a print catalog/finding aid into a digital one, with accuracy beyond the power of Optical Character Recognition (OCR) techniques alone.

The gap

The University of Washington Media Arcade/Reformatting and Digitization Lab is creating digital copies of approximately 1500 hours of audio from the Milo Ryan Phonoarchive. This sound collection includes the largest extent corpus of daily news broadcasts from World War II in the US. There are notable contributions by Edward R. Murrow and "Murrow's Boys", who transmitted news reports live from the theater of war in Europe.

The UW is preserving the sound recordings in a digital format. Rich descriptions of the contents of the tapes are needed for the collection to be discovered and used by researchers and members of the public.

Fortunately, the required information exists in print. However, simply scanning the pages and applying OCR does not provide sufficient quality reproduction of the text. This project utilizes algorithmic processing of the scanned text so that accurate, descriptive digital metadata can be attached to the sound recordings, greatly enhancing the approachability and utility of the historic radio broadcasts.

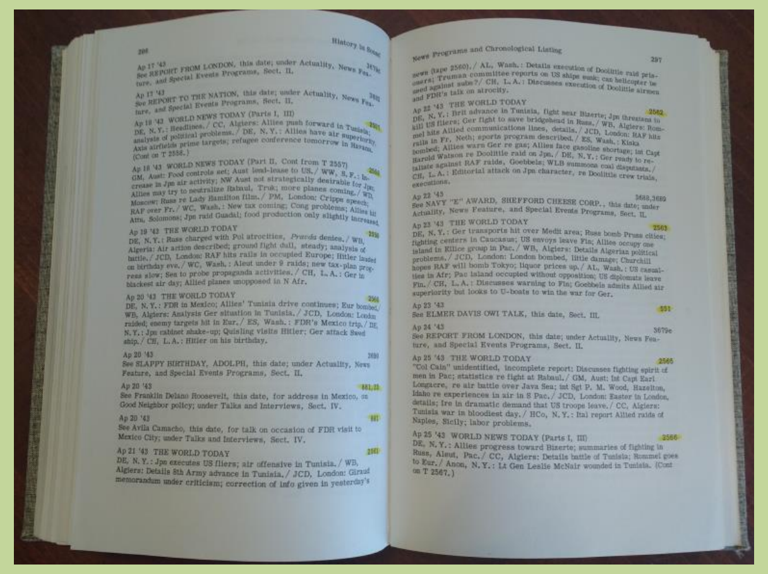
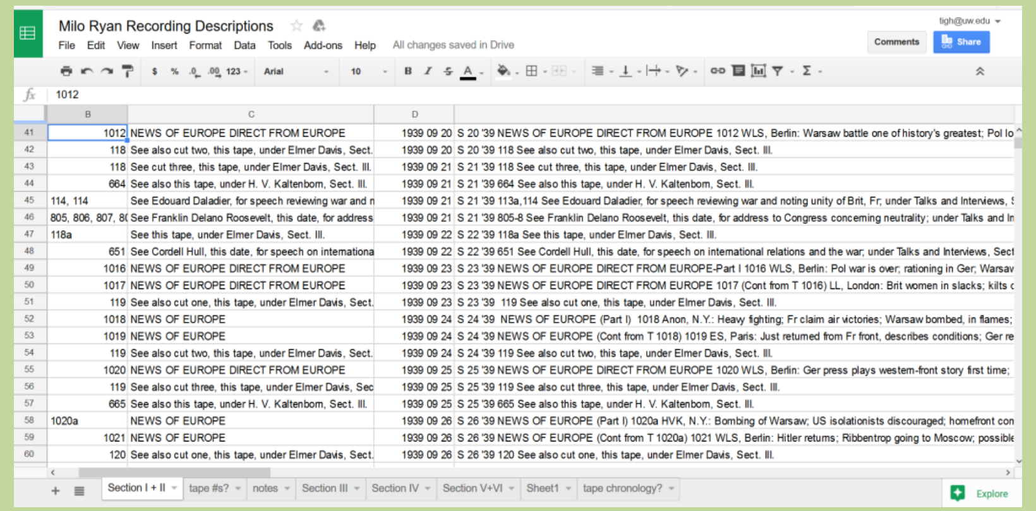


Figure 1. Original print finding aid.



Milo Ryan Recording Descriptions

fx 1012

	B	C	D	
41	1012	NEWS OF EUROPE DIRECT FROM EUROPE	1939 09 20	S 20 '39 NEWS OF EUROPE DIRECT FROM EUROPE 1012 WLS, Berlin: Warsaw battle one of history's greatest; Pol lo
42	118	See also cut two, this tape, under Elmer Davis, Sect.	1939 09 20	S 20 '39 118 See also cut two, this tape, under Elmer Davis, Sect. III.
43	118	See cut three, this tape, under Elmer Davis, Sect. III.	1939 09 21	S 21 '39 118 See cut three, this tape, under Elmer Davis, Sect. III.
44	664	See also this tape, under H. V. Kaltenborn, Sect. III.	1939 09 21	S 21 '39 664 See also this tape, under H. V. Kaltenborn, Sect. III.
45	114, 114	See Edouard Daladier, for speech reviewing war and n	1939 09 21	S 21 '39 113a,114 See Edouard Daladier, for speech reviewing war and noting unity of Brit, Fr; under Talks and Interviews, S
46	805, 806, 807, 80	See Franklin Delano Roosevelt, this date, for address	1939 09 21	S 21 '39 805-8 See Franklin Delano Roosevelt, this date, for address to Congress concerning neutrality; under Talks and In
47	118a	See this tape, under Elmer Davis, Sect. III.	1939 09 22	S 22 '39 118a See this tape, under Elmer Davis, Sect. III.
48	651	See Cordell Hull, this date, for speech on internationa	1939 09 22	S 22 '39 651 See Cordell Hull, this date, for speech on international relations and the war; under Talks and Interviews, Sect
49	1016	NEWS OF EUROPE DIRECT FROM EUROPE	1939 09 23	S 23 '39 NEWS OF EUROPE DIRECT FROM EUROPE-Part I 1016 WLS, Berlin: Pol war is over; rationing in Ger; Warsav
50	1017	NEWS OF EUROPE DIRECT FROM EUROPE	1939 09 23	S 23 '39 NEWS OF EUROPE DIRECT FROM EUROPE 1017 (Cont from T 1016) LL, London: Brit women in slacks; kilts c
51	119	See also cut one, this tape, under Elmer Davis, Sect.	1939 09 23	S 23 '39 119 See also cut one, this tape, under Elmer Davis, Sect. III.
52	1018	NEWS OF EUROPE	1939 09 24	S 24 '39 NEWS OF EUROPE (Part I) 1018 Anon, N.Y.: Heavy fighting; Fr claim air victories; Warsaw bombed, in flames;
53	1019	NEWS OF EUROPE	1939 09 24	S 24 '39 NEWS OF EUROPE (Cont from T 1018) 1019 ES, Paris: Just returned from Fr front, describes conditions; Ger re
54	119	See also cut two, this tape, under Elmer Davis, Sect.	1939 09 24	S 24 '39 119 See also cut two, this tape, under Elmer Davis, Sect. III.
55	1020	NEWS OF EUROPE DIRECT FROM EUROPE	1939 09 25	S 25 '39 NEWS OF EUROPE DIRECT FROM EUROPE 1020 WLS, Berlin: Ger press plays western-front story first time;
56	119	See also cut three, this tape, under Elmer Davis, Sec	1939 09 25	S 25 '39 119 See also cut three, this tape, under Elmer Davis, Sect. III.
57	665	See also this tape, under H. V. Kaltenborn, Sect. III.	1939 09 25	S 25 '39 665 See also this tape, under H. V. Kaltenborn, Sect. III.
58	1020a	NEWS OF EUROPE	1939 09 26	S 26 '39 NEWS OF EUROPE (Part I) 1020a HVK, N.Y.: Bombing of Warsaw; US isolationists discouraged; homefront con
59	1021	NEWS OF EUROPE	1939 09 26	S 26 '39 NEWS OF EUROPE (Cont from T 1020a) 1021 WLS, Berlin: Hitler returns; Ribbentrop going to Moscow; possible
60	120	See also cut one, this tape, under Elmer Davis, Sect.	1939 09 26	S 26 '39 120 See also cut one, this tape, under Elmer Davis, Sect. III.

Figure 2. New Digital finding aid.

Methods and Materials

The original print catalog, *History in Sound* by Professor Milo Ryan, comprises some 580 pages of printed text, which has been scanned and processed with OCR. This text contains numerous errors not appearing in the original book. These include extra breaks between lines and omitted or incorrect characters.

The OCR processed text was studied for systemic errors, and Python scripts were implemented to eliminate them. The Python scripts were also utilized to extract metadata features such as program titles and transmission dates from the body text.

Results

Through a combination of programming and manual inspection of the catalog entries, errors in the text were reduced to fewer than 1 out of 100 catalog fields based on randomized sample tests. The final product is a database of 4,763 digital catalog records of the tapes which faithfully reproduce the text from the print catalog.

Error Type Corrected	Count
Extra line breaks	~7000
Punctuation interrupting a word	~2000
Lowercase s for uppercase S	~500
Numeral 0 for letter O	~500
Wrong character(s) instead of letter	~250
Bracket { instead of parenthesis (	~200
Whitespace instead of letter	~200
Other	~200

Table 1. Summary of errors in OCR text repaired during this project.

Figure 3. The original tapes are being digitized via reel to reel machines.

Discussion

Several steps were necessary for this project:

1) Evaluate the original OCR text by eye.
2) Use of regular expressions to find-and-replace pervasive errors in the text.
3) Programming of scripts to fix more intricate pervasive errors and extract metadata from program descriptions.
4) Human review of catalog entries to assure matching to original text. While algorithmic methods improved the OCR text, about 20% of the entries had to be hand corrected after processing.

Impact and Future Work

The completed finding aid is being used by the UW Media Arcade/Reformatting and Digitization Lab for already, to catalog the archival holdings of the Milo Ryan Phonoarcive online at ArchiveSpace.

The methods developed in this project may be used to create digital versions of other print catalogs, finding aids, dictionaries, and encyclopedias. They will be especially useful for encoding the second print finding aid for this collection, *History in Sound Part 2*.

Contact

Tigh Bradley
University of Washington Information School
Email: tigh@uw.edu