End-to-End Robokeying of "Born Paper" Obituaries

Patrick Schone, Heath Nielson (FamilySearch, patrickjohn.schone@ldschurch.org, nielsonhe@familysearch.org)

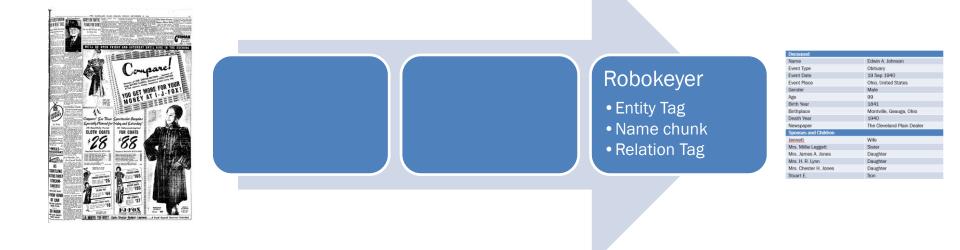




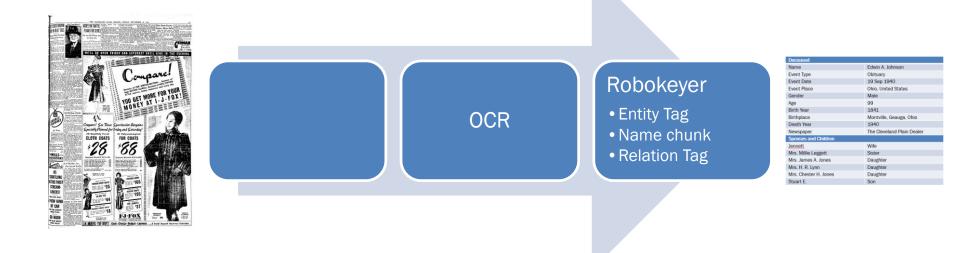


Deceased					
Name	Edwin A. Johnson				
Event Type	Obituary				
Event Date	19 Sep 1940				
Event Place	Ohio, United States				
Gender	Male				
Age	99				
Birth Year	1841				
Birthplace	Montville, Geauga, Ohio				
Death Year	1940				
Newspaper	The Cleveland Plain Dealer				
Spouses and Children					
Jennett	Wife				
Mrs. Millie Leggett	Sister				
Mrs. James A. Jones	Daughter				
Mrs. H. R. Lynn	Daughter				
Mrs. Chester H. Jones	Daughter				
Stuart E.	Son				

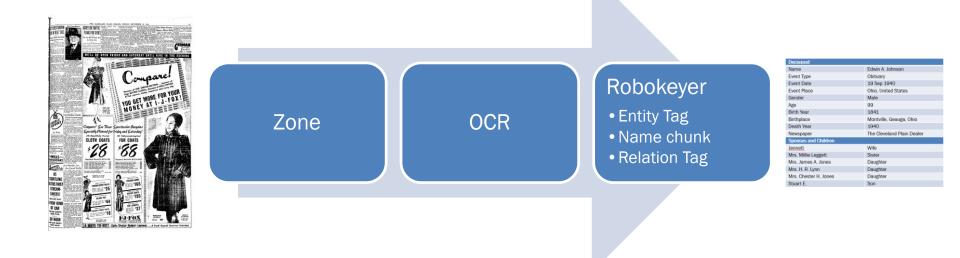














Zoning



TY'SBEST KINOWN By IN BLUE' DIES OUBLE w 7.1 MILLS **8** STARTLING **STHE FIRST** STREAM! UNERSI NEW KIN OF CAR HE THE LOWEST CO NASH

Archbishop Hold

Third Term Wron

Are Found Gu

ED IN SLIP (

THE CLEVELAND PLAIN DEALER, FRIDAY, SEPTEMBER 20

P	AGREE ON TRAFFIC AND A STATE OF THE ADDRESS OF THE	And there is the according to the second sec	For the followed out whice use is de- from the accurate law is designed. The memory will be addressly the the second second second second second with the exclusion. As use is it is the second second second second second to accurate the second second second to accurate the second	Ohio State Grotto Opens Akon Rolly The Pike Buck Reset Marco C. State State Marco C. State State State State Andreads and the State State Andread Andread Andreads and the State Andread Andread Andreads and the State Andread	of the Const Hade 21 an of or Alterna of the Samer of the
---	---	--	---	---	--





Zoning Challenges

n Capitol tional ba nt's fivesought to if shoring tion today nt across after he

gree from Dans tohe pr Provident desunvithree and ie giant, intra from e nations Merico C an plunge-Senator him hack sited Cald sly White Cold y adminis- anticipated gness to to pack h bolster-this line o Com-ico

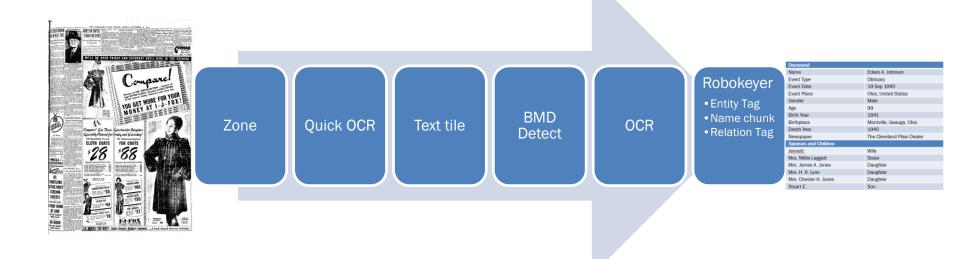
- Newspaper content is very dense
 - Distance between words within a column can be greater than distance between columns.



Content Filtering

- QuickOCR
 - Less accurate
 - More performant
- Text tiling
 - Group adjacent zones together
 - Uses a cosine similarity metric to predict when blocks from different zones should merge
- BMD detector
 - Identify any content containing Birth/Marriage/Death information
 - Uses support vector machines on ngrams of characters to predict which blocks of data appear to be BMDs and which appear to NOT be.







Results

- Proof of concept which met our expectations
- Would require more work to improve accuracy
- Production-based system would require 90-95% F-score
- We believe target is attainable

