Crowdsourced Transcription of Handwritten Mental Health Records

Unmil P. Karadkar
unmil@ischool.utexas.edu

The University of Texas at Austin
School of Information

Central State Hospital

- First mental health institution for African Americans
- Operational 1870 to present day
- Extensive administrative and medical records
- Records collection of historical significance
- Handwritten and printed documents

Challenge

Enable access to historical handwritten records without compromising patient privacy

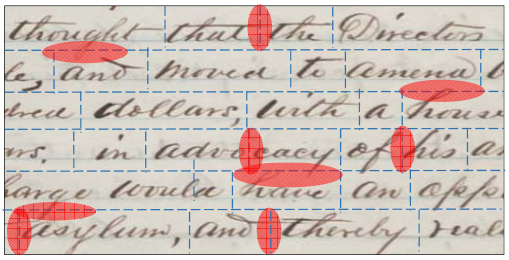
Approach

Segment documents into word images
Limit the context available to human readers
Crowdsource transcription of textual units

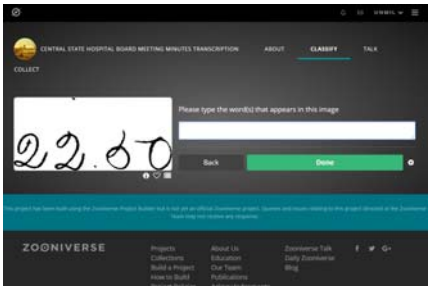
The Digital Collection

- ~500,000 pages
- 400 dpi, full color TIFFs, ~10 TB on disk
- Administrative records
  Acts of legislature, Land acquisition, Architectural blueprints, Annual reports to the state governor, **Board meeting minutes**
- Patient records
  Admission, Treatment, Furloughs, Death, Discharge, Ward books
- Institutional records
  Sign-in sheets, Newsletters, Photographs, News article clippings

Cursive, handwritten text is impossible to OCR and difficult to segment



Our segmentation algorithm correctly identifies around 90% of words, generates several false positives due to artifacts of scanning



Images, identifiers
Panoptes API
Text, user, quality data
MongoDB

Our transcription workflow uses Zooniverse, a free, open source crowdsourcing engine

Significance

Open source infrastructure (unlike reCaptcha)
First privacy-sensitive transcription software

Ongoing work

Improve text detection algorithms
Integrate validated text with document metadata

Acknowledgements

Andrew W. Mellon Foundation
Wuxi Li, Mona Mishra

National Association of State Mental Health Program Directors
Substance Abuse and Mental Health Services Administration
Institute for Urban Policy Research and Analysis, UT Austin