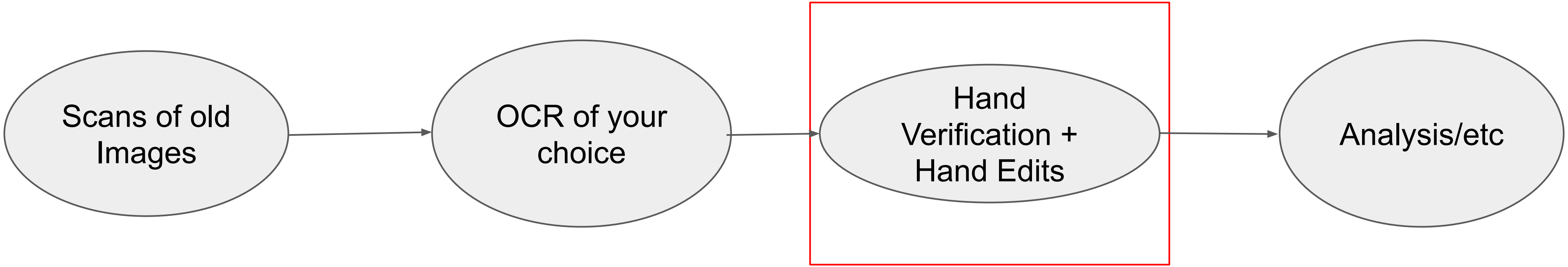
Record Linkage GUI

Todd Nobles, Natalie Turner, Terresa Tran, Honglam Van, Julia Zhu

Background

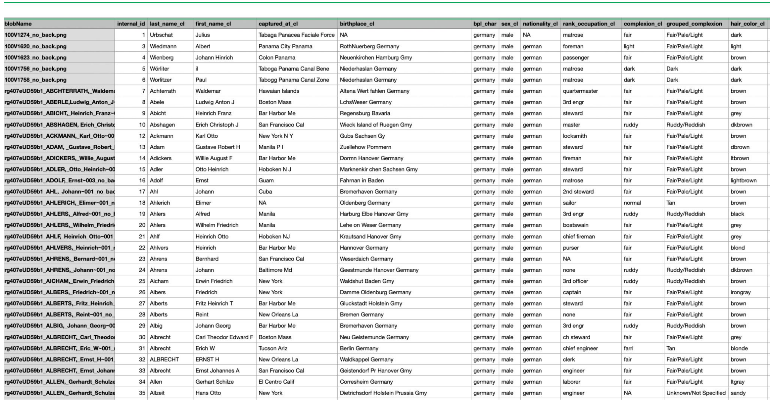
- Large scale quantitative analysis of archival data involves a number of steps that are tedious, error prone, and create difficulties for reproducibility
- Our tool assists with the first two of these problems by improving the workflow for users who are turning images of historical documents into analysis ready datasets



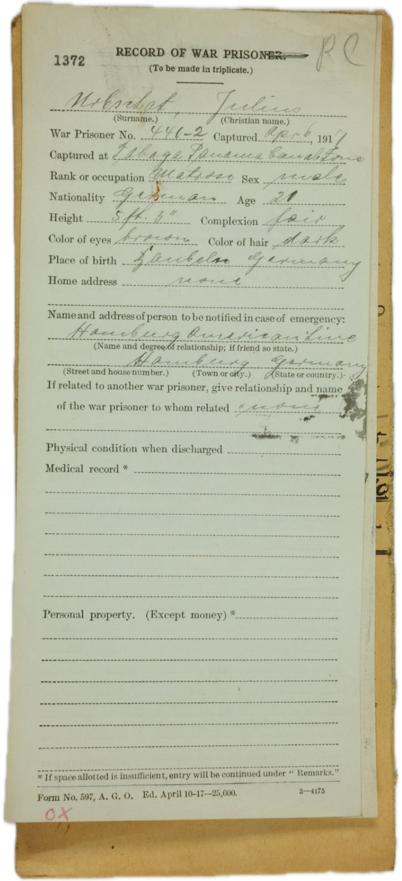
Data Used

We use two data sources:

1. Images of archival documents
2. A CSV of the computer extracted information from these images



blobName	internal_id	last_name_cl	first_name_cl	captured_at_cl	birthplace_cl	bpl_char	sex_cl	nationality_cl	rank_occupation_cl	complexion_cl	grouped_complexion	hair_color_cl
100V1274_no_back.png	1	Urbschat	Julius	Tabaga Panacea Faciale Force	NA	germany	male	NA	matrose	fair	Fair/Pale/Light	dark
100V1620_no_back.png	3	Wiedmann	Albert	Panama City Panama	RothNuerberg Germany	germany	male	german	foreman	light	Fair/Pale/Light	light
100V1623_no_back.png	4	Wienberg	Johann Hinrich	Colon Panama	Neuenkirchen Hamburg Gmy	germany	male	german	passenger	fair	Fair/Pale/Light	brown
100V1756_no_back.png	5	Wörliter	il	Taboga Panama Canal Bene	Niederhaslan Germany	germany	male	german	matrose	dark	Dark	dark
100V1758_no_back.png	6	Worlitzer	Paul	Tabogg Panama Canal Zone	Niederhaslan Germany	germany	male	german	matrose	dark	Dark	dark
rg407eUD59b1_ABCHTERRATH,_Waldema	7	Achterrath	Waldemar	Hawaian Islands	Altena Wert fahlen Germany	germany	male	german	quartermaster	fair	Fair/Pale/Light	brown
rg407eUD59b1_ABERLE,Ludwig_Anton_J-	8	Abele	Ludwig Anton J	Boston Mass	LchsWeser Germany	germany	male	german	3rd engr	fair	Fair/Pale/Light	brown
rg407eUD59b1_ABICHT,_Heinrich_Franz~	9	Abicht	Heinrich Franz	Bar Harbor Me	Regensburg Bavaria	germany	male	german	steward	fair	Fair/Pale/Light	grey
rg407eUD59b1_ABSHAGEN, Erich_Christo	10	Abshagen	Erich Christoph J	San Francisco Cal	Wieck Island of Ruegen Gmy	germany	male	german	master	ruddy	Ruddy/Reddish	dkbrown
rg407eUD59b1_ACKMANN,_Karl_Otto~00	12	Ackmann	Karl Otto	New York N Y	Gubs Sachsen Gy	germany	male	german	locksmith	fair	Fair/Pale/Light	brown
rg407eUD59b1_ADAM, _Gustave_Robert_	13	Adam	Gustave Robert H	Manila P I	Zuellehow Pommern	germany	male	german	steward	fair	Fair/Pale/Light	dbrown
rg407eUD59b1_ADICKERS,_Willie_August	14	Adickers	Willie August F	Bar Harbor Me	Dormn Hanover Germany	germany	male	german	fireman	fair	Fair/Pale/Light	ltbrown
rg407eUD59b1_ADLER,_Otto_Heinrich~00	15	Adler	Otto Heinrich	Hoboken N J	Marknenkir chen Sachsen Gmy	germany	male	german	steward	fair	Fair/Pale/Light	brown
rg407eUD59b1_ADOLF,_Ernst~003_no_ba	16	Adolf	Ernst	Guam	Fahrnan in Baden	germany	male	german	matrose	fair	Fair/Pale/Light	lightbrown
rg407eUD59b1_AHL,_Johann~001_no_bac	17	Ahl	Johann	Cuba	Bremerhaven Germany	germany	male	german	2nd steward	fair	Fair/Pale/Light	brown
rg407eUD59b1_AHLERICH,_Elimer~001_n	18	Ahlerich	Elimer	NA	Oldenberg Germany	germany	male	german	sailor	normal	Tan	brown
rg407eUD59b1_AHLERS,_Alfred~001_no_	19	Ahlers	Alfred	Manila	Harburg Elbe Hanover Gmy	germany	male	german	3rd engr	ruddy	Ruddy/Reddish	black
rg407eUD59b1_AHLERS,_Wilhelm_Friedri	20	Ahlers	Wilhelm Friedrich	Manila	Lehe on Weser Germany	germany	male	german	boatswain	fair	Fair/Pale/Light	grey
rg407eUD59b1_AHLF,_Heinrich_Otto~001_	21	Ahlf	Heinrich Otto	Hoboken NJ	Krautsand Hanover Gmy	germany	male	german	chief fireman	fair	Fair/Pale/Light	grey
rg407eUD59b1_AHLVERS,_Heinrich~001_	22	Ahlvers	Heinrich	Bar Harbor Me	Hannover Germany	germany	male	german	purser	fair	Fair/Pale/Light	blond
rg407eUD59b1_AHRENS,_Bernard~001_n	23	Ahrens	Bernhard	San Francisco Cal	Weserdaich Germany	germany	male	german	NA	fair	Fair/Pale/Light	brown
rg407eUD59b1_AHRENS,_Johann~001_no	24	Ahrens	Johann	Baltimore Md	Geestmunde Hanover Germany	germany	male	german	none	ruddy	Ruddy/Reddish	dkbrown
rg407eUD59b1_AICHAM,_Erwin_Friedrich	25	Aicham	Erwin Friedrich	New York	Waldshut Baden Gmy	germany	male	german	3rd officer	ruddy	Ruddy/Reddish	brown
rg407eUD59b1_ALBERS,_Friedrich~001_n	26	Albers	Friedrich	New York	Damme Oldenburg Germany	germany	male	german	captain	fair	Fair/Pale/Light	irongray
rg407eUD59b1_ALBERTS,_Fritz_Heinrich_	27	Alberts	Fritz Heinrich T	Bar Harbor Me	Gluckstadt Holstein Gmy	germany	male	german	steward	fair	Fair/Pale/Light	brown
rg407eUD59b1_ALBERTS,_Reint~001_no_	28	Alberts	Reint	New Orleans La	Bremen Germany	germany	male	german	none	fair	Fair/Pale/Light	brown
rg407eUD59b1_ALBIG,_Johann_Georg~00	29	Albig	Johann Georg	Bar Harbor Me	Bremerhaven Germany	germany	male	german	3rd engr	ruddy	Ruddy/Reddish	brown
rg407eUD59b1_ALBRECHT,_Carl_Theodo	30	Albrecht	Carl Theodor Edward F	Boston Mass	Neu Geistemunde Germany	germany	male	german	ch steward	fair	Fair/Pale/Light	grey
rg407eUD59b1_ALBRECHT,_Eric_W~001_	31	Albrecht	Erich W	Tucson Ariz	Berlin Germany	germany	male	german	chief engineer	farri	Tan	blonde
rg407eUD59b1_ALBRECHT,_Ernst_H~001_	32	ALBRECHT	ERNST H	New Orleans La	Waldkappel Germany	germany	male	german	clerk	fair	Fair/Pale/Light	brown
rg407eUD59b1_ALBRECHT,_Ernst_Johann	33	Albrecht	Ernst Johannes A	San Francisco Cal	Geistendorf Pr Hanover Gmy	germany	male	german	engineer	fair	Fair/Pale/Light	brown
rg407eUD59b1_ALLEN,_Gerhardt_Schulze	34	Allen	Gerhart Schilze	El Centro Calif	Corresheim Germany	germany	male	german	laborer	fair	Fair/Pale/Light	ltgray
rg407eUD59b1_ALLEN,_Gerhardt_Schulze	35	Allzeit	Hans Otto	New York	Dietrichsdorf Holstein Prussia Gmy	germany	male	german	engineer	NA	Unknown/Not Specified	sandy



1372 RECORD OF WAR PRISONER. RC

(To be made in triplicate.)

Urbschat, Julius
(Surname.) (Christian name.)

War Prisoner No. 440-2 Captured Apr 6, 1917

Captured at Taboga Panama Canal Zone

Rank or occupation Matrose Sex male

Nationality German Age 21

Height 5 ft. 3" Complexion fair

Color of eyes brown Color of hair dark

Place of birth Zaubeln Germany

Home address none

Name and address of person to be notified in case of emergency:

Hamburg American Line
(Name and degree of relationship; if friend so state.)

Hamburg Germany
(Street and house number.) (Town or city.) (State or country.)

If related to another war prisoner, give relationship and name of the war prisoner to whom related none

Physical condition when discharged

Medical record *

Personal property. (Except money) *

* If space allotted is insufficient, entry will be continued under "Remarks."

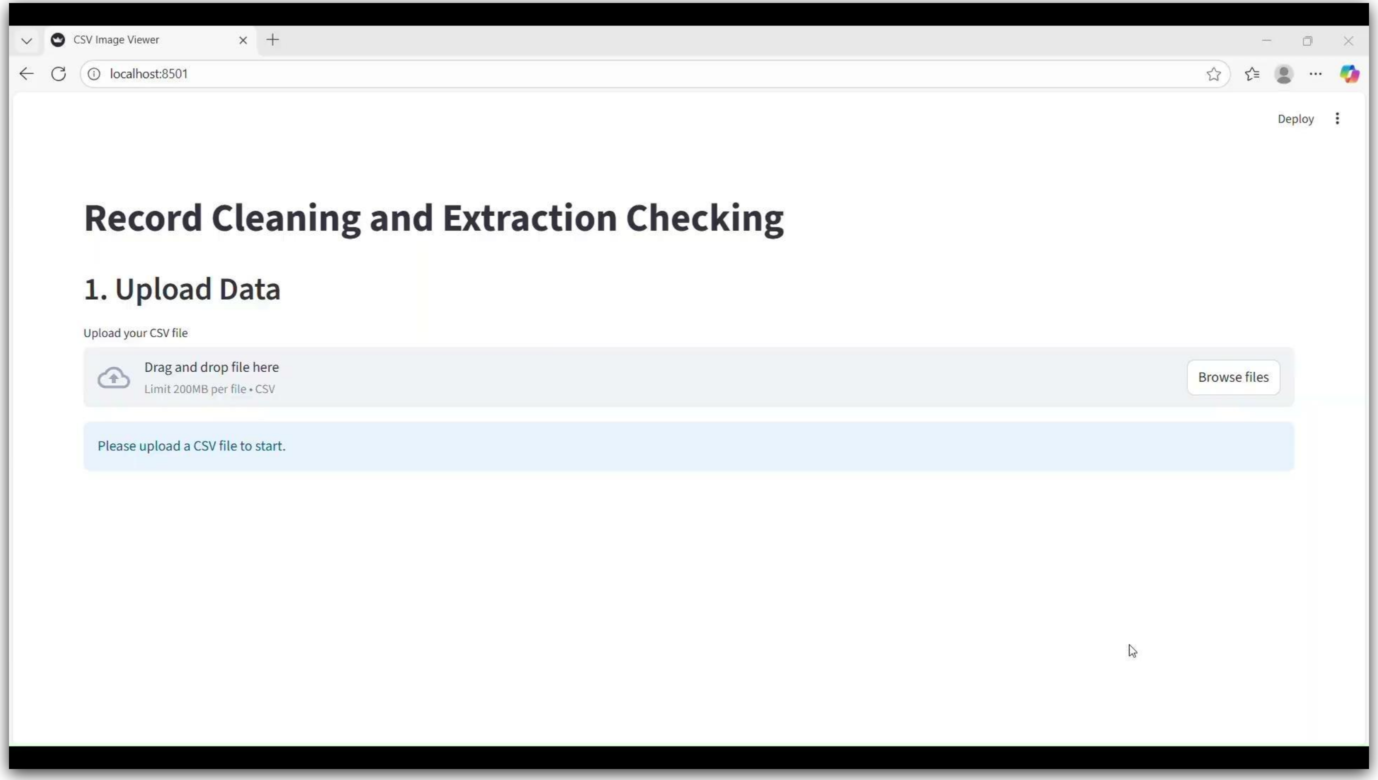
Form No. 597, A. G. O. Ed. April 10-17—25,000. 3—4175

OX

Use Cases

- Researchers working on **digitizing** and **transcribing** historical forms
    - Compare side by side text fields extracted from an image and the original image for accuracy
    - Manually edit the text for a field that was extracted from an image and save a new dataset with the corrected data
- Researchers with **little to no experience in coding**
    - Simple interface that allows user to quickly check information extracted from images without needing to work with file paths or code

Demo



Record Cleaning and Extraction Checking

1. Upload Data

Upload your CSV file

Drag and drop file here

Limit 200MB per file • CSV

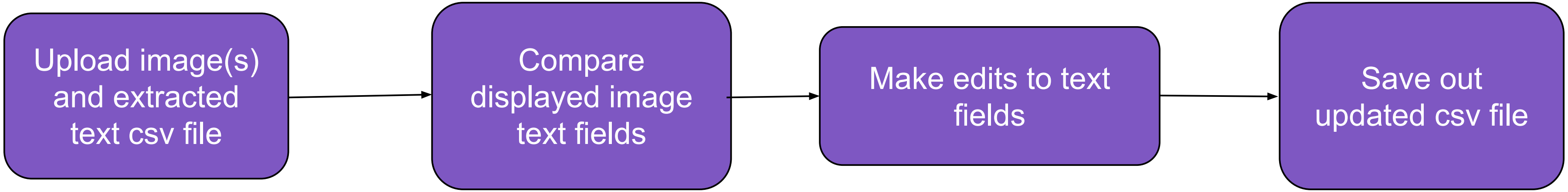
Browse files

Please upload a CSV file to start.

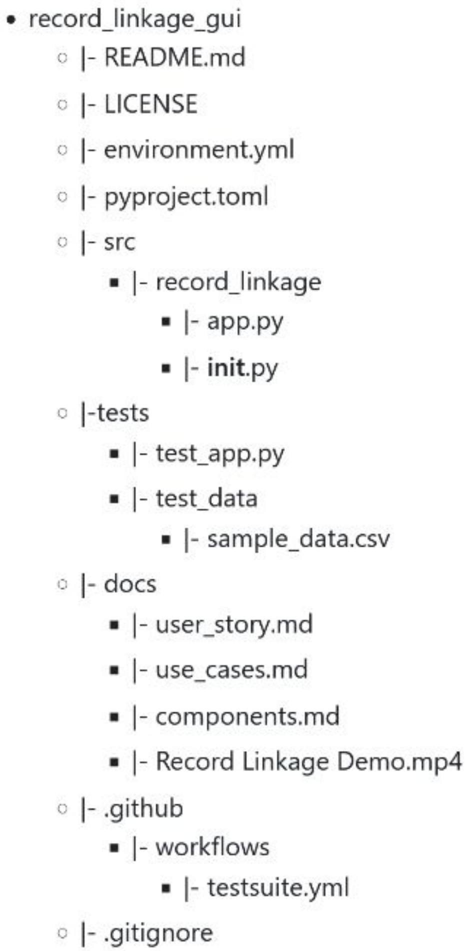
Design

Core components:

- User file upload
- Text display
- Image display
- Editable text fields


Upload image(s) and extracted text csv file
Compare displayed image text fields
Make edits to text fields
Save out updated csv file

Repository Structure



- record_linkage_gui
  - |- README.md
  - |- LICENSE
  - |- environment.yml
  - |- pyproject.toml
  - |- src
    - |- record_linkage
      - |- app.py
      - |- **init**.py
  - |-tests
    - |- test_app.py
    - |- test_data
      - |- sample_data.csv
  - |- docs
    - |- user_story.md
    - |- use_cases.md
    - |- components.md
    - |- Record Linkage Demo.mp4
  - |- .github
    - |- workflows
      - |- testsuite.yml
  - |- .gitignore

Lessons Learned and Future Work

- Sometimes documentation of specific components of a UI framework is not sufficient for newcomers to the tool. Full working examples are more helpful
- Our timing was off: we moved to testing in class before we had a UI, so we had to rush to build something testable.
- Our testing didn't always fit the standard testing discussed in class
    - We tested underlying functions with pytest, but to test Streamlit's UI behavior we did manual verification
- **Future work**
    - Improving scalability, currently a user would be limited to as many images as their RAM can store in working memory
    - Building the second tool in the Record Linkage Toolkit to help users adjudicate between potential matches across datasets by hand
    - Further testing once Streamlit has more capabilities to mock user inputs and interactions with the application