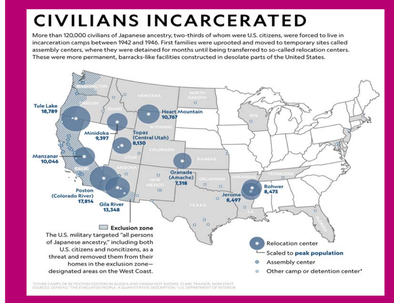
Seeing the Forest and the Trees: Creating Controlled Vocabularies for Historical Collections

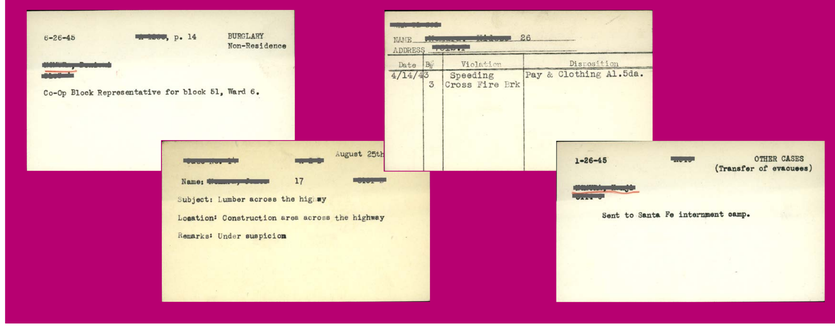
Margaret Rose Hunt (MLIS), **Hannah Frisch** (MLIS), **Hilary Shiue** (MLIS), **Tahura Turabi** (InfoSci / Japanese)

Historical Context



Executive Order 9066 created exclusion zones for citizens and resident aliens on the United States' West Coast. One of these internment camps was Tule Lake. These WRA camps were hastily organized, but the exclusionary nature of Tule Lake resulted in documentation about incidents that occurred in the camp between 1942-1946.

Example Documentation



Method

The uncleaned text in the spreadsheet was taken from the OCR process. The historical record was maintained by creating and editing a duplicate column of the spreadsheet, labeled *"New Offense."* Data cleaning was performed manually through OpenRefine, an open source software for data cleaning, using the original controlled vocabulary, while also adding terms to the controlled vocabulary spreadsheet as we encountered new terms.

Group members would convene weekly and discuss categories that were difficult to understand or lacked detail. When adjusting terms, each incident was given a **General Offense Category** and a **Specific Offense Category**, represented as General Offense_Specific Offense (for example: *Theft_Petty*). Once the data of all boxes were cleaned, they were all compiled into a large spreadsheet and cleaned en masse. The increased standardization meant that the terms that needed to be adjusted after the first round of cleaning would be more standardized.

OpenRefine: Clustered and Cleaned Example

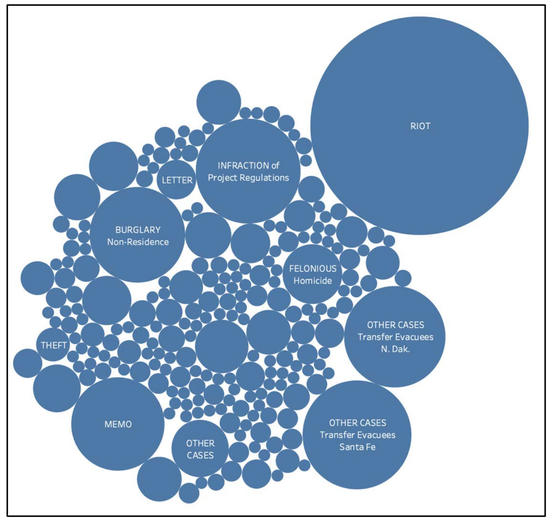
Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
2	3	• VIOLATION LIQUOR LAWS Possession & Manufacturing (2 rows) • VIOLATION LIQUOR LAWS Manufacturing & Possession (1 rows)	☑	Violation of Liquor Laws_Posse
2	29	• DISORDERLY Conduct Disturbance Peace (28 rows) • DISORDERLY CONDUCT Disturbance Peace (1 rows)	☑	Disorderly Conduct_Disturbing
2	2	• OTHER OFFENSES Violation oF Federal Selective Service Act (1 rows) • OTHER OFFENSES Violation of Federal Selective Service Act (1 rows)	☑	Federal Violation_Selective Se

OFFENSE	NEW.OFFENSE
OTHER CASES Record only	Administrative_Record
Out of bounds	Violation of Project Regulations_Trespassing
ACCIDENT M.V.	Accident_Motor Vehicle
OTHER CASES Anonymous Letter	Administrative_Camp Report

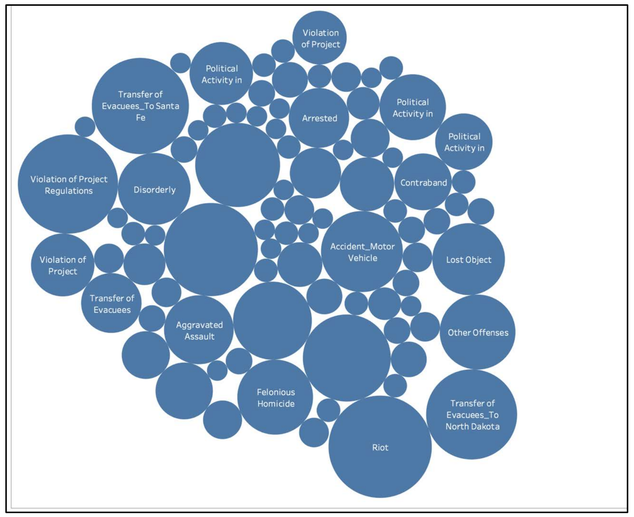
OpenRefine: Old and New Offense Columns after Cleaning

Cleaned General and Specific Offense Categories

Before Cleaning



After Cleaning



No incident lines were deleted in this process

Challenges

Human-created data, often intended for use in specific situations, tend to lack a standardized format, which entails difficulty in analysis due to its lack of computational amenability. Challenges that arose include:

Non-Standard terms, Abbreviations, and Capitalization

Inconsistencies in terminology resulted in a lack of association between incidents that are considered to be within the same category of offense. For example, "V.P.R" and "Violation of Project Regulations" both indicate the same category of offense, but initially would have been considered separate categories.

OFFENSE	NEW.OFFENSE
Viol. Project Regulations	Violation of Project Regulations
VIOL. PROJ. REG.	Violation of Project Regulations
V.P.R.	Violation of Project Regulations

Opaque Terms

Situations such as incidents lacking specific details or using obscure terminology created the problem of having several separate categories that should have been united under one. For example, thefts were often recorded as "[object name] Theft", designating the incident under that single category. This initially reduced the count for incidents under the "Theft" category, which provided an inaccurate reflection of that offense type.

OFFENSE
THEFT
THEFT Public bldg.
Theft Fr. Rooming-house
Theft Lumber
Theft Petty
Theft
Theft
THEFT Bicycle

Further Work & Impact

Controlled vocabularies can provide standardization that increases the likelihood of seeing the relations, patterns and frequencies of the records, incidents types, and historical events. With controlled vocabularies, we are able to minimize "noises" in data visualization; perform targeted research based on events, dates, etc.

Nevertheless, further details can be added to two large categories, *"Other Offenses"* and *"Violation of Project Regulations,"* in order to give more dimension to these cards. *"Administrative"* cards, that were related to camp operation, are another category that can be addressed in future work. While Incident Cards were created under official narratives, we can cross reference Administrative cards to see the power dynamics within the camp, i.e. if certain administrative order gave rise to riots, and to give voices to Japanese Americans who were incarcerated.

COLLEGE OF INFORMATION STUDIES

Dr. Richard Marciano

MARYLAND'S iSCHOOL