A Framework for Unlocking and Linking WWII Japanese American Incarceration Biographical Data

Lencia Beltran  
W. Madison Randall Library  
U. of North Carolina Wilmington  
Wilmington, USA  
beltanl@uncw.edu

Emily Ping O’Brien  
George C. Gordon Library  
Worcester Polytechnic Institute  
Worcester, USA  
opobrien@wpi.edu

Greg Jansen  
College of Information Studies  
University of Maryland  
College Park, USA  
jansen@umd.edu

Richard Marciano  
College of Information Studies  
University of Maryland  
College Park, USA  
marciano@umd.edu

Abstract—Entity Resolution (ER) is increasingly being used to identify and link names across archival collections. We describe a framework for unlocking and linking biographical data from WWII Japanese American Incarceration Camps using Entity Resolution and other computational approaches. We demonstrate the construction of social graphs that link people, places, and events and which support further scholarship and reveal hidden stories in historical events, especially given contested archival sources. Finally, we show the power of computational analysis to recreate event networks and represent movement of people using maps. This type of modeling is captured through interactive Jupyter Notebooks that integrate these various elements and document our interpretation of Japanese American experiences and events at the Tule Lake concentration camp.

Keywords—Computational Archival Science (CAS), Japanese American WWII Incarceration Camps, Entity Resolution, Digital Infrastructure, Linked Archives.

I. TOWARDS A PROSOPOGRAPHY OF WWII JAPANESE AMERICAN INCARCERATION

Our project explores digital history methods based on Computational Archival Science (CAS) [1] that use large quantities of data for historical research. In methodological terms, we are focusing on aspects of prosopography, which is sometimes defined as “a means of profiling any group of recorded persons linked by any common factor” [2]. One may thus think of prosopography as an investigation of a historical group (or of several interconnected groups) that is linked by some common factor, not focusing on a small number of individual biographies but using significant quantities of source material to carry out a collective study of their lives, looking at patterns of activities and interrelationships.

Computational prosopography has emerged with social network analysis, allowing us to reveal the interactions between individuals and groups. Our particular focus is the use of relationship mining, information extraction and visualization to investigate prosopographical questions in the history of Japanese American Incarceration, where the source materials are (a) very large, requiring the computational approach, and (b) very heterogeneous and ‘messy’. The aim is to generate and visualize a corpora of prosopographical information, represented as networks of entities such as people, places and events, in order to support the understanding of WWII Japanese American Incarceration.

We are interested in connecting the movements of displaced people inside individual camps and across a larger dizzying and little-known universe of fifty-nine incarceration sites (including Concentration Camps, Temporary Assembly Centers, Immigration Detention Stations, DOJ Internment Camps, U.S. Army Internment Camps, Citizen Isolation Centers, U.S. Federal Prisons, and additional facilities). We seek to identify how incarcerated peoples connected with each other, especially towards individual, small group, and communal acts of resistance, how they moved around surveillance practices, and how their movements were perceived by state authorities. We draw this hidden history of the WWII Japanese-American incarceration from newly released archival sources that, until now, had not been publicly available. March 28, 2021 marked the 75th anniversary of the closing of the last of these camps, Tule Lake in Northern California. Not unlike the Stasi Files which in 1992 made public the detailed records that were kept on thousands of East German citizens targeted as anti-government troublemakers, in the US, the War Relocation Authority (WRA) created Camp Files on individual incarcerees. We focus on the over 25,000 narrative reports prepared by camp investigators, police officers, and directors of internal security, relating cases of alleged “disorderly conduct, rioting, seditious behavior,” etc. at each of the ten major concentration camps. These reports include detailed information on the names and addresses in the camps of the persons involved, the time and place where the alleged incident occurred, an account of what happened, and a statement of action taken by the investigating officer, and even photos.

The picture that emerges by processing these records from the inside out by extracting the information content, and creating models of people and events across space and time, is a potential “history-changer” in unlocking a more complete story of what happened in the camps. Our work leverages advances made by Dr. Anne Knowles in her groundbreaking work on the Holocaust Geography Collaborative [3], where she demonstrates the value of creating spatial and temporal models that capture forced displacement patterns at various scales and the all-important of representation. Densho’s recently released “Sites of Shame” spatially represents the complex network of detention sites throughout the U.S. where more than 125,000 individuals of Japanese ancestry were held without trial.
In this historical moment when there is a heightened attention to misinformation, this project also highlights the transformation of information to knowledge [4]. This approach deepens our understanding of the way individuals, as well as groups connected with each other through movement, often towards resistance, in state-controlled environments, and the way knowledge is created. We believe this project could be a model for other narratives of “serial forced displacement” in America [5], a term that describes the consequences of repetitive, government-policy driven coercive upheaval of groups and includes: segregation, redlining, urban renewal, mass criminalization, disaster displacement, and more recently family separation immigration policy.

II. COMPUTATIONAL TREATMENTS OF BIOGRAPHICAL DATA

A seminal Biographical Data in a Digital World Conference Series was held from 2015 to 2019, which included three events: BD15, BD17, and BD19. This series was driven by computational advances which now offer biographers and historians new opportunities for both tracing individuals in isolation, in small groups, or as part of larger networks that can be researched and visualized in new ways. The inaugural BD15 conference identified four main themes [6]: (i) Bringing Biographical Data Online, (ii) Analyzing Biographical Data with Computational Methods, (iii) Creating Group Portraits and Networks and (iv) Visualization and Representation. It is worth noting that the questions that emerged included:

- How do we extract and analyze biographical data (with Named Entity Recognition--NER, or manually added metadata)?
- How do we link biographical datasets?
- How do we visualize biographical data online?
- How does the digital turn change traditional biography?
- How reliable is the biographical data (content and manipulation of the content) for what kind of research questions?
- How do we identify and distinguish different individuals within these huge datasets of biographical data (named entity disambiguation)?
- What bigger questions can we answer with all this data?

A particularly memorable paper from the BD15 event was the “Traces through Time” case-study [7], in which the authors used statistical methods to refine algorithms for linking biographical data. The project focuses on the challenge of connecting people with levels of accuracy through the identification and linking of individual names across collection boundaries. The approach differs from traditional techniques in that it attempts to automatically generate these linkages at scale, associating computed confidence measures to links using “fuzzy” matching of attribute values. The framework is meant to help accommodate data variations such as missing data, inaccurate dates (DOB of “14-7-2193”), alternative names (“Henry” vs. “Harry”), inferring the switching of letters in names (“Percy T.” vs. “Percy J.”), etc. The authors point out that in the case of name comparisons, the variation between name transcriptions can be due to many factors, including regional spelling variations, recording medium (handwritten vs. types), involuntary errors during data capture, spelling mistakes while writing or typing the original document, and involuntary errors during transcription, including those caused by difficult handwriting.

Conrad & Williams [8] further explore the use of graph databases to link “everyday” voices and people in archives by leveraging machine readable data and metadata to identify and display relationships between persons, places, dates, events, etc. across items and collections. This is a form of computational biography.

Cultural Institutions dedicated to the preservation of memory, based on traumatic events, typically emphasize remembrance through diaries, artifacts, and other documentation such as photographs, video, audio and written testimonies. This is eminently true for the Yad Vashem World Holocaust Remembrance Center in Jerusalem Israel, which continues to collect the names and stories of Holocaust victims. In fact, one of its major initiatives is the Names Recovery Project, which not only collects names, but tries to unlock complex biographical records using innovative technologies that integrate distributed and multilingual archival sources. Yad Vashem uses an Entity Resolution (ER) process to cluster names into groups of similar records, followed by an ML-based comparison process, enhanced with human crowdsourcing input [9].

Similarly, Densho (a Japanese term meaning “to pass on to the next generation,” or “to leave a legacy”), a grassroots memory organization in Seattle, WA, dedicated to preserving, educating, and sharing the story of World War II-era incarceration of Japanese Americans, has embarked on big biographical data projects. While its digital archive includes thousands of historic photographs, documents, newspapers, letters, visual oral histories and transcribed interviews, its primary holdings include intake and exit name registries from Incarceration Camps, resulting in complex name disambiguation challenges.

These memory organizations, large and small, are increasingly exploring the use of “emergent technologies” [10] to process “collections as data”; predicated on computational treatments of cultural heritage collections, using methodologies such as digital curation, Artificial Intelligence (AI), and Machine Learning (ML). Reflecting on the impact of rapid technological changes in the archival profession, Arie-Erez et al. highlight the benefits of computing points of interconnectivity between collections in order to link different types of materials and documentation through enriched access points including: names, subject headings, geographical locations, etc. [11].

In this paper we experiment with an Entity Resolution (ER) technique called fuzzy matching, to unlock and link biographical data from WWII Japanese American Incarceration Camps. Entity Resolution [12] is defined as the “Problem of identifying and linking/grouping different manifestations of the same real-world object, e.g. different ways of addressing the same person in text (names, email addresses, Facebook accounts).” Examples where ER is traditionally applied
include linking Census records, counter terrorism, and spam detection. The authors state that ironically ER has many duplicate names which include: record linkage, fuzzy match, reference reconciliation, approximate match, reference matching, entity clustering, object consolidation, household matching, object identification, merge/purge, householding, duplicate detection, coreference resolution, deduplication, identity uncertainty, hardening soft databases, and doubles.

III. THE CASE OF THE JAPANESE-AMERICAN WWII INCARCERATION CAMP RECORDS

In 1942 a network of 10 incarceration camps was created from California to Arkansas (see Fig. 1). Over 120,000 civilians of Japanese ancestry, two-thirds of whom were U.S. citizens, were deported into incarceration camps between 1942 and 1946. Major federal records associated with the War Relocation Authority (WRA), the agency established to handle the forced relocation and detention of Japanese-Americans during World War II, include:


2. The “Final Accountability Rosters of Evacuees at Relocation Centers, 1944-1946, also known as FAR, with camp outtake records of evacuees at the time of their final release or transfer.

3. Various WRA (Record Group 210) records with over 100 record series.

4. The National Archives “Internal Security Case Reports” Index Cards, a very significant WRA (Record Group 210 from 1941-47) records series.

The WRA records are permanent records of the National Archives, while other relevant records are with agencies such as the FBI, CIA, Department of Interior, etc.

![Fig. 1. Network of 10 WWII Incarceration Camps](https://www.nationalgeographic.com/magazine/2018/10/japanese-internment-then-now-portraits)

The Index Cards reference narrative reports prepared by camp investigators, police officers, and directors of internal security, relating cases of alleged “disorderly conduct, rioting, seditious behavior,” etc. at each of the 10 camps, with detailed information on the names and addresses in the camps of the persons involved, the time and place where the alleged incident occurred, an account of what happened, and a statement of action taken by the investigating officer. There are 25,045 index cards, 63% of which (15,648) come from the Tule Lake concentration camp. The Advanced Information Collaboratory (AIC) was granted research access to a number of these cards. Fig. 2 shows a typical Index Card. The date at the left of the top line is the “Incident Date”. To the right of the date is the “Case File ID” followed by the “Offense Type”. The person’s name in the second line is the index term (redacted in the image). To the right of the index term is the person’s “Family Number ID”, and to the far right his “Residence ID”. The information below that line is the Remarks section. It is important to note this information is not consistently recorded in the index cards and additional details are available in the associated incident reports.

![Fig. 2. “Incident” card for an incarcerated in the Tule Lake camp.](https://www.nationalgeographic.com/magazine/2018/10/japanese-internment-then-now-portraits)

IV. ENTITY RESOLUTION IN CAMP DATA

One of the principal challenges with the analysis of the Japanese-American internee data is that there is not a unified registry or any consistently applied unique identifiers for the people who were subjected to incarceration and forced migration. Establishing a unified registry of individuals and their families will enable further discrete links to archival context, such as events, incident reports, news stories, and internal and external camp locations. All such biographical and organizational context is composed of statements made about individuals and their families. In order to make such statements in data products, we need to refer to individuals and their families with unique identifiers. Fortunately, we have two large collections of individual and family data in WRA Form 26 (WRA) and the FAR. As stated previously, the WRA data contain camp intake records and the FAR data contain camp outtake records. If we can link the individuals referenced by these records, then we can create a unique identifier for each individual in the camps and establish the primary migration events that made up their wartime experience, namely their point of origin, arrival event, transfer events, and their exit event, including citizenship status and destination.

Entity resolution is the practice of detecting and creating explicit links between different data sources with partial information. In the case of these camp records, the entities we wish to resolve are people. We want to know when a person described by a FAR record is the same person that is described by a WRA record, or, by extension, any other FAR record.
There are many approaches to entity resolution, but they all involve looking for similarities in the information fields of both records. In the following sections we will explain how we achieved an acceptable rate of entity resolution through a combination of techniques. Our data science work in Python and Pandas is documented in a Jupyter notebook on GitHub [19], which is open to comments, suggestions, and improvements in the form of code contributions.

**Segmenting the Data.** The FAR data contains 151,728 records of individuals exiting camps. The WRA data documents 109,211 individuals arriving into the camp system. If we compare every record in the FAR with every record in the WRA, then that adds up to roughly sixteen billion combinations. Each combination demands several calculations to score similarity, which include first name and year of birth comparisons. After scoring, the scores are compared against each other in order to link the most similar records. All of that adds up to many more computations than we can complete in a reasonable period of time. Therefore, we found it necessary to chop this data up into smaller chunks or segments within which the same individuals are expected to appear.

Our first segmentation approach was to use the family ID number that exists in both datasets. This identifier is a sequence of letters and numbers that sometimes includes hyphens and spaces.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>2999</td>
<td>30-SP-223</td>
</tr>
<tr>
<td>17122</td>
<td>11-030</td>
</tr>
<tr>
<td>30888</td>
<td>H-168</td>
</tr>
<tr>
<td>24289</td>
<td>11-091</td>
</tr>
<tr>
<td>21492</td>
<td>H-130</td>
</tr>
<tr>
<td>25979</td>
<td>30-SP-8</td>
</tr>
<tr>
<td>31967</td>
<td>30-SP-164</td>
</tr>
<tr>
<td>30412</td>
<td>11-125</td>
</tr>
<tr>
<td>3196</td>
<td>30-SP-12</td>
</tr>
</tbody>
</table>

*Fig. 3. Sample family IDs*

Family ID is recorded inconsistently at times and often it is missing entirely. However, the majority of our data does have a useful unique family ID. Grouping our records by the family ID field reduces our problem to computing and comparing scores within each family, usually less than ten individuals that have a particular family ID in the WRA and FAR data. So that means we will have at most 100 combinations to score and compare at one time. After processing each family group, we can then discard the scoring data, using much less memory. An important caveat with this approach is that some individuals may be recorded in different families at different times. Those people will simply not get linked when the data is segmented in this way.

**Segment by Family ID and Camp ID.** Furthermore, when looking at the FAR data, you will sometimes observe the same individual appear twice, having left two different camps at different times due to a camp to camp transfer. It is ideal to have only one possible match to any WRA row, because that means that we can eliminate the WRA row from further matches within the group. Indeed, when only one WRA record and one FAR record are left unlinked in a family group we might establish an entity link despite a very low score, since we have ruled out the other possible matches within that family group.

The FAR camp ID field can help with the problem of individuals appearing twice due to camp transfers. By segmenting the FAR data by both family ID AND camp ID, we can ensure that each individual appears once in the group and will have just one link within the WRA data. Using these groupings to segment the data and our scoring algorithm, described below, we were able to propose a link between almost 80% of the FAR records and a matching WRA record.

**Segment by Phoneme.** Being familiar with the camp data sources and the inconsistency of the family ID field, we thought that we might be able to link additional records by trying another segmentation approach using the last names. Last names are not unique to a single family and are repeated many times in the camp data. However, the multi-family groups that share a last name will suffice to reduce the number of potential matches from all the WRA and FAR rows to only several hundred people. As with the family ID, individuals are still likely to remain in the same last name groups in both the FAR and the WRA data. Since a single last name is often spelled in multiple ways by clerks at the camps, we decided to use a sound index or “soundex” tool to create an index of last names based on the sequence of phonemes. Spellings like “Abbey” and “Abe” have the same soundex value and are considered in the same group. This segmenting strategy means that more individuals are lumped into the same groups, which leads to more possible combinations between the records and more computation. Computation of proposed links when data was segmented this way took approximately twenty times as long as the first computation. There was also significant additional computation used to create the soundex index of family names for both datasets. Using this segmentation approach, we were able to propose links for an additional 7% of the FAR records, leaving only around 13% or twenty thousand out of one hundred and fifty thousand records as yet unlinked to an individual in the WRA.

**Scoring similarity between two records.** Within the groups of families or phonetically unique last names, we used first names, other first names, and birth year to determine if two records were similar. First names as well as any other first names were compared against each other using the FuzzyWuzzy Python module. FuzzyWuzzy uses Levenshtein Distance to calculate the difference between sequences. The Levenshtein distance is a string metric for measuring the difference between two sequences. This means that an exact match is not always needed and helps to account for misspellings and typos. In Fig. 4 and Fig. 5 you can see that “hana” and “hana”, circled in red, refer to the same person, born in 1900. Those names are given a .75 match ratio by FuzzyWuzzy. In the name comparison, we compared all the WRA names with all of the FAR names for two records. The closest match, or the highest similarity ratio between 0 and 1, was taken as the name portion of the total score.
A birth year portion of the total score was calculated as either 0 (no match) or 1 (match) if the birth years in the two records were within 1 year of each other. An exact match was not required as we have seen many examples of inexact birth years in the data. In blue you can see that “Tami” and “Tomee” have similar first names, though these are different people. In that case the birth year will increase the score and yield the correct match. You can also see that in the case of Tomee that the other first name column is needed to establish a name match. Finally, in gold you can see that two children, “Shuzo” and “Tami”, are born in the same year. In this case the first name similarity will further increase the score of the correct match.

<table>
<thead>
<tr>
<th>famid</th>
<th>ln</th>
<th>fn</th>
<th>on</th>
<th>byear</th>
</tr>
</thead>
<tbody>
<tr>
<td>27510</td>
<td>7423</td>
<td>abe</td>
<td>agnes</td>
<td>harumi</td>
</tr>
<tr>
<td>27511</td>
<td>7423</td>
<td>abe</td>
<td>hana</td>
<td>nan</td>
</tr>
<tr>
<td>27512</td>
<td>7423</td>
<td>abe</td>
<td>janie</td>
<td>mikiye</td>
</tr>
<tr>
<td>27513</td>
<td>7423</td>
<td>abe</td>
<td>shuji</td>
<td>nan</td>
</tr>
<tr>
<td>27514</td>
<td>7423</td>
<td>abe</td>
<td>shuzo</td>
<td>nan</td>
</tr>
<tr>
<td>27515</td>
<td>7423</td>
<td>abe</td>
<td>tami</td>
<td>nan</td>
</tr>
<tr>
<td>27516</td>
<td>7423</td>
<td>abe</td>
<td>tomee</td>
<td>tomoye</td>
</tr>
</tbody>
</table>

Fig. 4. FAR records for Abe family 7423

<table>
<thead>
<tr>
<th>famid</th>
<th>ln</th>
<th>fn</th>
<th>on</th>
<th>byear</th>
</tr>
</thead>
<tbody>
<tr>
<td>73</td>
<td>7423</td>
<td>abe</td>
<td>hana</td>
<td>None</td>
</tr>
<tr>
<td>80</td>
<td>7423</td>
<td>abe</td>
<td>harumi</td>
<td>None</td>
</tr>
<tr>
<td>113</td>
<td>7423</td>
<td>abe</td>
<td>janie</td>
<td>None</td>
</tr>
<tr>
<td>258</td>
<td>7423</td>
<td>abe</td>
<td>shuji</td>
<td>None</td>
</tr>
<tr>
<td>260</td>
<td>7423</td>
<td>abe</td>
<td>shuzo</td>
<td>None</td>
</tr>
<tr>
<td>276</td>
<td>7423</td>
<td>abe</td>
<td>tami</td>
<td>None</td>
</tr>
<tr>
<td>289</td>
<td>7423</td>
<td>abe</td>
<td>tomoye</td>
<td>None</td>
</tr>
</tbody>
</table>

Fig. 5. WRA records for Abe family 7423

After these partial scores were calculated, they were combined in a total similarity score. A name match supported by a birth year match would yield a total score equivalent to the name similarity score. A birth year match without a name match would yield a score of only .5, making it a slightly lower score than most of the plausible name matches, which were scored about .75 by FuzzyWuzzy.

A typical scoring grid of the possible links within a family group is shown in Fig. 6 with the vertical axis showing a WRA record index and the horizontal axis being the FAR record index. You can see that some combinations are scored at .5, which indicates a birth year only match. Two siblings were born in the same year, so they have that similarity to each other, but the similarity score is not as high as the 1.0 score that they have with their own records in both datasets. The exact details and code for the scoring algorithm, as well as additional data cleaning steps not described here, can be found in the Jupyter notebook [19]. After the scoring grid was calculated, the highest scores for each FAR record were chosen and if that score exceeded a reasonable threshold, then a link was proposed.

<table>
<thead>
<tr>
<th>row</th>
<th>27510</th>
<th>27511</th>
<th>27512</th>
<th>27513</th>
<th>27514</th>
<th>27515</th>
<th>27516</th>
</tr>
</thead>
<tbody>
<tr>
<td>73</td>
<td>0.0</td>
<td>0.75</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>80</td>
<td>1.0</td>
<td>0.00</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>113</td>
<td>0.0</td>
<td>0.00</td>
<td>1.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>258</td>
<td>0.0</td>
<td>0.00</td>
<td>0.0</td>
<td>1.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>260</td>
<td>0.0</td>
<td>0.00</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>276</td>
<td>0.0</td>
<td>0.00</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.5</td>
<td>1.0</td>
</tr>
<tr>
<td>289</td>
<td>0.0</td>
<td>0.00</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Fig. 6. Grid with similarity scores

Comparison with the curated gold standard. In order to check on the completeness and accuracy of this entity resolution method, we wanted to compare our results with those obtained by the researcher on our team most familiar with these datasets, Dr. Marciano. Dr. Marciano used a subset of 102 names of male internees who were transferred from Tule Lake to North Dakota on 7/03/45, following an incident that we might describe as a protest, but that authorities in the camp termed a “riot”. Out of a total of 204 FAR records, Dr. Marciano was able to find 199 matching records in the WRA dataset. The entity resolution approach was able to match 168 of these same FAR records with the WRA records. The difference between the manual method and the automated linking is about 15%. This difference may be attributed to several causes, including some truncated last names in the WRA records due to limitations in the original punched card data format.

The next step for these proposed links is to publish a graph dataset with unique identifiers for all the individuals and the migration events that they experienced. This graph will be made available as JSON linked data and as a Neo4j graph database file. It can become the basis for further work on data infrastructure to support community access, curation, and contribution as outlined below.

V. INTEGRATING INCARCERATION ARCHIVES FOR RESEARCH

In contrast to European humanities projects, the US has not managed to successfully sustain larger projects of integrated infrastructures. Examples of European infrastructures include: DARIAH ([http://www.dariah.eu](http://www.dariah.eu)) and CLARIN.
VI. DEPORTATION TO THE DOJ CAMP OF BISMARCK

In order to illustrate what this kind of integration might look like and the potential discoveries that would be enabled, we explored the data surrounding a historical event, the Bismarck deportation. This test event from the incident cards is the deportation of 100 men from the Tule Lake WRA concentration camp to the Fort Lincoln DOJ Camp in Bismarck, North Dakota on July 3, 1945. One hundred individual Incident Cards capture this event at the person level, and the associated police report features a letter listing all one hundred individuals. We create individual “Person” nodes and connect them to a “Fort Lincoln, Bismarck, ND” DOJ Camp location node through a “xfer-To” link. Through this process we are able to create knowledge networks from individual events, linking them into larger interconnected graphs. In our initial ontology, we create the following categories:

- **Person** -- with data from WRA Form 26, and FAR
- **City** -- which includes Pre-evacuation, Birth, and Port cities: with data from WRA Form 26, FAR, and the “Sailing Lists of Repatriations and Persons Returning to Hawaii and Japan”.
- **Camp** -- which include the DOJ and WRA camps.

We then automatically upload the compiled information into the graph database. See Fig. 7 which already shows 194 nodes and 700 relationships.

Such a knowledge base allows us to then query the graph to extract individual camp biographies. We call these “computational biographies”. Fig. 8 illustrates how a query on “Tatsuo Ajisaka” reveals major life and incarceration events including: birthplace (light green node: Kyushu, Japan), pre-evacuation city (dark green: Venice, California), Temporary Assembly Center of entry (orange node: Manzanar), 1st WRA Camp (purple node: Manzanar), 2nd and final WRA Camp (purple node: Tule Lake), DOJ Camp of deportation (light blue node: Fort Lincoln, Bismarck ND), and finally Port City from which he was deported back to Japan (light brown node: Portland, OR).
Having constructed the core camp life stories of the 100 Japanese Americans, we then further mine the Incident Cards to provide more details on their trajectories within the camps. We illustrate this with two Incident Cards for “Itaru INA,” one of which is shown in Fig. 9.

We model these as a new type of node:

- **Event** -- which connects to Person nodes through an “in_Event” link

This demonstrates how increasingly detailed knowledge graphs can be constructed as archival fragments are discovered across collections and incrementally connected to the graph database.

Thus, Itaru Ina’s computational biography would now include two dark blue nodes indicating his connection to two significant camp events.

Finally, we demonstrate one additional computational biography graph where the person is not connected to an Assembly Center or a first WRA Camp, as he is directly transferred to the Tule Lake WRA concentration camp.

VII. NETWORKED REPRESENTATIONS

The disambiguation of named entities and a unified registry of individuals allows us to use computational methods to share biographical stories and bring new clarity to events. Keeping computational prosopography in mind, we examine relationships between internees and events at Tule Lake using social network analysis. Social network analysis studies “the behavior of the individual at the micro level, the pattern of relationships (network structure) at the macro level, and the interactions between the two. Social networks are both the cause of and result of individual behavior” [16]. For this case, social network analysis involved extracting data from WRA records and applying graph algorithms found within networking tools. NetworkX, an open source Python package used to create and study dynamic structures and functions of complex networks [17], was selected for its ability to integrate graphing and modeling functions directly in Jupyter notebooks. Another open source Python graphing library, Plotly, was explored to analyze the same data and verify findings. The approach to leverage both the NetworkX framework and Plotly...
visualizations in Jupyter notebooks, provides a model for students and researchers to apply to their own archival science research.

Fig. 13. Social network model of reported incidents at the Tule Lake concentration camp between Oct. 30th - Nov. 9th, 1943.

Social network models were created from the Incident Card dataset transcribed from the National Archives Index Cards. Using the Pandas software library in Python, a dataframe was created to prepare the data for computational analysis. Based on specified conditions within the dataframe, visualizations were generated using the NetworkX function .from_pandas_edgelist to create nodes and edges, and the .draw function to generate the social network model. In Fig. 13, the large peach-colored nodes of the network model represent dates and the small blue-colored nodes represent reported incidents. This figure focuses on November 4th, 1943, when a large congregation of incarcerated Japanese Americans at Tule Lake was deemed a “riot” by government officials. On that day, as groups began to gather, the project director of the incarceration camp called in the military police and imprisoned in ‘bullpen’ area of the hastily assembled stockade” [18].

Fig. 13 provides a closer look into the activities surrounding the “riot” through the extraction of Incident Card data in the dataframe based on the “riot” date, as well as data from five days before and five days following the event. Investigating a date node on November 1st reveals reported incidents of “theft.” Returning to the Python dataframe and setting conditions for incidents of “theft” occurring on November 1st, our results show nine incarcerated Japanese Americans lived or were treated and can lead to further exploration of reported incidents in the Incident Card data. Another date node for November 4th shows reports of incidents designated “other cases breach of trust.” Results from the Python dataframe based on these conditions display data for 15 cards for individuals who failed to return tools taken from the construction department at some point before November 4th. The number of cards from these incidents is also verified in Fig. 14. A similar dataframe that includes the number of reported incidents per day was applied using the Plotly .bar function to create a bar chart. Setting the same conditions for five days before and after the “riot” event, the figure shows the total number of incidents that were reported from October 30, 1943 to November 9, 1943, not including the “riot” event itself. The reported incidents are charted on the x-axis and the count, or number of cards, is charted on the y-axis. The date for each incident is distinguished by different colors which are identified in the legend or by hovering over each bar in the Jupyter notebook [19]. In Fig. 14, we can confirm there were 9 incidents of “theft” on November 1, 1943 and 15 incidents of “other cases breach of trust” on November 4, 1943.

Through examination of these results, one can conclude that the “thefts” and “other cases breach of trust” were clear displays of the rising discontent in the camp regarding access to food and supplies. Further research into these acts of resistance might shed light on how and why the “riot” event unfolded at Tule Lake. It also leads to questions regarding the informal gathering of the accused incarcerated Japanese Americans and additional or other intentions they may have had beyond what was reported in the cards. By examining the Incident Card dataset using Python and Plotly, and incorporating NetworkX functions to create social network models, we can easily view different interactions between individuals over time and how they relate to various events (as reported by government officials), ultimately presenting additional opportunities for analysis and research.

VIII. Spatial Representations

Spatial analysis is another unique approach to visually understanding the movement or spaces of how Japanese Americans lived or were treated and can lead to further exploration of reported incidents in the Incident Card data. Through spatial analysis, we can answer the where questions and model new information and relationships. For this case study, we wanted to know where Japanese Americans, those considered influential figures in the reprisal of January 4, 1944, lived before arriving at Tule Lake, including where and the number of times they were relocated.

For the application and analysis of geospatial data of the Incident Cards, we chose to use Folium. Folium is an open-source Python library that allows for the visualization and manipulation of spatial data and supports Image, Video,
GeoJSON, and TopoJSON overlays (Folium). The Incident Card data processing was completed using Pandas in a Jupyter notebook. The data was then filtered based on two criteria: offense (riot) and case number (A-7), using the ‘pandas.series.str.contains’ function (e.g., df.series.str.contains('riot')). We found that there were inconsistencies in how the case number was recorded because incidents were registered on the cards by different individuals. To get around this challenge, we applied regular expressions within the contains function to parse out and capture the many variations of case number 'A-7'. After filtering, the offenses were sorted using ‘value_counts()’ and selected based on the frequency (total number of times a name appeared in the cards) and if they could be found in the WRA and FAR datasets. As a result, 25 names were selected and used to search through the WRA, FAR, and Incident Card datasets to discover five central locations (the point of origin, the name of the assembly center, the first camp, the final departure state, and barrack address in Tule Lake).

In the Jupyter notebook [19], map visualizations were created using ‘folium.Map’ and assigning longitude and latitude values, and setting zoom parameters. By applying folium plugins, we generated circle markers to display the number of individuals relocated to their respective first camp locations (see Fig. 15). In Fig. 15, the size of the circle corresponds to the number of individuals relocated to that location.

Fig. 15. Out of the 25 selected names a large number were relocated to Jerome before relocating to Tule Lake in 1943.

Another way to visualize multiple movements is by creating layers. By having layers for each location, we can control which layer is displayed at one time to observe patterns in the migration of this particular group of Japanese Americans (see Fig. 16). For example, through interactive mapping, we can follow the movements of Japanese Americans who resided in Hawaii and observe they were directly transferred from their place of origin to Sand Island and shortly after to Jerome, which is a much different pattern than those who reside in the U.S.

We can also utilize aspects of Folium to add images to markers to develop a story or give additional background on a given location, as we did in Fig. 17. The idea is to use spatial data in a way that links space and time to specific individuals and events to discover how and where groups connected, where events occurred, and ultimately to learn how Japanese Americans were treated.

Fig. 16. Movement of 25 Japanese Americans from place of origin to their final relocation between 1942 - 1945.

Fig. 17. A photograph from the National Park Service of Tule Lake in 1943.

**IX. CONCLUSIONS AND FUTURE WORK**

In summary, our investigations show the foundational importance of building a unified person registry, through the demonstrated entity resolution techniques, evaluated against a best effort manual gold standard, that supports the prosopographical approach, namely the enhancement of historical context by creating meaningful links to biographical data across multiple datasets. The use of social network analysis to graph dates and incidents around the November 4, 1943 “riot” event offers an alternate view into the life and treatment of incarcerated Japanese Americans. By leveraging NetworkX and Plotly functions with Python to create network models and diagrams, we delve deeper into recorded events and provide opportunities to observe the data in new light. Furthermore, we have shown the value of using spatial data as an additional lens for the visualization of data to give more context to obscure topics. By bringing spatial analysis into this project, we have observed the movement (five relocations) of a small group of Japanese Americans. We can identify where they lived while at
Tule Lake, including significant locations where notable events have taken place based on data from the Incident Cards.

Future work should focus on a published unified incarceree registry that uniquely identifies the individuals and families that are described in WRA Form 26 and FAR records, including with it the methods used for entity resolution as captured in a computational notebook. Additional scholarship should further explore Python graphing libraries to produce intricate social network models and create a more comprehensive representation of the movement of individuals and groups (e.g., Dr. Marciano’s dataset of 100 Japanese Americans deported to Bismarck, North Dakota, July 3, 1945).

X. ACKNOWLEDGMENTS

We wish to acknowledge the support of Geoff Froh at Densho.org. We also wish to acknowledge the support of two interconnected Institute of Laura Bush 21st Century Librarian (LB21), National Digital Infrastructures and Initiatives Museum and Library Services (IMLS) projects: LIS Education and Data Science Integrated Network Group (LEADING) project (RE-246450-OLS-20) and Piloting an Online Collaborative Network for Integrating CT into Library and Archival Education and Practice (RE-246334-OLS-20). These projects are advancing data science education and infrastructure across the library, archives, and museum ecosystem. Together, these initiatives are providing resources for faculty teaching across these communities, and preparing the next generation of information professionals and enabling early and mid-career professionals to develop skills through immersive learning experiences.

REFERENCES


