Computational Curation of a Digitized Record Series of WWII Japanese-American Internment

William Underwood
Digital Curation Innovation Center
College of Information Studies
U. Maryland, College Park
underwood@umd.edu

Richard Marciano
Digital Curation Innovation Center
College of Information Studies
U. Maryland, College Park
marciano@umd.edu

Sandra Laib
Independent Researcher
GA Tech. Research Inst. (retired)
Atlanta, GA US
sandra.laib@gmail.com

U. Maryland College of Information Studies student team:

Undergraduate (Info. Sci.):
Carl Apgar
Luis Beteta
Waleed Falak
Marisa Gilman
Riss Hardcastle
Keona Holden
Yun Huang

Graduate (MLIS):
David Baasch
Brittni Ballard
Tricia Glaser
Adam Gray
Leigh Plummer

Graduate (MIM):
Zeynep Diker
Mayanka Jha
Aakanksha Singh
Namrata Walanj

***Abstract*—This paper describes the linguistic analysis of index note cards from record series of the World War II Japanese-American Internment Camps that are in the custody of the National Archives. It also describes the use of GATE Developer, and an extension of ANNIE, a GATE plugin, in linguistic processing of information specific to index note cards in order to extract metadata supporting access and archival decisions regarding record release and withdrawal. The content of the index cards will be interpreted as OWL/RDF statements. Those statements will be stored in a graph database and used with objects such as digital maps and photos to produce an interactive user interface to exhibit events at relocation centers.**

Keywords: NLP, NER, World War II Japanese-American Internment Camps, Computational Archival Science

I. Introduction

Two months after the attack on Pearl Harbor, President Franklin D. Roosevelt issued Executive Order 9066 that allowed the Secretary of War to designate military areas and order evacuation of all persons in those areas deemed to be a threat to national security. [1] Though the text of EO 9066 does not contain the word "Japanese," the intent was to create a program to remove 120,000 Americans of Japanese ancestry from their homes in coastal California, Oregon, and Washington State in the name of national security. Those affected were forced to leave their communities as the Federal government moved them to heavily guarded camps in isolated areas hundreds of miles away. The 10 major camps, euphemistically called "Relocation Centers" were located in California (Manzanar and Tule Lake), Arizona (Poston and Gila), Idaho (Minidoka), Wyoming (Heart Mountain), Colorado (Granada), Utah (Topaz), and Arkansas (Rohwer and Jerome).

The National Archives and Records Administration (NARA) is the repository of the records of this program. Record Group 210, Records of the War Relocation Authority, includes paper records of internal security cases and associated paper index cards for the 10 Relocation Centers. These records have not been released to the public due to access restrictions on some of the records.

There are four objectives of the research described in this paper.

1. Use computational linguistic analysis methods to extract item-level metadata from the Relocation Center index cards in order to supply archivists at the National Archives and Records Administration (NARA) with the information needed to determine access restrictions for items in this record group.
2. Curate the information in these cards by improving their quality (scanning, OCR, text correction, analysis, and extraction) and adding value to the repository of digital information by providing to NARA descriptive metadata and support of withdrawal decisions.
3. Explore archival analytics approaches through geo-processing and social networking analysis and geospatial processing of the database.
4. Provide an opportunity for iSchool Students to gain knowledge and experience in using methods of computational linguistics in analysis of archival textual record series through teamwork in the Digital Curation Innovation Center (DCIC).

We expect our research to contribute valuable insights to the emerging field of Computational Archival Science (CAS) [2]. CAS as described in this seminal paper suggests

the development of a trans-disciplinary field that blends computational and archival thinking. It is motivated by the laying out of eight thematic case studies covering: (1) Evolutionary prototyping and computational linguistics, (2) Graph analytics, digital humanities and archival representation, (3) Computational finding aids, (4) Digital curation, (5) Public engagement with (archival) content, (6) Authenticity, (7) Confluences between archival theory and computational methods - cyberinfrastructure and the Records Continuum, and (8) Spatial and temporal analytics. Our research clearly intersects with six of these themes: (1), (2), (3), (4), (5), and (8)

II. Collection Analysis

A. Analysis of Tule Lake Index Cards

There are over 25,000 paper index cards from all 10 Relocation Centers. The cards from the Tule Lake center comprise the bulk of the collection with close to 65% of the total, and to date the index cards associated with this center have been the focus of our research. The faculty and students working on this project have so far analyzed close to 500 of the index cards from the Tule Lake center. Figure 1 shows an example of one of these cards. [1]

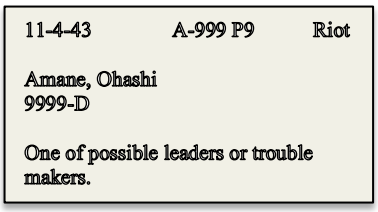


Figure 1. Sample index card to Case Reports.

The types of information on this card that are relevant to our research are the date, the case report ID (A-999), the relevant page number in the report (P9), the subject of the case report (offenses such as a Riot), the Japanese-American or Japanese internee name, the residence ID in the camp (9999-D) and a remarks section. These are the most frequent concepts on the index cards. However, other concepts occurring in the index cards include European-American names of administrative and staff members of the center, document identifiers (C-816), document types (letter, memo), names of relocation center facilities (hospital, stockade), subjects of other cases that are not offenses (Accidents and Transfers), job titles (chief cook, #35 Block Manager, secretary), organizations such as young Japanese men's militant groups (Hokoku Seinen Dan), relocation center organizations (Administrative Police, Agricultural Division), US Government organizations (U.S. Army, War Relocation Authority), locations (Honolulu, Hawaii, San Francisco, Calif.), time and time periods (2:45 pm, 36 days), Family numbers (F#8504), and Individual numbers (8504-A). The Remarks section of the index cards, also include relations such as membership in militant organizations and actions/events (arrested, interviewed by, sentenced to time period, and released).

[1] The names, case ids and residence ids on this sample card are fictitious, as

B. Styles of Index Cards

Index cards were found to have different styles. This is most likely due to the fact that they were indexing a record series consisting of files created and used in different offices and at different internment camps. The styles (or document types) of index cards need to be identified and defined in order to correctly interpret their content. Initial analysis of about 500 index cards indicates at least five major styles of index cards, with three sub-styles:

1. Index card to Case Reports
    a. With Other IDs
    b. With Multiple Names
    c. With Other Cases
2. Index card to Case Report using Keywords
3. Index card to Registration Arrests
4. Index card to Documents
5. Continuation of Remarks – Remarks on back of Index Card

The sequential structure of information on the card aids in identifying the style of the index card as well as the interpretation of its contents. For instance, the pattern for identifying an *Index Card to Case Reports* (example in Figure 1) is:

Date, Case Report Id, Subject is an Offense,
Japanese (-American) person name,
(optional Residence Identifier), Remarks

III. Digital Curation of the Collection

A. Digitization of the Index Cards

NARA has scanned the index cards producing JPEG images of the cards. The DCIC Center is using ABBYY FineReader [3] to create UTF-8 encoded text from the JPEG images. There are typos and misspellings on the original paper index cards as well as OCR errors that are being manually corrected by the project team, thus increasing the quality and utility of the record copies. Our OCR strategy utilizes the previous classification of the cards into families of styles, to create recognition layout templates that produce more structured text output.

B. Annotation and Information Extraction

GATE is the acronym for General Architecture for Text Engineering, a collection of computational methods for linguistic analysis of digital documents [4]. ANNIE is a serial controller for conditional processing resources (PRs) run over corpora. Processing resources are primarily automated methods of computational linguistics. ANNIE was originally developed to annotate expressions in newswires. ANNIE needs to be refined to correctly annotate expressions in other kinds of digital documents such as the index cards to security cases.

The out-of-the-box functionality of GATE provides a baseline level on top of which individual processing resources can be modified to suit the corpus being processed as well as the pipeline itself. For most purposes, the inclusion of a custom ANNIE Gazetteer and NE Transducer that applies to a specific collection of records would be sufficient to extract meaningful information from them. In the case of records related to Tule Lake Index Cards, we supplied a custom gazetteer ("Tule Gazetteer") and NE transducer ("Tule NE Transducer").

Figure 2 displays the so-called pipeline of GATE processing resources that ANNIE uses to annotate text in a GATE corpus.

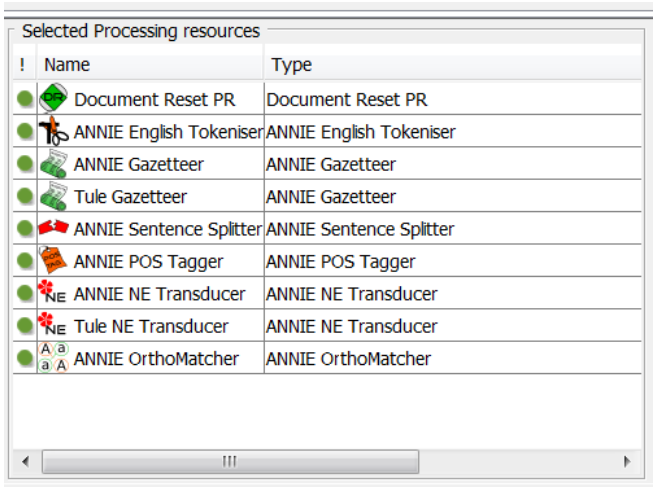


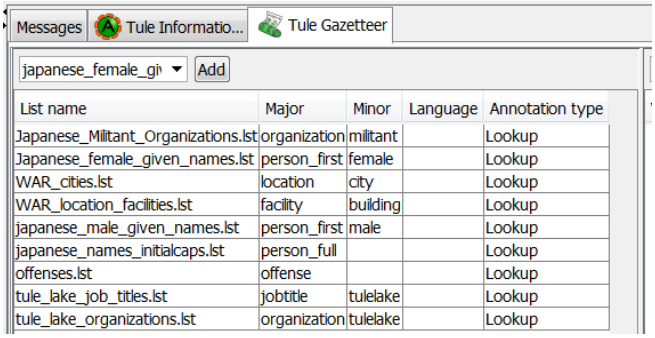
Figure 2. ANNIE Pipeline with additional PRs for Tule Lake corpora.

The 1st processing resource, Document Reset PR, sets the internal table recording of annotations to empty. As one identifies errors in the annotations one can update the JAPE rules ("Java Annotation Pattern Engine") and word lists and rerun the Pipeline against the current corpus.

The 2nd PR, ANNIE English Tokenizer, identifies the tokens in the text. The term "token" refers to the total number of words in the text, regardless of how often they are repeated. The term "type" refers to the number of distinct words in the text. Thus, the sentence "I do what I want to do" contains seven tokens, but only five types, as the tokens "I" and "do" are repeated.

The 3rd PR, ANNIE Gazetteer, is a wordlist lookup that matches the tokens in a document against terms in wordlists of such classes as person first names, surnames, city names, country names, and organization names. If there is a match, the text is annotated with a tag for the name of that class.

The 4th PR, Tule Gazetteer, is the same PR as the preceding but with wordlists created from the index cards. Figure 3 shows the names of these wordlists and the tags used to annotate the text.



List name	Major	Minor	Language	Annotation type
Japanese_Militant_Organizations.lst	organization	militant		Lookup
Japanese_female_given_names.lst	person_first	female		Lookup
WAR_cities.lst	location	city		Lookup
WAR_location_facilities.lst	facility	building		Lookup
japanese_male_given_names.lst	person_first	male		Lookup
japanese_names_initialcaps.lst	person_full			Lookup
offenses.lst	offense			Lookup
tule_lake_job_titles.lst	jobtitle	tulelake		Lookup
tule_lake_organizations.lst	organization	tulelake		Lookup

Figure 3. Wordlists specific to the index cards.

The 5th PR, ANNIE Sentence Splitter, splits the text into sentences by identifying tokens followed by punctuation such as periods and question marks.

The 6th PR, ANNIE POS Tagger, identifies the part-of-speech of the tokens by looking up the word type of the token in a dictionary to find its part-of-speech. If the token is not found in the lexicon, the rule-based "Hepple POS Tagger" is used to identify the probable part of speech.

The 7th PR, ANNIE NE Transducer, is a GATE PR called the "Java Annotation Pattern Engine" or JAPE. It applies JAPE rules to the annotated text to produce additional annotations. For instance, if a person's first name is followed by a proper noun, it concludes that the combination is a person's full name. It also produces such annotations as full location names made up of city and state or country names. NE is an abbreviation for Named Entity.

The 8th PR, Tule NE Transducer, is also the ANNIE Transducer, but with an additional set of rules to address concepts specific to the index cards. For instance, there are rules to identify and annotate case ids, page numbers, residence ids, and person names in which the surname appears before the given name. There are also rules to identify different styles of the index cards.

The 9th PR, ANNIE Orthomatcher, performs linguistic functions known as nominal co-reference. For instance, if an index card mentions "Tom Yoshio Kobayashi" and "Kobayashi", this method concludes that these person annotations refer to the same person and links the annotations.

C. Running the ANNIE Pipeline on a Digital Corpus

The OCRed text-corrected index cards for a Relocation Center can be loaded into GATE as a GATE Corpus. The ANNIE Pipeline we are constructing can then be applied to the GATE Corpus. GATE Developer provides a user interface for displaying the annotations created by the pipeline as color-coded highlighted Text. Figure 4 shows the color-coded annotations for the sample index card shown in Figure 1. It also shows the internal table in which the annotations are actually recorded.

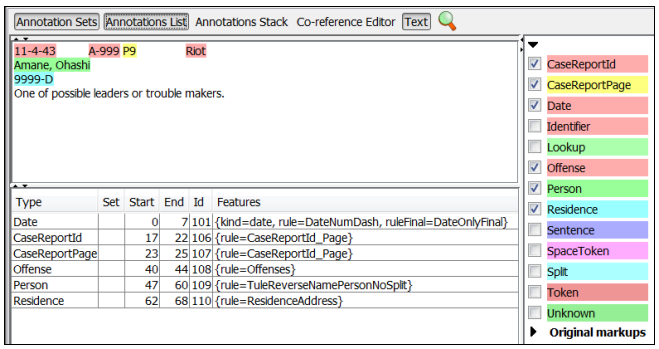


Figure 4. Color-coded Annotations and Annotation List for sample index card Shown in Figure 1.

D. Interpretation of the Index Cards

An OWL/RDF ontology will be developed to define the name meaning of the terms in the note cards [6]. A set of rules can be developed to interpret the annotations in terms of the ontology. A program will be written to export the information in the annotation table as a list of logical predicates or RDF statements. For example, the quasi-logical form (*qlf*) of the annotations of the Index card shown in Figure 4 is:

```
(qlf =[
  filename( "Box8-Tule_Lake-999.jpg" ),
  style( "Box8-Tule_Lake-999.jpg",
         "index_card_to_case_reports" ),
  case_no( "Box8-Tule_Lake-999.jpg","A-999" ),
  page_no( "A-999", "9" ),
  date( "A-999", "11-4-43" ),
  subject( "A-999", "Riot" ),
  person_name( "Amane, Ohashi" ),
  gender( "Amane, Ohashi", "male" )
  indexname( "Amane, Ohashi",
             "Box8-Tule_Lake-999.jpg")
  residence_no( "Amane, Ohashi", "999-D" )
])
```

These predicates express the meaning (semantics) of the terms annotated in the index card. The predicates in the *qlf* can be used to populate the graph database framework we describe in section IV.

E. Catalog Description and Support of Withdrawal Decisions

The National Archives has requested that we use metadata extracted from the index cards to construct the following descriptive metadata for each item.[2]

Box #	Item Title	Image Filename
1	Relocation Center: Last Name, First Name	SeriesNAID-BoxNumber-ImageNumber
2	Central Utah (Abraham): Adachi, George	1548726-01-0001.jpg

The National Archives has also requested that we identify index cards with information about internees under the age of 18 years old, which should not be released to the public. To support these decisions, we plan to utilize a NARA record series entitled *"Japanese-American Internee Data File], 1942 – 1946"*, which is a data file consisting of the following information.

> *"This series contains personal descriptive data about Japanese Americans evacuated from the states of Washington, Oregon, and California to ten relocation centers operated by the War Relocation Authority during World War II in the states of California (Tule Lake and Manzanar Centers), Idaho (Minidoka Center), Utah (Central Utah Center), Colorado (Granada Center), Arizona (Colorado River and Gila River Centers), Wyoming (Heart Mountain Center), and Arkansas (Rohwer and Jerome Centers). Each record represents an individual and includes the name; relocation project and assembly center to which assigned; previous address; birthplace of parents; occupation of father; education; foreign residence; indication of military service, public assistance, pensions, and physical defects; sex and marital status; race of evacuee and spouse; year of birth; age; birthplace; indication of the holding of an alien registration number and/or Social Security number, and whether the evacuee attended Japanese language school; highest grade completed; language proficiency; occupations; and religion." [5]*

It turns out that this data file is not "clean" enough to serve as the sole reference source and we are cross-referencing it with a second dataset, the FAR or Final Accountability Rosters of Evacuees at Relocation Centers. A relational database will be created from the combination of these two datasets, and generally speaking, information extracted from the index note cards such as person name, age, family id, individual id, and gender will be used to identify items that can be released to the public or that should be withdrawn.

IV. LINKING, VISUALIZING, AND ANALYZING EXTRACTED INFORMATION

In addition, we are interested in discovering untold stories hidden in the forest of index cards, by linked them together and visualizing them through social networking and interactive mapping techniques (GIS).

One of the student groups is exploring the use of the Neo4j platform, a graph database tool, which stores data and its relationships together physically. The nodes in the graph can be people, organizations, or events (essentially the entities we extracted in GATE and stored in the database). The edges can represent family connections or co-appearance of people on an incident card. Both nodes and relationships can be further tagged with attribute-value pairs. After "stitching" nodes together with a number of computed relationships, a social network can be built. This

[2] The name and case number are fictitious

will allow for a variety of analyses including: clustering, finding the shortest path between two nodes, calculating various measures of centrality and closeness, and recognizing hidden relationships in the network. The results of this type of network analysis may have strong social impacts, and when we are ready, we hope to engage with experts and survivors who can help guide the process in a meaningful and ethical way, taking into account the underlying sensitivities, and navigating through the inherent collection biases and propaganda. Figure 5 links together five types of nodes: *blue* (Internee names), *red* (Relocation Camp locations), *yellow* (Internee occupations), *purple* (common Family Number), and *green* (Assembly Center locations).

Figure 5. Graph visualization of internees who shared Family Number "14911".

Another student group has recreated an interactive map of the Tule Lake camp, down to individual buildings and outhouses. We also hope to use this approach to link archival fragments and extracted entities together, to visualize and recreate the clustering of spatial events and movement of people within the camp.

Figure 6. Interactive map of Tule Lake including individual buildings, barracks, and structures.

V. SUMMARY

It has been described how scanning, OCR and error correction can be used to improve the quality of paper records. It has also been shown how linguistic analysis and computational methods can be used to add value to the records by extracting metadata from the index cards to be used to create the descriptive metadata needed by NARA. It has also been shown how that metadata can be used to sort cards into those which can be released to the public and those which should be withdrawn. In addition, we have started archival analytics activities that involve social networking analysis and geospatial processing. Finally, it is through the involvement of students in such projects that they have the opportunity to gain knowledge and experience in using computational methods for curation of digital textual records.

ACKNOWLEDGMENTS

The Office of Innovation at the National Archives and Records Administration (NARA) has, prior to their release to the public, provided us access to the archived index cards to the internal security cases of the Japanese-American Relocation Centers. We also wish to acknowledge ongoing support from Maryland's iSchool to the Digital Curation Innovation Center (DCIC) and also major funding from the National Science Foundation's "Brown Dog" project (NSF Cooperative Agreement ACI-1261582), a $10.5M NSF/DIBBs-funded collaboration with the University of Illinois at Urbana-Champaign NCSA Supercomputing Center. The work in this paper has leveraged the Brown Dog funded dataCave and DRAS-TIC open-source platform (Digital Repository At Scale - That Invites Computation). More details at: http://dcic.umd.edu/about-us/infrastructure/.

REFERENCES

[1] Executive Order 9066 dated February 19, 1942. https://www.ourdocuments.gov/print_friendly.php?flash=true&page=transcript&doc=74&title=Transcript+of+Executive+Order+9066:++Resulting+in+the+Relocation+of+Japanese+(1942)

[2] Archival Records and Training in the Age of Big Data, Marciano, Lemieux, Hedges, Esteva, Underwood, Kurtz, Conrad, accepted for publication in 2018 in "Advances in Librarianship – Re-Envisioning the MLIS: Perspectives on the Future of Library and Information Science Education", Editors: Lindsay C. Sarin, Johnna Percell, Paul T. Jaeger, & John Carlo Bertot. See: http://dcicblog.umd.edu/cas/wp-content/uploads/sites/13/2016/05/Marciano_Kurtz_et-al-Archival-Records-and-Training-in-the-Age-of-Big-Data-final-1.pdf

[3] ABBYY FineReader. https://en.wikipedia.org/wiki/ABBYY_FineReader

[4] Cunningham, et al. *Developing Language Processing Components with GATE Version 8*. University of Sheffield Department of Computer Science. 17 November 2014.

[5] Japanese-American Internee Data File, 1942-1946 [Archival Database]; Records of the War Relocation Authority, Record Group 210; National Archives at College Park, College Park, MD. Also accessible via the Access to Archival Databases (AAD) at www.archives.gov/aad

[6] Protégé. Stanford University. http://protege.stanford.edu