Digital Curation of a World War II Japanese-American Incarceration Camp Collection: Implications for Sociotechnical Archival Systems

Richard Marciano[1], Myeong Lee[1], William Underwood[1], Sandra Laib[2], Zeynep Diker[1], and Aakanksha Singh[1]
[1]College of Information Studies, University of Maryland, College Park, United States
[2]Independent Researcher, Georgia Tech Research Institute (retired), Atlanta, United States
{marciano, myeong, underwod}@umd.edu, {sandra.laib, zdiker, aakanksha_singh93}@gmail.com

***Abstract*—We describe computational treatments of archival collections through a case study of World War II Japanese-American Incarceration Camps. Camp staff and police officers compiled so-called "internal security" reports relating to alleged cases of "disorderly conduct, assault, theft, loss of property, and accidents" in the camps, and an index to these reports comprising over 25,000 index cards to the reports. The sheer size of these collections is pushing archivists and researchers to consider new forms of processing for collections at scale. We discuss novel digital treatments that involve (1) processing incident cards at scale in an automated way, (2) preserving and documenting the record series and their operations using an Archival Information Package (AIP), and (3) developing social network-based archiving, analysis, and visualization of the collection. Since this type of project involves multiple partnerships, agreements, and social interactions with community groups and public-sector organizations, we discuss the possibility of the concept of *sociotechnical archival systems* as an implication of our approach.**

***Index Terms*—NLP, NER, digital curation, computational archival science (CAS)**

I. Introduction

An open problem in the archival community is to catalog large-scale archival collections as many of them are still uncatalogued [1]. Index card collections represent a common type of big archival collections [2]. For instance, the International Tracing Service (ITS) in Germany houses nearly 50 million central name index cards, documenting forced labor and the Holocaust in Nazi Germany and its occupied regions [3]. Also, the US National Archives in St. Louis alone houses many such series with potentially hundreds of millions of item-level records, which include World War I Award cards, VA cards, American Expeditionary Forces cards, Pay cards from the US Army, Philippine POW cards, World War II Individual Civilian Internment cards, Selective Service Registration cards, World War II Army Nurse cards, Berlin Airlift Device cards, Army Air Force Award cards, Chaplain File cards, Merchant Marine Officers Licenses cards, Civilian Service Record cards, Arlington National Cemetery cards, and many others.

These collections are of interest to archival scientists and practitioners not only due to the large amount of uncatalogued collections but also because of users' limited access to them. In traditional archival practices, at best, a large collection is catalogued in a way that provides its index as an access point for potential users and researchers. This is the approach carried out across historical card series that provide valuable information for researchers and genealogists who look for people's names, addresses, and other insights.

However, simply digitizing these kinds of cards with one or two access points such as organizing cards based on names limits users' capability to fully retrieve, use, and navigate archived information. Our goal, in addition to efficiently cataloguing such a big collection of cards, is to help unlock these types of records by enhancing and diversifying access points through *digital curation workflows* that combine incremental modeling, automated information extraction, entity-level markup, data analysis, and visualization. Ultimately, we aim to support archival processing, archival decision-making, and review analytic practices by providing new tools for professional archivists, which will also lead to better and faster outcomes for the public.

This is part of an emerging approach that we call Computational Archival Science (CAS) [4]. We illustrate this approach by curating, processing, analyzing, and visualizing the World War II Japanese-American Incarceration Camp card series (RG 210 Records of the War Relocation Authority at the National Archives and Records Administration in Washington D.C.). This collection has not yet been released to the public and includes paper records of internal security cases and associated paper index cards for ten incarceration centers in which Japanese-Americans were imprisoned during World War II.

We report details of the progress toward developing a tool for extracting descriptive metadata from OCR-ed images of the index cards that can be used to provide researchers diverse access points to these cards. Steps include: (1) cleanup of typos and OCR errors in the scanned images of the index cards, (2) development and performance evaluation of an Index Term Identifier to automatically identify index terms in the index cards and extract descriptive metadata needed to provide access to the index cards, (3) identification of additional resources needed to improve the performance of the Index Term Identifier, and (4) a refined format for descriptive metadata in light of the kinds of index terms discovered on the index cards. We then discuss the need to create an Archival Information Package (AIP) that addresses the preservation of the record series, the documentation of archival operations on

978-1-7281-0292-4/18/$31.00

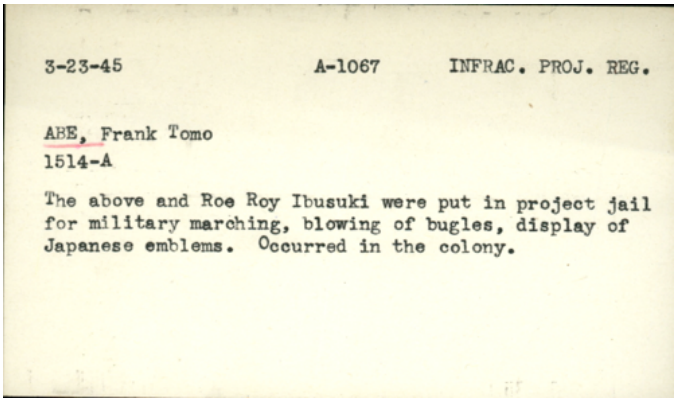

Fig. 1. An example of incident cards for Frank T. Abe in the Tule Lake camp.

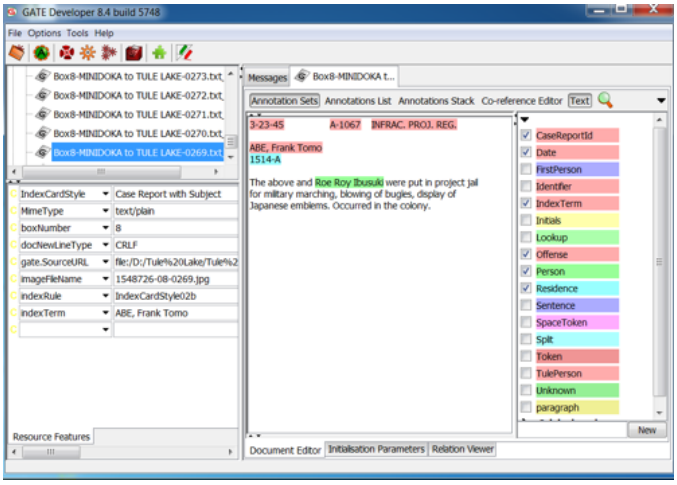

Fig. 2. An example of multiple annotations for Tule Lake Index Card 269.

the series, and the resources used in processing the series.

Finally, we discuss theoretical implications for the possibility of a new concept of *sociotechnical archival systems* since developing and sustaining these kinds of archival projects not only require advanced-level information technologies and infrastructures but also involve complex partnerships, agreements, alliances, and social interactions with community groups and non-profits, hence, constitute a larger system of actors, networks, artifacts, and their institutions [5]. We hope this conceptualization can be developed in a holistic way and one that provides vocabularies and assessment frameworks for future archival systems.

II. World War II Japanese-American Incarceration Camp Cards

We published an earlier study in 2017 on the linguistic analysis of index cards from record series of the World War II Japanese-American Incarceration Camps that are in the custody of the National Archives [6]. This earlier paper described the use of the open source GATE Developer software, and an extension of ANNIE, a GATE plugin, in processing the linguistic information of incident index cards. The extracted metadata from this processing supports user access to and archival decisions on records. The content of the index cards is planned to be interpreted as OWL/RDF statements, and these statements will be stored in a graph database and combined with other digital artifacts such as digital maps and photos in producing interactive user interfaces that exhibit events at Incarceration Centers.

There are over 25,000 paper index cards from all the ten Incarceration 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 from this center have been our research focus (e.g., Fig. 1). The text extraction GATE workflow that we are developing goes beyond automatically indexing the principal person names on the cards; this workflow also extracts entities in numerous categories such as Case Report IDs, Dates, Housing Identifiers, Types

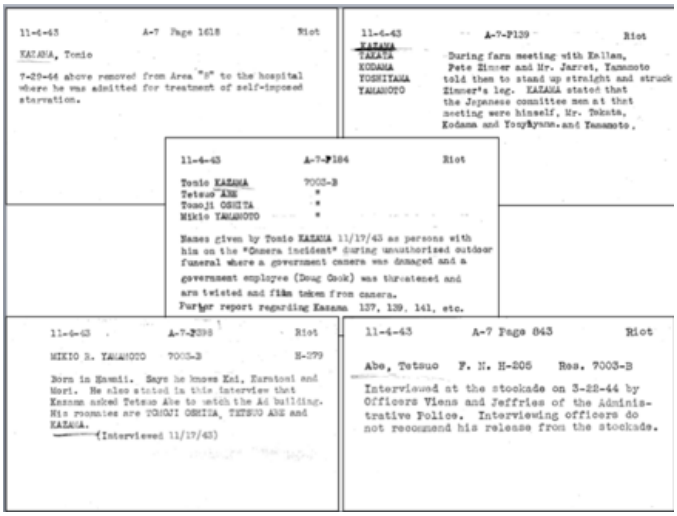

Fig. 3. Five incident cards on 11/04/1943 in the Tule Lake camp with overlapping names.

of Offenses, Organizations, Locations, etc. This multi-faceted tagging process is illustrated in Fig. 2.

In addition to indexing access points, it is possible that multiple cards can index the same people in different incident contexts. This pattern makes it possible to link individuals, incidents, and places as heterogeneous social networks, as illustrated in Fig. 3. Social network-based data modeling and analysis provide rich access protocols to users not only due to their abilities to surface hidden patterns that come from relational structures, but also their intractability that stems from modern graph databases and network visualization techniques.

Building upon these opportunities, we aim to create scalable infrastructures that can support the aforementioned operations against large datasets in a near real-time. This effort was possible through the Brown Dog project where deliverables include a new type of preservation repository called DRASTIC (i.e., Digital Repository At Scale That Invites Computa-

tion) [7]. The repository leverages a highly-scalable back-end NoSQL database, but more importantly supports the extraction of metadata attributes that can be associated with records and enable indexing and search. We are using DRAS-TIC to flexibly index and search on all the NLP/NER extracted terms, not just on the name index. DRAS-TIC is based on REST APIs and currently being integrated with the Fedora repository through an IMLS grant.

A. Computational Treatments of the Security Index Cards at Scale

The National Archives and Records Administration (NARA) is the repository of Record Group 210, Records of the War Relocation Authority (WRA). This record group includes paper records of internal security cases and their associated paper index cards for the ten Incarceration Camps in which Japanese-Americans were imprisoned during World War II. The workflow that we are developing for extracting descriptive metadata from OCR-ed images of the cards is as follows:

1) cleaning-up of typos and OCR errors in the scanned images of the index cards using crowdsourcing techniques,
2) development and performance evaluation of Index Term Identifiers to automatically identify index terms in the index cards and extract descriptive metadata that are essential in providing access points,
3) identification of additional resources needed to improve the performance of the Index Term Identifiers
4) format refinement for the descriptive metadata in light of the index term patterns discovered from the index cards, and
5) a plan for applying the resulting workflow across all 25,000 cards from 10 camps, incorporating privacy considerations.

B. Creating Name Gazetteers for Privacy Mining

The National Archives has requested the project team to identify index cards that include information about prisoners who were eighteen years old or less at the time of the event recorded on the index card. NARA's policy guides these cards not be released to the public (40% of the cards were vetted and shared with the DCIC in 2015). This presents an important Personally Identifiable Information (PII) computational test case. To support these decisions, detailed name gazetteers of internees are being built using two NARA record series: (1) the *Japanese-American Internee Data File, 1942-1946* with records of evacuated Japanese-Americans and (2) the *Final Accountability Rosters of Evacuees at Relocation Centers, 1944-1946* with records of evacuees at the time of their final release or transfer. We are very grateful to the directors of Densho.org[1] for supporting these efforts. The computational processing of potential PII information is part of the final workflow and determines the releasability of the cards to the public. This process also involves applying redaction processing on the digitized cards.

[1]Preserving, educating, and sharing the story of World War II-era incarceration of Japanese Americans: http://densho.org/

III. Preservation of Record Series, Documentation of Archival Operations on the Series, and Resources Used in Processing the Series

As new archival products are being developed from digital records, the procedures used in creating digital records from paper records need to be well documented and preserved. There are similarities between scientific products derived from scientific data and archival products derived from digital representations of record series. Scientific data needs to be corrected for noise introduced by scientific instrumentation. In a similar way, OCR-ed text from scanned images of paper records needs to be corrected for OCR errors. Also, both OCR-ed copies of paper records and born-digital records need to be corrected for typographical errors. The procedure for correcting measurement noise and OCR errors needs to be documented in a transparent way so that users can understand the differences between the original records and their derived records. Part of our approach to this issue is to crowdsource the corrections.

The documentation of archival procedures is being investigated in the context of OAIS Archival Information Packages (AIP). This package contains what is being preserved. Specifically, AIP contains (1) the scanned images of the index cards and a catalog of index terms of this series of images; (2) documentation of the processes of scanning paper index cards, OCR-ing the scanned images, and correcting the OCR errors and typos; (3) the linguistic analysis of the cards that identifies the classes of the objects that are referred to in the cards; (4) the styles of the cards; (5) the resources (Gazetteers and JAPE rules) created to automatically interpret the contents of the cards in order to generate the Catalog of index terms for the images of index cards; (6) the rules used to extract the metadata to support review of the cards; and finally, (7) references to external documents and data that are related to this record series but not included in this package.

Practically, this kind of process can be achieved through the use of a scalable archival platform as well. An example of the use cases is Jansen and Marciano's work on addressing the issue of long-term preservation of and access to digital government information [7]. This case shows how the preservation process can be enhanced through storing an infrastructure-independent representation of the raw data together with a model-dependency graph (an executable graph of database view mappings). To illustrate this approach, a case-study, the Florida Ballots Project for the 2000 Presidential election, was used. These workflows, tools, and infrastructures allow for the design of decision-support tools and services for improving government transparency and promoting citizen access to eGOV data.

IV. Implications for Sociotechnical Archival Systems

The pursuit of computational archives and the preservation of archival operations follow the CAS methodology described in [4]. In particular, these operations can be categorized as "evolutionary prototyping and computational linguistics"

among the CAS themes. At the time of the submission of this paper, we had validated the text extraction GATE workflow based on the analysis of two of the 21 boxes of Incident Cards, and we are incrementally adjusting the approach following the agile process paradigm. In other words, this workflow can be enhanced and refined as more records from boxes are added. More use cases and data will help refine the pattern matching rules and make the recognition performance more robust.

At the same time, these advances in archival processing techniques are barely possible without establishing well-functioning social subsystems such as collaborations and partnerships between archivists, domain experts, and cultural organizations. This iterative process that happens for refining the digital curation workflow often requires domain experts' nuanced interpretations of the digitized materials and the biases that are embedded in archival processing. Community groups, non-profits, and cultural organizations are the ones who understand the subtleties of government-produced content in all its biases and challenges, and provide domain knowledge that is essential in refining archival processing (e.g. Densho.org which under funding from several JACS grants is compiling the most precise name gazetteers to date). Meanwhile, it is also possible that advances in digital curation reshape stakeholders' social dynamics and archiving strategies, as part of the workflows is automated in an efficient way in the pipeline. Using crowdsourcing techniques for the correction of OCR errors is an example of the inter-dependency between technical and social subsystems where they altogether comprise a better archival system (while both subsystems co-shape each other) [8].

In this sense, developing a digital curation workflow in an iterative way can be conceptualized as a process to designing and implementing a sociotechnical system, which not only views new technologies as an innovative part of the system, but also understands operational processes, stakeholders' roles, and their interactions as part of the larger archival system [9]. The iterations we have gone though so far incorporate archivists, community activists, survivors of the Japanese-American camps and their family members, privacy and ethics experts, records managers, and computational linguistic scientists. Through this iteration, the overall archival practices evolve and are optimized by enhancing both social and technical subsystems. These need to be deepened.

Managing and establishing such heterogeneous actors, social networks, their norms/cultures, and archival artifacts for sustaining these types of projects are not trivial, especially in the era of big archival collections. Based on the theoretical implications for the sociotechnical archival systems, further studies need to develop assessment frameworks for better understanding the innovation process in the contexts of digital curation for big archival collections.

V. Conclusion

We used the WWII incident card collection as the case study to demonstrate automated generation of item-level metadata at scale. This particular case is emblematic of the whole family of record series that potentially covers hundreds of millions of records about people and organizations, thus posing continuing challenges for future archival practices. Our approach is based on computational treatments of archival collections and associated practices that changed and were shaped by both technical solutions and social mechanisms. By understanding this process as sociotechnical development of archival systems, we hope to provide a systematic framework for assessing digital curation workflows.

Acknowledgement

This project was conducted under a Sep. 2017 to May 2018 Research Agreement from NARA to the Digital Curation Innovation Center of the University of Maryland. The views and conclusions in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of NARA, or the U.S. Government. The Office of Innovation at NARA had released 40% of the cards back in 2015 for use by the DCIC and this last year's research agreement provided us access to the other 60% of the index cards for computational experimentation at NARA only. We acknowledge ongoing support from Maryland's iSchool and major funding from the NSF "Brown Dog" project (NSF Cooperative Agreement ACI-1261582). More recently, we obtained an IMLS grant [LG-71-17-0159-17], which helps us port Fedora on top of DRAS-TIC, an open-source platform.

References

[1] M. Greene and D. Meissner, "More product, less process: Revamping traditional archival processing," *The American Archivist*, vol. 68, no. 2, pp. 208–263, 2005.

[2] D. Labinsky and D. Hardin, "It's in the Cards: Finding Family Members in National Archives at St. Louis' Card Series, 2015 Virtual Genealogy Fair," 2015. [Online]. Available: https://www.archives.gov/files/calendar/genealogy-fair/2015/handouts/session-4-labinsky-hardin-presentation.pdf

[3] B. C. G. Lee, "Line detection in binary document scans: A case study with the international tracing service archives," in *Big Data (Big Data), 2017 IEEE International Conference on*. IEEE, 2017, pp. 2256–2261.

[4] R. Marciano, V. Lemieux, M. Hedges, M. Esteva, W. Underwood, M. Kurtz, and M. Conrad, "Archival records and training in the age of big data," in *Re-envisioning the MLS: Perspectives on the Future of Library and Information Science Education*. Emerald Publishing Limited, 2018, pp. 179–199.

[5] B. Latour, "Where are the missing masses? the sociology of a few-mundane artifacts," in *Shaping Technology/Building Society: Studies in Sociotechnical Change*, W. Bijker and J. Law, Eds. Cambridge, Mass.: MIT Press, 1992, ch. 10, pp. 151–180.

[6] W. Underwood, R. Marciano, S. Laib, C. Apgar, L. Beteta, W. Falak, M. Gilman, R. Hardcastle, K. Holden, Y. Huang *et al.*, "Computational curation of a digitized record series of WWII Japanese-American Internment," in *2017 IEEE International Conference on Big Data (Big Data)*. IEEE, 2017, pp. 2309–2313.

[7] G. Jansen and R. Marciano, "DRAS-TIC Measures: Digital Repository at Scale that Invites Computation (To Improve Collections)," *CNI Fall 2016 Project Briefing*, 2016. [Online]. Available: https://www.cni.org/topics/digital-curation/drastic-measures-digital-repository-at-scale-that-invites-computation-to-improve-collections

[8] W. J. Orlikowski, "Sociomaterial practices: Exploring technology at work," *Organization Studies*, vol. 28, no. 9, pp. 1435–1448, 2007.

[9] P. M. Leonardi, "Materiality, sociomateriality, and socio-technical systems: What do these terms mean? how are they different? do we need them," *Materiality and organizing: Social interaction in a technological world*, vol. 25, 2012.