DOI: 10.2352/issn.2168-3204.2016.1.0.135

Unlocking the Archives of Displacement and Trauma: Revealing Hidden Patterns and Exploring New Modes of Public Access through Innovative Partnerships and Infrastructure

Diane M. Travis (*), Myeong Lee (*), Magdalena Rojas (*), Allison Gunn (*), Anuj Nimkar (*), Gregory Jansen (*), Nicholas Diakopoulos (+), Richard Marciano (*)
University of Maryland College of Information Studies (iSchool) (*) and Phillip Merrill School of Journalism (+)
4105 Hornbake Building, South Wing
University of Maryland
College Park, MD 20742
301-405-9535 and 301-314-9145 fax

Abstract

This paper describes innovative partnerships: university - federal agency (between the University of Maryland and the Office of Innovation at the National Archives and Records Administration - NARA) and university - industry (between the College of Information Studies or "iSchool" at the University of Maryland and Archive Analytics Solutions Ltd.) where we are developing automated scalable workflows that involve digitization, OCR, information extraction, and linking into interactive maps and graph databases, and where digital preservation and archiving are performed using an innovative NoSQL Cassandra-based archival catalog and NetApp-based peta-scale storage infrastructure. This is a contribution to linking sensitive dispersed cultural resources involving the archives of displacement and trauma.

Background and Motivation

The goal of the "Revisiting Segregation Through Computational History" initiative at the UMD iSchool Digital Curation and Innovation Center (DCIC) is to develop new methods of digitizing and curating socially and politically important historical records in order to promote accessibility. To further this initiative, the project titled "Revisiting Segregation through Computational History: The Case of the World War II Japanese American Tule Lake Segregation Center" was launched. The purpose of this project was to explore the integration of archival and user-contributed data as well as investigate and prototype a GIS platform that links people, places, and events from distributed sources. The approach was inspired by the Digital Harlem Project [1], which developed a digital platform to query, map, and visualize legal records of ordinary citizens in Black Harlem between 1915 and 1930. The Tule Lake Project investigates and prototypes a platform that also links people, places, and events from distributed sources in a way that brings the archives to life, to be used as a resource for exploration and storytelling. We hope that the integration and access of these sensitive records through analytics and visualization, akin to a form of digital repatriation, will contribute to the healing and empowerment of the members of the community. As Eric Ketelaar writes, while "the violation of human rights is documented in the archives, the citizen who defends himself appeals to the archives."[2]

In 1942, President Franklin D. Roosevelt signed Executive Order 9066 when the United States entered the Second World War. The Order allowed all people on the west coast of the United States of Japanese descent to be taken from their homes and placed in containment centers by government officials. There was a fear of spying and terrorism attacks from individuals already within the U.S. borders. Over 120,000 people were placed in internment and segregation camps. There were ten camps in total in the United States. In addition, the Tule Lake Internment Camp became a Segregation Center after a 'loyalty' questionnaire was issued and the people who were deemed "disloyal" were crowded there. In the six years that the camp operated, not only did several federal government agencies create a vast quantity of records but the individuals interned there did as well. This project focuses on linking the varying types of records and data to explore the experiences of the internees for their descendants, historians, and researchers.

The data created about and within the Japanese-American Internment Campus are incredibly heterogeneous and are scattered across different repositories. As an example, resources that are located at the National Archives and Records Administration (NARA) include Camp "incident cards" and associated files (which describes alleged disciplinary infractions of internees which contain information about the people involved, dates, events, and locations), the Japanese-American Internee database of 109,000 names, 4,100 online War Relocation Authority (WRA) photos, and architectural camp maps and records. Located in the Densho Digital Archive are photographs, newsletters produced in the camps, and oral history files. The University of California Japanese American Relocation Digital Archives (JARDA) contains personal diaries, letters, photographs, drawings, and US War Relocation Authority materials, including camp newsletters, final reports, and photographs.

The rest of the paper covers the Infrastructure and Processes involved and the Innovative Partnerships developed. In addition, we frame the larger notion of the Archives of Displacement and Trauma, and conclude with Future Directions.

Infrastructure and Processes

There are several technical dimensions to this project: (1) the provisioning of the archiving and storage platform; (2) the establishment of digitizing equipment and ingestion processes; and (3) the development of tools and techniques for storing, processing, and visualizing the data.

Archiving and Storage Platforms

One of the main DCIC research projects is a $10.5M NSF/ DIBBs-funded initiative called "Brown Dog" with the University of Illinois NCSA Supercomputing Center and industry partners, NetApp and Archive Analytics Solutions, Ltd. The DCIC is creating

DOI: 10.2352/issn.2168-3204.2016.1.0.135

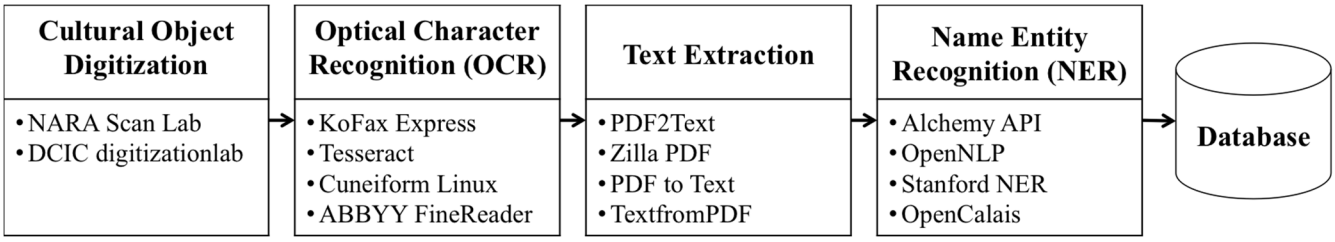

FIGURE 1. OPTIMIZED, SCALABLE INGESTION WORKFLOW

a data observatory to provide access to Big Records training sets and teach students practical digital curation skills using the infrastructure. The archiving infrastructure consists of three parts in the DCIC to support (1) large-scale data storage and analytics; (2) archival research; and (3) teaching and learning. The three systems are named *dataCave*, *virtualFarm*, and *vclCloud*.

An innovative infrastructure called the "*dataCave*" has been created to support hosting large-scale data. It is a peta-scale archival storage and analytics facility powered by NetApp storage and commercial Cassandra-based Indigo software for long-term archival storage and preservation. This storage allows up to 720 TB for archiving. With Indigo, digital curation workflows that invoke content extraction services can be automatically triggered on archival backend content.

In order to facilitate local research by allowing researchers to create individual virtual machines, "*virtualFarm*" has been created. It is a 14.4 TB platform, High Availability (HA) VMware environment. The purpose of this platform is not to host a large-scale data, but to enable many researchers to store and process mid-level data. In other words, whenever a new project is designed, a new virtual machine can be created to support the project by isolating the computing environment from other projects. This minimizes the amount of efforts to set up many physical machines while allowing each research project to have its own platform. In this reason, *virtualFarm* focused on the reliable and flexible management of virtual machines rather than scaling up the storage size.

"*VclCloud*" is an iSchool virtual lab environment consisting of a dashboard and virtual machine configuration functionalities based on Amazon Web Services (AWS). It was designed for supporting teaching and learning in a class environment, allowing easy creation of virtual machine instances with different configurations and OS's (e.g., Windows and Linux). In addition to the flexibility, the system also facilitates interactions between professors and students by allowing cloud monitoring and control sharing.

Digitizing Equipment and Ingestion Processes

To digitally curate historical materials and store them in the archiving platforms, digitizing equipment and processes are essential. The *digitizationLab* has been created to support digitizing activities and develop digitization processes. The *digitizationLab* is a lab space with various kinds of scanners from film-to-digital image converters to document scanners. For example, overhead ScanSnap SV600 scanners make it possible to digitize different kinds of documents such as books by minimizing distortions that come from the physical characteristics of the material.

Having this diverse equipment, digitization processes for different kinds of artifacts such as old postcards and books are under development by researchers and students. The recent work of the DCIC was the development of ingestion workflows targeting incident cards from the Japanese American WWII Incarceration Camp. Due to the large amount of incident cards with an inconsistent format and limited resources, it was necessary to develop an optimized and scalable process to properly archive the data. The ingestion strategy is shown in ***Figure 1***. This workflow is applicable to other materials, since digitizing artifacts has common features such as inconsistent format, fragmented data, and limited resources.

Tools and Techniques for Storing, Processing, and Visualizing the Data

In addition to physical and infrastructural settings, mapping interfaces and graph archives are being investigated as interaction and linking platforms. *Mapping interfaces* have the potential to accommodate a wide range of potential users from members of the Japanese American community, visitors to the National Parks, historians, genealogists, and researchers. The team decided to use the map of the Tule Lake Camp as a primary interface. A digitized georeferenced blueprint of the camp was created where individual features (i.e. living quarters, laundry, hospital) were coded using QGIS, an open-source geographical information system (GIS). This provides a hook for linking dispersed digital objects such as specific photos and newsletter stories.

This approach using GIS brings advantages in archiving data: (1) geographical data curation; (2) visualization; (3) geospatial queries; and (4) geo-tagged data navigation. GIS provides easy-to-use interfaces and tools for digitally curating historical data. For example, ArcGIS and QGIS have features that allow to geo-reference an old map on the online map and to draw polygons on the map to make it interactive. Furthermore, GIS makes it easy to visualize the stored data in combination with its own or third-party Javascript packages such as Leaflet or OpenLayers. Another advantage of GIS platforms is that it allows advanced geospatial queries based on the features of geospatial databases. For example, a query such as "find the closest artifact to the line between city A and city B" is possible if the data are stored in geospatial databases. Ultimately, users can navigate archival data more intuitively and extensively with the help of geospatial database features, enhanced queries of GIS, and visualization tools. The combination of front-end technologies and back-end systems may facilitate not only general users' search experience, but also researchers' data exploration. Therefore, the DCIC is trying to investigate as many GIS tools and technologies as possible to provide useful and impactful ways of storing, processing, and visualizing archival data.

Graph archives are also being evaluated. Pioneered by Tobias Blanke at King's College London as part of the EHRI European Holocaust Research Infrastructure project [3], they represent a new way linking dispersed cultural resources involving people, places, time, and events, into very large social networking-like graphs. Indeed they are based on NoSQL graph databases, the kinds used with Facebook's Social Graph, Google's Knowledge Graph, and Twitter's Interest Graph. Modeling the displacement of people with graph databases allows us to ask new comprehensive historical

DOI: 10.2352/issn.2168-3204.2016.1.0.135

questions: who was connected to whom at a moment in time and place, what relationships and relocation patterns were present, to name just a few.

FIGURE 2. A GEOREFERENCED MAP OF THE TULE LAKE CAMP

Innovative Partnerships

The success of this project was due to the innovative partnerships that were created and utilized. There were three types of partnerships facilitating this research: (1) university - federal agency, (2) university – industry, and (3) university – community.

In particular, the research team worked closely with:

(1) Staff from the Office of Innovation and Research Services (both from NARA) including John Martinez, Markus Most, Martha Murphy, Chris Naylor, and staff from the National Park Service (NPS) from the Tule Lake Unit including Larisa Proulx.

(2) Experts from Archive Analytics Solutions Ltd. including John Burns, Jerome Fuselier, Paul Watry, and initial support from CyArk.

(3) Prominent Tule Lake historians including Barbara Takei, Satsuki Ina, colleagues from King's College London including Mark Hedges, Tobias Blanke, experts from the US Holocaust Memorial Museum including Michael Levy, Michael Halley Goldman, members of the Japanese American camp survivor community, colleagues from the iSchool including Michael Kurtz, Greg Jansen, Katie Shilton, colleagues from UMD including Nick Diakopoulos, and finally GIS experts including Scott Madry from Informatics International Inc.

The partnership with NARA was innovative in the sense that expertise, resources, and funding were shared. The iSchool and NARA negotiated an innovative partnership where the iSchool leased scanning equipment for NARA to digitize the records, provided resources for NARA staff to OCR content, and prepared lookup databases for selection of incident cards.

The partnership led to the:

- Scanning and OCR the "incident cards".
- Culling of the names from the OCR text and captured in a database.
- Checking of the names against the full database of internees (also part of RG210), which includes birthdates.
- Identification of the names as being those of juveniles (under 18) and their redaction from the database of names and from the scanned card images before their delivery to the iSchool DCIC Center.
- Checking of the data ensure that FOIA b (6) redactions are made. Names of any juveniles were to be redacted from data and images.

Tule Lake experts and survivors meeting with the student team helped formulate sound and appropriate research questions through interactive sit-down sessions.

FIGURE 3. DR. SATSUKI INA MEETING WITH STUDENT TEAM ON MAY 4, 2015

During this visit neo4j graph database representations of patterns of Tule Lake deportation were demonstrated. See Figures 4 and 5.

FIGURE 4. VISUALIZATION OF PEOPLE DEPORTED TO TULE LAKE CLUSTERED BY CITY OF ORIGIN.

DOI: 10.2352/issn.2168-3204.2016.1.0.135

FIGURE 5. VISUALIZATION OF THE MOVEMENT OF SATSUKI INA'S FAMILY WITHIN THE TULE LAKE CAMP SYSTEM.

In addition on May 4, 2015, the iSchool and the DCIC hosted a movie screening on the Tule Lake Segregation Center. The documentary was "From a Silk Cocoon," on the World War II internment camp of Japanese Americans, included Q&A with producer, writer, and director Dr. Satsuki Ina and featured Dr. Ina's own family's experience in the Tule Lake Segregation Center.

The Archives of Displacement and Trauma

While the infrastructure and technical processes are non-trivial aspects of this work, there is another essential dimension. The collections of displacement and trauma (of which Tule Lake is part of) must be seen as part of a larger context of emotional and political frameworks. Members of a community who survive a large-scale trauma are changed by that experience. While these changes may look different in different individuals, the language choices researchers choose when interacting with these community members and the efforts to gain trust are universally critical when engaging the community. As seen in Bowers and Yehuda's recent work on intergenerational transmission of trauma, the children and grandchildren of the Japanese American survivors will evidence similar interactions with the world. [5] Research with communities of trauma will continue to be set in the larger context of emotional and political turmoil even when it is several generations past the traumatic event.

As of such, in this project the utmost thought and respect of the actual people whose stories were told in these diverse records went into the design and implementation of the project. Members of the Japanese American community, including notable professor and psychotherapist Dr. Ina Satsuki, were involved in every stage of the project. After the prototype was built, stakeholders from several groups engaged in an in-person meeting. This had never been done before with these stakeholders and it was critical for this project's success and sustainability.

There a shift to increasingly make archival materials available through digitization and online access. There are inherent issues in particular when this shift is applied to archives of displacement and trauma. There are concerns of privacy and safety of the community members. There are concerns of trust and who controls the narrative of the community. There are concerns of who controls the access of the records. And if great care is not taken when records are processed and made available to the public, then archival institutions can become instruments of further traumatization of the oppressed and displaced community.

Directions for Future Work

The DCIC focuses on "Big Records" management and archival analytics. These goals align closely with the Human Computer Interaction and Information Management (HCI & IM) NITRD Program focus area. NITRD is the Networking and Information Technology Research and Development Program (NITRD), part of the Office of Science and Technology at the White House and NARA is one of its member agencies.

Areas of future development include:

- Looking at frameworks for managing records documenting human rights abuse [4].
- Studying the requirements of other archives of displacement, relocation, and trauma in order to help drive infrastructure development. We are currently engaged with several other collections: urban transformations (redlining, gentrification, and urban renewal) and looted Holocaust WWII assets.
- Gaining a better understanding of access modalities concerning PII issues.
- Developing new "archival forensics" methods based on scale and automation, but also on the handling of record fragments, "archival dust," and information gaps.
- Applying the lessons learned from graph database modeling and visualization to incident cards at scale.

Acknowledgements

This work has been partially funded through a UMD/FIA Seed Grant (The Future of Information Alliance), and a NSF/DIBBs award (ACI-1261582). Staff and students involved in the FIA pilot included: Richard Marciano, Michael Kurtz, Greg Jansen, Susan Winter, and Andrew Barker, James Howland, Emily Keithly, Elizabeth Tobey, Karen Mawdsley, Dilip Bharadwaj, and Diane Travis.

References

[1] "Digital Harlem, Everyday Life 1915-1930," http://digitalharlem.org/.

[2] Eric Ketelaar, "Archival Temples, Archival Prisons: Modes of Power and Protection," Archival Science 2: 221-238, 2002.

[3] Tobias Blanke, Conny Kristel, "Integrating Holocaust Research," posted Dec. 27, 2014 on BigHumanitiesData blog, https://bighumanities.files.wordpress.com/2014/12/ehri-international-journal-of-ah-computing-final-pdf.pdf.

[4] Michelle Caswell, "Toward a Survivor-Centered Approach to Records Documenting Human Rights Abuse: Lessons from Community Archives," Archival Science 14: 307-322, 2014.

[5] Mallory Bowers, Rachel Yehuda, "Intergenerational Transmission of Stress in Humans," Neuropsychopharmacology, 41: 232-244, 2016.

Author Biography

Diane M. Travis is an Information Studies Doctoral Student at the College of Information Studies, University of Maryland. She has a Masters of Library Science, with a specialization in Government Information Management and Services, from the University of Maryland.

Myeong Lee is an Information Studies Doctoral Student at the College of Information Studies, University of Maryland. His interests are in urban dynamics, community information, and geospatial analytics.

DOI: 10.2352/issn.2168-3204.2016.1.0.135

Magdalena Rojas is a graduate student at the University of Maryland pursuing a Master's in Information Management with a focus on Archives and Digital Curation. She has a BA in Art History, Criticism and Conservation from UMD College Park.

Allison Gunn is a graduate student in the HiLS (History and Library Science) program at UM, where she studies American Southern Jewish history and digital curation in cultural institutions. She works as a research assistant at UMD and as a guide for the National Park Service.

Anuj Nimkar is a graduate student at the University of Maryland pursuing a Master's in Information Science with a focus on Data Analytics. He has worked as a Data Analytics intern with ERNST and Young in their Forensic Technology and Discovery Services team.

Gregory Jansen is a software architect in the DCIC Center at the UMD College of Information Studies (iSchool) and focuses on the building of platforms to manage digital objects, including access, digital preservation, and curation.

Nicholas Diakopoulos is an Assistant Professor of computational journalism at the Philip Merrill College of Journalism and he received a Ph.D. from the School of Interactive Computing at Georgia Tech. He focuses on the research, design, and development of computational media applications.

Richard Marciano is Director of the Digital Curation Innovation Center, a Professor in the iSchool at the University of Maryland College Park. His interests are in Big Records and archival analytics.