Automating the Detection of Personally Identifiable Information (PII) in Japanese-American WWII Incarceration Camp Records

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

William Underwood  
College of Information Science  
University of Maryland  
College Park, USA  
underwood@umd.edu

Mohammad Hanaee  
College of Information Studies  
University of Maryland  
College Park, USA  
mhanaee1@terpmail.umd.edu

Connor Mullane  
College of Information Studies  
University of Maryland  
College Park, USA  
cmullan1@umd.edu

Aakanksha Singh  
College of Information Studies  
University of Maryland  
College Park, USA  
asingh39@terpmail.umd.edu

Zayden Tethong  
College of Information Studies  
University of Maryland  
College Park, USA  
ztethong@terpmail.umd.edu

Abstract—We describe computational treatments of archival collections through a case study involving World War II Japanese-American WWII Incarceration Camps. We focus on automating the detection of personally identifiable information or PII. The paper also discusses the emergence of computational archival science (CAS) and the development of a computational framework for library and archival education. Computational Thinking practices are applied to Archival Science practices. These include: (1) data creation, manipulation, analysis, and visualization (2) designing and constructing computational models, and (3) computer programming, developing modular computational solutions, and troubleshooting and debugging. We conclude with PII algorithm accuracy, transparency, and performance considerations and future developments.

Keywords—Computational Thinking, Digital Curation, Computational Archival Science (CAS), PII, Privacy

I. CONSIDERATIONS ON PERSONALLY IDENTIFIABLE INFORMATION (PII)

In this paper, we focus on the curation, processing, and analysis of the records of the WWII Japanese-American Incarceration Camp card series 52 (from Record Group 210: Records of the War Relocation Authority), located at the National Archives and Records Administration (NARA) in Washington D.C. This series has not yet been released to the public and includes index cards containing information about internal security cases for the ten incarceration centers in which Japanese-Americans were imprisoned during World War II [1].

Currently, public access to these records is restricted to protect personally identifiable information (PII). NARA’s Privacy Policy Program 160 defines PII as “any information that can be used to distinguish or trace an individual’s identity, such as their name, Social Security Number, biometric records, etc. alone, or when combined with other personal or identifying information that is linked or linkable to a specific individual, such as date and place of birth, mother’s maiden name, etc.”

PII is specified in the Freedom of Information Act (FOIA), 5 U.S.C. §552, which allows for most federal records to be disclosed to the public unless they are exempt from disclosure by one of nine exemptions. Exemption 6 (also known as FOIA (b)(6) restriction) relates to documents which are “personnel and medical and similar files, the disclosure of which would constitute a clearly unwarranted invasion of personal privacy.”

2 RG210 Series 52 is currently FOIA (b)(6) restricted. In addition, NARA’s regulations establish a presumption that PII concerns end at the 75-year mark, per 36 CFR 1256.56 generally and especially subsection (a)(2).

Archivists are increasingly recognizing the importance of information privacy as seen through the Society of American Archivists’s (SAA) Core Values Statement and Code of Ethics 4, which focuses on access and use that is consistent with personal privacy. “Archivists select, preserve, and make available primary sources that document the activities of … individuals” and in the process “place access restrictions on collections to ensure that privacy and confidentiality are maintained, particularly for individuals and groups who have no voice or role in the collections’ creation, retention, or public use.” Similarly, one of the ten points from the


2 FOIAdvocates, See: http://www.foiadvocates.com/exemptions.html

3 The Electronic Code of Federal Regulations (e-CFR) is divided into 50 titles that represent broad areas subject to Federal regulation. Title 36, Chapter XII, Subchapter C, Part 1256, describes access to records and donated historical materials Part 1256.56 relates to information that would invade the privacy of a living individual. In accordance with FOIA (b)(6), NARA withholds records about a living individual that would reveal details of a highly personal nature that, if released, would cause a clearly unwarranted invasion of personal privacy. See: https://www.ecfr.gov/cgi-bin/text-idx?SID=212c31b61ee5ea2d0d03ca4ab31e4a9&mce=tre&node=pt36.3.125

International Council on Archives’ (ICA) Code of Ethics focuses on the need to respect “the personal privacy of individuals who created or are the subjects of records, especially those who had no voice in the use or disposition of the materials.”

In an earlier effort to help with the release of the 25,000+ cards in this series, NARA’s Special Access and FOIA office devised a PII release policy for adults referenced in cards only, out of concern that cards involving younger individuals be interpreted as “juvenile criminal” records. This policy was used by staff at the Office of Innovation to release part of the collection to the Digital Curation Innovation Center (DCIC) at the University of Maryland iSchool. Nearly 10,000 cards were transferred, representing 40% of the series. The other cards were to be redacted, but this was never completed. Redaction is the process of removing exempted information from a copy of a record destined to the public.

In this paper, we report on the use of the 10,000 released cards for the purpose of demonstrating the creation and testing of an algorithm that automates the detection of PII. Balancing privacy and access is an important topic in the age of digital curation, where datafication and the use of big data can lead to compromising privacy through the application of predictive analytics. As noted in [3] “powerful tools for extracting information from data have the potential to significantly enhance the human condition. However, … this comes with the potential for misuse.”

II. THE EMERGENCE OF COMPUTATIONAL ARCHIVAL SCIENCE

The last two years have seen the emergence of the concept of “Collections as Data” in cultural heritage institutions, where computational methods and tools are increasingly leveraged to enhance library and archives collections:

“Combined with an increasing flow of born digital items, digital library collections have come to represent a rich community resource for users... Yet a focus on replicating traditional ways of interacting with collections in a digital space does not meet the needs of the researcher, the student, the journalist, and others who would like to leverage computational methods and tools to treat digital library collections as data.”

David Weintrop et al. [4] further refine computational thinking concepts by envisioning a set of computational practices consisting of four main categories: (1) data practices, (2) modeling and simulation practices, (3) computational problem solving practices, and (4) systems thinking practices. Computational thinking is a form of problem solving that uses modeling, decomposition, pattern recognition, abstraction, algorithm design, and scale [5].

Similarly, the Digital Curation Innovation Center at the U. Maryland iSchool is developing a larger agenda to “infuse computational thinking” into archival science, as demonstrated in their work on Computational Archival Science (CAS) [6], defined as:

An interdisciplinary field concerned with the application of computational methods and resources to large-scale records / archives processing, analysis, storage, long-term preservation, and access, with the aim of improving efficiency, productivity and precision in support of appraisal, arrangement and description, preservation and access decisions, and engaging and undertaking research with archival materials.

In this paper, we attempt to demonstrate the relevance of computational thinking concepts adapted from the computational practices taxonomy to the automation of the detection of juvenile PII in Japanese-American WWII Incarceration Camp Records.

III. DEVELOPING A COMPUTATIONAL FRAMEWORK FOR LIBRARY AND ARCHIVAL EDUCATION

The University of Maryland received a grant from the Institute of Museum and Library Services (IMLS) to host a major planning workshop of library and archival educators and technologists. This Workshop will identify the foundation and building blocks for an integrated library and archival Master’s educational curriculum centered on computational treatments of large complex collections. This is designed to prepare the next generation of librarians and archivists to meet the evolving needs of professionals working with digital collections. This Workshop will be held in conjunction with the 2019 iConference meeting hosted at the University of Maryland in April 2019 and builds on two years of CAS

---


meetings and interactions.\(^8\)

Our paper demonstrates the first steps of mapping computational thinking concepts to archival science practices. We hope this approach will lead to identifying canonical collections, analytical methodologies, open source software applications, and developing an initial set of end-to-end repurposeable lesson plans to be disseminated and shared with the archival community at large.

In preparation for this workshop, next we attempt to adapt the Weinrop computational practices taxonomy to the problem at hand, using the first three categories. Our team comprises iSchool students in the undergraduate Information Science program and Master’s students in the MLIS (Master of Library and Information Science) and MIM (Master of Information Management) programs.

\(A.\) Creating Data

"The increasingly computational nature of working with data in " archival science "underscores the importance of developing computational thinking practices in the classroom." “Part of the challenge is teaching students that answers are drawn from the data available.” “In many cases” archivists "use computational tools to generate data... at scales that would otherwise be impossible.”

The National Archives requested that the project team identify index cards with information about internees 18 years old or younger at the time of the event recorded on the index card so that these cards could be redacted. To support these release decisions, detailed name gazetteers of internees are needed. These include: (1) the “Japanese-American Internment Data File, 1942 – 1946”, with records of evacuated Japanese-Americans, also known as WRA Form 26, and (2) the “Final Accountability Rosters of Evacuees at Relocation Centers, 1944-1946, also known as FAR, with records of evacuees at the time of their final release or transfer. We are very grateful to staff from Densho\(^9\) for supporting these research efforts by sharing their research datasets.

\(B.\) Manipulating Data

"Computational tools make it possible to efficiently and reliably manipulate large and complex archival holdings. Data manipulation includes sorting, filtering, cleaning, normalizing, and joining disparate datasets."

We use data cleaning software such as OpenRefine to clean and normalize column values. Figure 5 shows how incident date formats are standardized.

---

\(^8\) The Computational Archival Science (CAS) Portal, see: [http://dcicblog.umd.edu/cas/](http://dcicblog.umd.edu/cas/)

C. Analyzing Data

“There are many strategies that can be employed when analyzing data for use in an archival context, including looking for patterns or anomalies, defining rules to categorize data, and identifying trends and correlations.”

Once we calibrate and iteratively tune the NLP/NER module, which involves the creation of recognition patterns and rules, we are able to automatically process all 10,000 incident cards.

D. Visualizing Data

“Communicating results is an essential component of understanding archival data and computational tools can greatly facilitate that process. Tools include both conventional visualizations such as graphs and charts, as well as dynamic, interactive displays.”

As part of the interactive server-based shared version of Jupyter Notebook used in this project (see section I for details), we use a simple bar chart plotting function as part of the Python matplotlib Library. This allows us to quickly assess the distribution of dates across all three input datasets.
E. Designing Computational Models

“The ability to create, refine, and use models of phenomena is a central practice.” “Models can include **flowcharts and diagrams**.” “Part of taking advantage of computational power... is designing new models that can be run on a computational device.” “There are many reasons that might motivate designing a computational model, including wanting to better understand a phenomenon under investigation, to test out a hypothesis.” “Students... will be able to define the components of the model, describe how they interact, decide what data will be produced by the model.”

The flowchart in Figure 6 shows an algorithm that loops through the consideration of metadata from each of the index cards. It checks to see if the index term is a Japanese name. If it is, the person is an internee. If not, the index term is a non-Japanese name (e.g., an administrator or staff name), an organization name, a job title, or a topic. Next, based on the lookup of the Japanese name in the FAR registry, it computes the internee’s age from the incident date and the internee’s birthdate to determine whether the named person is a minor or not. If the internee is not found in the FAR, a secondary lookup is performed on the WRA Form 26 name registry. The initial algorithm leads to three outcome values: **Releasable**, **Not Releasable** (meaning the card should be flagged for possible PII), **Not Determined** (meaning that the PII detection process was inconclusive).
F. Constructing Computational Models

"An important practice is the ability to create new or extend existing computational models. This requires being able to encode the model features in a way that a computer can interpret."

Pseudocode is used in this phase. This is a plain English and readable notation resembling a simplified programming language and used in program design.

```
FOR EACH of the 10,000 cards:
  IF cardname is Japanese:
    MATCH cardname in the FAR registry
    IF MATCH is true:
      COMPUTE date difference
      IF difference > 18:
        RELEASE card
      ELSE:
        DO NOT RELEASE card: possible PII
    ELSE:
      MATCH cardname in the Form 26 registry
      IF MATCH is true:
        COMPUTE date difference
        IF difference > 18:
          RELEASE card
        ELSE:
          DO NOT RELEASE card: possible PII
      ELSE:
        PII Status NOT DETERMINED
  ELSE:
    RELEASE card: WRA Staff
```

Figure 9: Flowcharting an initial PII detection algorithm

Figure 10: Pseudocode of the PII Detection Algorithm
G. Computer Programming

"Enabling students to explore" archival problems “using computational problem solving practices such as programming, algorithm development, and creating computational abstractions.” "The ability to encode instructions in such a way that a computer can execute them is a powerful skill for investigating" archival problems. Programs include ten-line Python scripts.”

Students map the pseudocode of Fig. 10 into Python code.

```python
for i in range(0, 10000):
    if interest(cardName):
        print_cardInfo(i)
    if card_Name in FAR_list:
        FAR_lookup(i)
        PII_DateCheck(i)
        if age(i) > 18:
            print('RELEASE')
        else:
            print('NO RELEASE')
    elif card_Name in FORM26_list:
        FORM26_lookup(i)
        PII_DateCheck(i)
        if age(i) > 18:
            print('RELEASE')
        else:
            print('NO RELEASE')
    else:
        print('NOT DETERMINED')
else:
    print('RELEASE')
```

Figure 11: Python script of the PII Detection Algorithm

We also make use of the powerful pandas 10 software library for Python, designed for data manipulation, cleaning, and analysis.

H. Developing Modular Computational Solutions

"When working toward a specific" archival “outcome, there are often a number of steps or components involved in the process; these steps, in turn, can be broken down in a variety of ways that impact their ability to be easily reused, repurposed, and debugged. Developing computational solutions in a modular, reusable way has many implications. By developing modular solutions, it is easier to incrementally construct solutions, test components independently, and increase the likelihood that components will be useful for future problems.”

Figure 11 illustrates the use of abstraction and functional programming through the use of modular components such as: PII_DateCheck(), FORM26_lookup(), and FAR_lookup(). This allows for reusable chunks of code that can be tested locally. The larger program is the composition of these modules, which makes it both more readable and maintainable.

I. Troubleshooting and Debugging

"Troubleshooting broadly refers to the process of figuring out why something is not working or behaving as expected. There are a number of strategies one can employ while troubleshooting a problem, including clearly identifying the issue, systematically testing the system to isolate the source of the error, and reproducing the problem so that potential solutions can be tested reliably.”

To facilitate group debugging, we use an interactive server-based shared version of Jupyter Notebook. "The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more.”

IV. PII Algorithm Performance

In 1942 a network of 10 relocation centers was created from California to Arkansas. See Figure 12.

The complete 10 Camp "Incident Card" collection consists of a total of 25,045 index cards, spanning 21 boxes:

<table>
<thead>
<tr>
<th>Camp</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topaz UT</td>
<td>1%</td>
</tr>
<tr>
<td>Poston AZ</td>
<td>5%</td>
</tr>
<tr>
<td>Gila River AZ</td>
<td>6%</td>
</tr>
<tr>
<td>Granada CO</td>
<td>3%</td>
</tr>
<tr>
<td>Heart Mountain</td>
<td>2%</td>
</tr>
<tr>
<td>Manzanar CA</td>
<td>9%</td>
</tr>
<tr>
<td>Minidoka ID</td>
<td>9%</td>
</tr>
<tr>
<td>Tule Lake CA</td>
<td>62%</td>
</tr>
<tr>
<td>Rohwer/Jerome AK</td>
<td>2%</td>
</tr>
</tbody>
</table>

Figure 13: Total Index Card Counts for 10 Camps

In summary:

- 62% of the cards (or 15,648) are from Tule Lake, CA
- 92% of the cards are from 5 camps (Tule Lake CA, Minidoka ID, Manzanar CA, Gila River AZ, and Poston AZ, respectively)

---

10 Pandas Cheat Sheet for Data Science in Python, see: https://www.datacamp.com/community/blog/python-pandas-cheat-sheet


12 Project Jupyter, see: http://jupyter.org/
Of these 25,045 cards, only 9,996 (almost 10,000 or 40%) were released to the DCIC Center. This represents a substantial portion of each of the 10 camps. The numbers are as follows:

- Topaz: 186/364 (or 51%),
- Poston: 309/1,202 (or 26%),
- Gila River: 605/1,578 (or 38%),
- Granada: 186/763 (or 24%),
- Heart Mountain: 249/533 (or %47),
- Manzanar: 736/2,146 (or 34%),
- Minidoka: 767/2,343 (or 33%),
- Tule Lake: 6,769/15,648 (or 43%), and
- Rohwer/Jerome: 189/468 (or 40%).

![Figure 14: Index Card Counts Released to DCIC](image)

Preliminary testing at the time of the publication of this paper was conducted on the incident cards released for Tule Lake from Box8 (Tule Lake spans Box8 through Box21).

![Figure 15: Pictures of Box8](image)

Of the 247 Tule Lake records in Box8, the PII algorithm performed as follows:
- 92% or 228 of the 247 cards were identified as releasable (with 3 found in the secondary Form26 lookup)
- Of the 19 that were not released:
  - 12 (5%) are minors (with 11 found in the FAR lookup and 1 in the Form 26 lookup) with nine 18-year olds, one 14-year old, one 6-year old or one 1-year old
  - 7 (3%) are “Not Determined” (neither found in the FAR nor in the Form26)

This means that 240 (97%) were PII triaged as “Release” or “Don’t Release”. We are in the process of extending the testing to all 14 Tule Lake boxes (14 of 21), where our Tule Lake sample represents 43% or 6,769 of the 9,996 cards being tested in the lab.

From Box8 we processed 247 index cards, 7 of which resulted in being labeled as “Not Determined”, meaning that the name lookup did not yield a match in the FAR or Form 26 registries. We believe the number of “Not Determined” cards can be further reduced by checking the validity of the names on the cards. For example, card 470 is listed as Hayato AKABEPPU and it is very likely that the real name is OKABEPPU Hayata, which can be found in the FAR.

V. CONCLUSIONS AND FUTURE WORK

This paper demonstrates the value of: (1) supplementing human resources in labor intensive archival review by introducing computation to the archival review of records with probable PII restrictions, and (2) introducing library and archival science students to computational thinking practices that can be applied to archival practices.

Regarding the specifics of this study, cards like card 546, which references AKIYAMA Tome, is very likely to match the FAR entry of AKIYAMA Tomeshichi. An alternative to dealing with ambiguous cards is to pre-process the entire card collection and create deeper profiles for the individuals referenced in the cards using richer data (such as Residence IDs and Family numbers on the cards to infer missing information). This will provide deeper context and create stronger recognition patterns, a process we will balance carefully with privacy concerns. Also, we are only using the Tule Lake FAR registry in this first instantiation of the PII detection algorithm and 9 additional FAR registries will soon be available from Densho (one for each of the other camps), creating a greater probability that the individuals on the incident cards will be referenced across other camp registries as well.

The next iteration of the algorithm will also go beyond processing the indexed name and we will reach deeper and try to assess whether other individuals referenced in the remarks section pose privacy challenges. We will also explore ways of making this PII detection algorithm more transparent and produce arguments that address “algorithmic accountability”. The focus is on avoiding a black-box design and providing enough feedback for the procedure to be audited.

ACKNOWLEDGMENT

We wish to acknowledge 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 is helping us port Fedora on top of DRAS-TIC, an open-source platform (Digital Repository At Scale - That Invites Computation). More details at: [http://dcic.umd.edu/about-us/infrastructure/](http://dcic.umd.edu/about-us/infrastructure/). A second IMLS grant [RE-73-18-0105-18] focuses on planning a workshop of library and archival educators and technologists.

REFERENCES


