Reframing Digital Curation Practices through a Computational Thinking Framework

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

Sarah Agarrat
Computer Science & Math.
University of Maryland
College Park, USA
sagarrat@terpmail.umd.edu

Hannah Frisch
College of Information Studies
University of Maryland
College Park, USA
hfrisch@g.umd.edu

Margaret Rose Hunt
College of Information Studies
University of Maryland
College Park, USA
mrhunt@umd.edu

Kanishka Jain
College of Information Studies
University of Maryland
College Park, USA
kjain307@umd.edu

Genevieve Kocienda
College of Information Studies
University of Maryland
College Park, USA
gjkocienda@mac.com

Hannah Krauss
College of Information Studies
University of Maryland
College Park, USA
tkrauss@terpmail.umd.edu

Chenxi Liu
College of Information Studies
University of Maryland
College Park, USA
lcx@umd.edu

Mary McKinley
College of Information Studies
University of Maryland
College Park, USA
mmckin@terpmail.umd.edu

Danish Mir
College of Information Studies
University of Maryland
College Park, USA
dmir1@umd.edu

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

Emery Patterson
College of Information Studies
University of Maryland
College Park, USA
epatters@umd.edu

Debashish Pradhan
College of Information Studies
University of Maryland
College Park, USA
devp@terpmail.umd.edu

James Santos
College of Information Science
University of Maryland
College Park, USA
jamesnsantos@gmail.com

Britton Schams
College of Information Studies
University of Maryland
College Park, USA
bschams@umd.edu

Hilary Szu Yin Shiue
College of Information Studies
University of Maryland
College Park, USA
hshiue@umd.edu

Andy Jose Silva
College of Information Studies
University of Maryland
College Park, USA
andysilva301@gmail.com

Mayhah Suri
College of Information Studies
University of Maryland
College Park, USA
mayhah.suri@gmail.com

Tahura Turabi
Info. Sci & Japanese
University of Maryland
College Park, USA
tahura@umd.edu

Mirielle Vasselli
College of Information Studies
University of Maryland
College Park, USA
mirielle.vasselli@gmail.com

Jiale Xu
College of Information Studies
University of Maryland
College Park, USA
jlxu2016@umd.edu

***Abstract*—We describe the value of reframing digital curation practices through a computational thinking (CT) framework. Using a case study that demonstrates computational treatments of World War II Japanese-American Incarceration Camp Records, we demonstrate the applicability of CT with respect to: (1) Detecting personally identifiable information, (2) Developing name registries, (3) Integrating vital records, (4) Designing controlled vocabularies, (5) Mapping events and people, and (6) Connecting events and people through networks. The work was carried out by 5 teams of students in an 8-week digital curation exploration and development sprint.**

Keywords—Computational Thinking, Digital Curation, Computational Archival Science (CAS), Japanese American WWII Incarceration Camps.

I. Digital Curation and the Emergence of Computational Archival Science (CAS)

"The use of emergent technologies have profoundly altered the nature of archives, by disrupting how information is created, recorded, captured, encoded, curated, shared, made available and used" (E. Goudarouli, 2019). These changes are affecting libraries and archives of all sizes. This is apparent in the decision of the National Archives and Records Administration (NARA) to stop accepting paper-based records at the end of 2022 (Fedscoop, 2019). Similarly, The Library of Congress has just launched a $1M Mellon-funded project called "Computing Cultural Heritage in the Cloud (CCHC)" (Mellon, 2019), the goal of which is to experiment with digital collections as Big Data. In the IMLS "Collections as Data" project *(LG-73-16-0096-16),* computational treatments of collections are also illustrated where "a Digital Humanities researcher engages in term frequency visualization, topic modeling, and network analysis across thousands and sometimes even millions of

978-1-7281-0858-2/19/$31.00 © 2019 IEEE

items." Beyond text data "the scope of data extends to images, moving images, sound, web archives, and beyond."

Preparing MLIS students for this changing landscape is essential. While courses in archival methods typically cover records appraisal, arrangement, description, preservation, and access, there is a new digital curation imperative as observed in the "Archival Records and Training in the Age of Big Data" paper (Marciano et al.) [1]:

> *Digital curation [increasingly]… extend[s] archiving and preservation by adding value to digital objects, through indexing; adding metadata, annotation, or markup of various forms, including semantic markup/ontologies (using both manual and automated methods); enhanced discovery and access (including retrieval, visualization); and facilitating interoperability and integration… Digital curation is concerned with curating digital objects and information in all their varied guises. Such digital assets have a key role to play not only in scholarly research, or in the domains traditionally associated with the management of information, such as libraries, archives, and other memory institutions, but also in a much broader range of institutions and activities.*

The Digital Curation Innovation Center (DCIC) at the U. Maryland iSchool is developing a larger digital curation agenda that explores the computational move towards "Big Cultural Data", as demonstrated in their work on Computational Archival Science (CAS), 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 [1].*

II. Mapping Computational Thinking to Library and Archival Science Education and Research

The DCIC is now conducting research with the explicit goal of exploring how Computational Thinking (CT) practices can be incorporated into Masters-level curricula in Library and Archival Science Education and Research (CT-LASER). An on-line repository called CASES (Computational Archival Science Educational System: http://cases.umd.edu) is being explored for storing and providing access to open-source cloud-based Jupyter notebooks that record the results of performing archival tasks on digital records. The goal is to enable a collaborative network of educators and practitioners who can learn from one another through the sharing and dissemination of computational case studies and lesson plans.

David Weintrop et al. [2] further refine CT concepts by envisioning a set of computational practices covering: (1) data, (2) modeling and simulation, (3) computational problem solving, and (4) systems thinking. CT is a form of problem solving that uses modeling, decomposition, pattern recognition, abstraction, algorithm design, and scale [3]. We provide a summary of these 22 CT practices spread across these 4 categories next. We have started the remapping of these concepts to archival science.

Towards presenting a self-contained paper, the rest of this section provides an italicized paragraph taken from [2] that describes the meaning of each of the 22 computational thinking practices, where we also substituted the original mathematics and science terms for archival science terms, highlighting them in bold to demonstrate the relevance of mapping computational thinking practices to archival science practices.

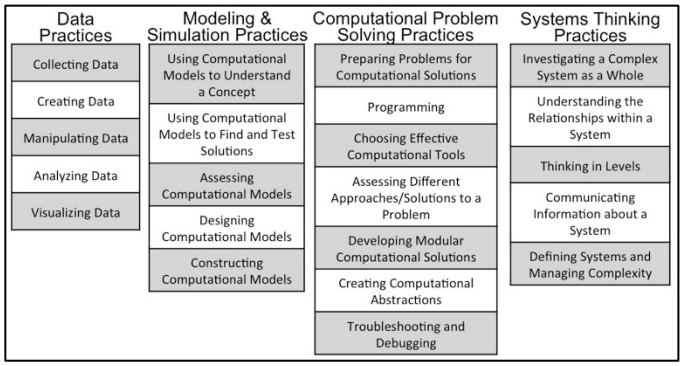


Figure 1: Computational thinking in math and science practices taxonomy

A. Data Practices

The nature of how data are collected, created, analyzed, and shared is rapidly changing primarily due to advancements in computational technologies.

1. *Collecting Data*
   "*Data are collected through observation and measurement. Computational tools play a key role in gathering and recording a variety of data across many different* ***archival*** *endeavors. Computational tools can be useful in different phases of data collection, including the design of the collection protocol, recording, and storage***."**
2. *Creating Data*
   "The increasingly computational nature of working with ***archival data*** *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."*
3. *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."*
4. *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."*
5. *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."*

B. Modeling & Simulation Practices

The ability to create, refine, and use models of **archival** phenomena is a central practice… Models can include flowcharts and diagrams.

1. *Using Computational Models to Understand a Concept*
   "Computational models that demonstrate specific ideas or phenomena can serve as powerful learning tools. Students can use computational models to deepen their understanding of **archival science**.
2. *Using Computational Models to Find and Test Solutions*
   "Computational models can be used to test hypotheses and discover solutions to problems. They make it possible to test many different solutions quickly, easily, and inexpensively before committing to a specific approach."
3. *Assessing Computational Models*
   "Students who have mastered this practice will be able to articulate the similarities and differences between a computational model and the phenomenon that it is modeling."
4. *Designing Computational Models*
   "Part of taking advantage of computational power… is designing new models that can be run on a computational device. Students… will be able to define the components of the model, describe how they interact, decide what data will be produced by the model."
5. *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."

C. Computational Problem Solving Practices

Problem solving is central to **archival** inquiry.

1. *Preparing Problems for Computational Solutions*
   "While some problems naturally lend themselves to computational solutions, more often, problems must be reframed so that existing computational tools can be utilized. Strategies for doing this include decomposing problems into subproblems, reframing new problems into known problems for which computational tools already exist, and simplifying complex problems so the mapping of problem features onto computational solutions is more accessible."
2. *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."*
3. *Choosing Effective Computational Tools*
   "Students who have mastered this practice will be able to articulate the pros and cons of using various computational tools and be able to make an informed, justifiable decision."
4. *Assessing Different Approaches/Solutions to a Problem*
   "When there are multiple approaches to solving a problem, or multiple solutions to choose from, it is important to be able to assess the options and make an informed decision about which route to follow. Even if two different approaches produce the same correct result, there are other dimensions that should be considered when choosing a solution or approach such as cost, time, durability, extendibility, reusability, and flexibility."
5. *Developing Modular Computational Solutions*
   "Students who have mastered this practice will be able to develop solutions that consist of modular, reusable components and take advantage of the modularity of their solution in both working on the current problem and reusing pieces of previous solutions when confronting new challenges."
6. *Creating Computational Abstractions*
   *"The ability to create and use abstractions is used constantly across **archival science** undertakings, be it creating computational abstractions when writing a program, generating visualizations of data to communicate an idea or finding, defining the scope or scale of a problem or creating models to further explore or understand a given phenomenon."*
7. *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."

D. Systems Thinking Practices

Systems thinking analyses… focus on an inclusive examination of how the system and its constituent parts interact and relate to one another as a whole.

1. *Investigating a Complex System as a Whole*
   "Students who have mastered this practice will be able to pose questions about, design and carry out investigations on, and ultimately interpret and make sense of, the data gathered about a system as a single entity… Computational tools such models and simulations are especially useful in such investigations."
2. *Understanding the Relationships within a System*
   "Computational tools are useful for conducting such inquiry as they can provide learners with controls for isolating different elements, investigating their behaviors, and exploring how they interact with other components of the system."
3. *Thinking in Levels*
   "Students who have mastered this practice will be able to identify different levels of a given system, articulate the behavior of each level with respect to the system as a whole, and be able to move back and forth between levels, correctly attributing features of the system to the appropriate level."
4. *Communicating Information about a System*
   "Students who have mastered this practice will be able to communicate information they have learned about a system in a way that makes the information accessible to viewers who do not know the exact details of the system from which the information was drawn."
5. *Defining Systems and Managing Complexity*
   "Students who have mastered this practice will be able to define the boundaries of a system so that they can then use the resulting system as a domain for investigating a specific question as well as to identify ways to simplify an existing system without compromising its ability to be used for a specified purpose."

In papers [4], [5], and [6] we provide a case for CT to LASER mapping with concrete examples. However, while these innovative developments are taking place, there are still many challenges for students in developing digital curation projects from scratch and limited exemplars. In this paper, we suggest how a number of core digital curation areas can be reframed through the common lens of CT practices. This approach is informed by a case study on the computational treatments of Japanese-American WWII Incarceration Camp Records.

III. The Case of the Japanese-American WWII Incarceration Camp Records

In partnership with Densho**,** whose mission is to **"**Preserve, educate, and share the story of World War II-era incarceration of Japanese Americans: http://densho.org/), we are exploring computational treatments of WWII historical archival datasets.

In 1942 a network of 10 incarceration camps was created from California to Arkansas (see **Figure 2**). 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:

(1) The *"Japanese-American Internee Data File, 1942 – 1946"*, with camp intake records of evacuated Japanese-Americans, also known as WRA Form 26.

(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.

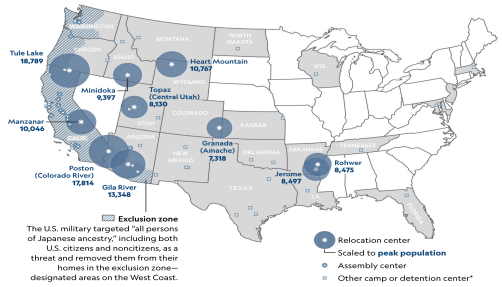
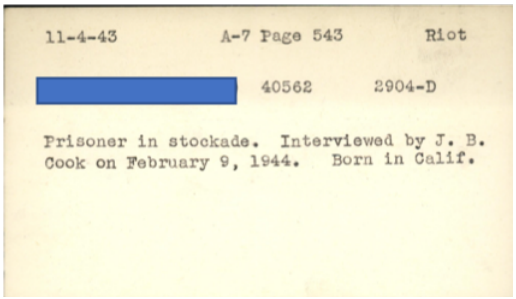


Figure 2: Network of 10 WWII Incarceration Camps [1]

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 camp. The DCIC was granted research access to a number of these cards. **Figure 3** 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.


11-4-43 A-7 Page 543 Riot

[redacted] 40562 2904-D

Prisoner in stockade. Interviewed by J. B. Cook on February 9, 1944. Born in Calif.

Figure 3: "Incident" card for an incarceree in the Tule Lake camp.

IV. Digital Curation Student Sprints

The work conducted in this paper was carried out through a "Digital Curation Sprint" in the fall of 2019, where 22 students were selected at the U. Maryland iSchool to work together over an 8-week period leading to a public event and poster presentation on October 30, 2019 called *"Resistance at Tule Lake: A Conversation with the Filmmaker and iSchool Digital Curators (and Film Viewing)."* Students were split across five interdisciplinary teams in order to promote College interaction by combining students from various programs: Undergraduate programs [iSchool InfoSci Information Science, CS Computer Science, and Japanese], and Graduate programs [MLIS Master of Library and Information Science, MIM Master of Information Management, and HCIM Human-Computer Interaction Master]. All interactions took place in the Digital Curation Innovation Center (DCIC) lab space (http://dcic.umd.edu), driven by the DCIC's goals to:

> *"Sponsor interdisciplinary projects that explore the integration of archival research data, user-contributed data, and technology to generate new forms of analysis and historical research engagement, particularly in the arenas of social justice, human rights, and cultural heritage."*

V. Detecting Personally Identifiable Information

This section recaps student work conducted over an earlier and distinct 8-week sprint in the fall of 2018 and described in greater detail in [5]. Students were: Mohammad Hanaee (InfoSci), Connor Mullane (MLIS), Aakanksha Singh (MIM), and Zayden Tethong (MLIS).

A. Digital Curation Significance

Personally Identifiable Information (PII) is a core digital curation topic when providing public access to records that contain information that can reveal an individual's identity directly (name, SSN, etc.), or indirectly through the linking of other personal information (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] 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).

B. Case Study Connections

These computational activities were matched to descriptions of the 22 computational thinking practices. It was found that these activities correspond to ten of the computational thinking Practices. See **Figure 4**.

[1] From: https://www.nationalgeographic.com/magazine/2018/10/japanese-internment-then-now-portraits/

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

C. Computational Thinking Patterns

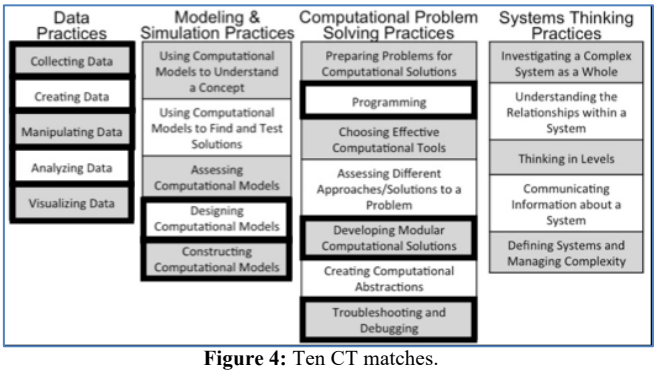


Figure 4: Ten CT matches.

The ten relevant CT Practices include:

- *Five **Data Practices**:* (1) *Collecting Data* – a scanner collected digital images of index cards (2) *Creating Data* -- *Abbyy FineReader* was used to create digital text from scanned images of paper index cards; (3) *Manipulating Data – Open Refine* was used to clean and normalize the data; (4) *Analyzing Data* in which The *General Architecture for Text Engineering (GATE)* was used to analyze the digital text and extract metadata by creating rules to perform Named Entity Recognition (NER); and (5) *Visualizing Data* -- The *Python* programming library called *mathplotlib was used* to create graphs and charts to visualize and understand the results of the analysis.
- *Two **Modeling & Simulation Practices***: (1) *Designing Computational Methods* -- A flowchart was created that represents the input metadata, the input FAR and WRA Form26 databases, the computations on the data and the decisions necessary to conclude whether an index card has PII requiring restriction on release; and (2) *Constructing Computational Models* -- Pseudocode was constructed from the flowchart.
- *Three **Computational Problem Solving Practices***: (1) *Programming* -- The pseudocode was encoded into the *Python* programming language, (2) *Developing Modular Computational Solutions* -- Python functions were developed for looking up dates in the FAR and WRA Form26 databases and for comparing birthdates from the databases with dates from the index cards; and (3) *Troubleshooting and Debugging* – A Jupyter Notebook was used to support debugging of the Python program for recognizing PII.

VI. Developing Name Registries

Student team: Andy SILVA (Info.Sci), Emery PATTERSON (MLIS), Mary McKINLEY (MLIS) [team leader].

A. Digital Curation Significance

Name registries relate to authority records in libraries by providing standardized forms of names or terms. There are a number of very significant Name Registries initiatives, including: (1) *The Library of Congress Authorities*, a service used by librarians that specifies authoritative forms of names (for persons, places, meetings, and organizations), towards assisting in finding materials ((https://authorities.loc.gov); (2) *The ISAAR (CPF) standard,* "International Standard Archival Authority Record for Corporate Bodies, Persons and Families", which provides guidance on the creation of elements of an authority record, including a name as an authorized, parallel, or alternate form, the goal being the linking of these archival authority records to archival materials and other archival resources; (3) *Social Networks and Archival Context (SNAC),* an online resource that helps users discover biographical and historical information about persons, families, and organizations that created or are documented in historical resources (primary source documents) and their connections to one another. and (4) *Yad Vashem Names Database,* with its Central Database of Shoah Victims' Names (or Names Database) with a primary purpose of helping to recover the names and reconstruct the life stories of each individual Jew murdered in the Shoah. In addition, the Documents Archive also contains lists of residents, deportees, victims, war criminals and collaborators.

B. Case Study Connections

In the records for the Japanese-American internment camps, the names registry needs to be the authority file for the individuals in the camp across the various kinds of records, so that each individual person can be reliably identified and traced through all of their experiences. To accomplish this goal, it is necessary to find a way to match records across the two mass record types that every person in the camps would have had, the WRA Form26 and FAR. The challenge for this process is that individuals were not assigned unique identifiers, so there is no single field that can be used to connect the records from the two sets. In order to address this challenge, Densho and the DCIC have created a Jupyter Notebook Python script that attempts to identify matching records through different matching strategies. The first and most successful matching strategy has been Family Group Number and Birth Year. However, because historical data is never perfect, even after data cleaning, more than one strategy is necessary to approach matching the records.

The basic protocol for identifying possible matching schema is fairly simple, but entirely unique for the record sets in question: first, identify each piece of information or field that occurs in both datasets, preferably in the same (or easily transformed) formats. Then identify which combinations are likely to return one individual consistently. For this project, Family Number-Birth Year has been the most successful pairing, but other options include listed names (First, Last, and Other)-Birth Year and Family Number-listed names. Another combination that may be particularly useful for this project in particular subsets is Pre-Evacuation Town and Birth Year.

Students designed a divide-and-conquer strategy each adopting a subset of the FAR records based on the camp entry code status. This group's deliverables will positively impact the productivity of all four other groups.

C. Computational Thinking Patterns

These computational activities were matched to descriptions of the 22 computational thinking practices. It was found that these activities correspond to thirteen of the computational thinking Practices. See **Figure 5**.

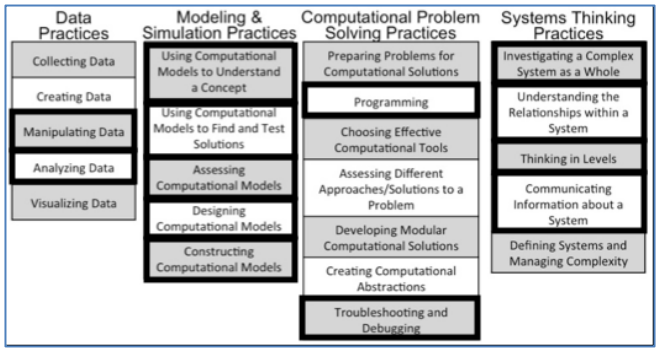


Figure 5: Thirteen CT matches.

The thirteen relevant CT Practices include:

- *Two **Data Practices***: (1) *Manipulating Data – Open Refine* was used to clean and normalize the data; and (2) *Analyzing Data* through facets and joins where we discover the points of connection between datasets.
- *Five **Modeling & Simulations Practices***: where we (1) *Use* an existing computational Jupyter Notebook from Densho to Understand how to join existing files, (2) *Test* the joins, (3) *Assess* computational approaches, (4) *Design* new matching strategies, and (5) *Construct* new models.
- *Two **Computational Problem Solving Practices***: (1) *Programming* and (2) *Debugging* where the computational Python Jupyter Notebook was enhanced, re-run and debugged.
- *Four **Systems Thinking Practices**:* (1) *Investigating* a complex matching system, (2) *Understanding* the relationships within this system, (3) *Thinking in Levels* through the splitting of the FAR into 26 individual files based on the camp entry code status, and (4) *Communicating* the information with Densho.

VII. Integrating Vital Records

Student team: James SANTOS (MLIS/HiLS), Genevieve KOCIENDA (MLIS) [team leader], Kanish JAIN (MIM).

A. Digital Curation Significance

Vital records typically comprise birth, marriage, and death records. They are normally produced by local authorities (counties and cities) and not the federal government (with the exception of military records). What NARA calls vital records refers to something very different: records disaster mitigation and recovery outside of normal operations conditions. We focus on the former local definition.

B. Case Study Connections

The vital records pertaining to the camps, and those of Tule Lake camp in particular, provide a way to draw relationships between the conditions in the Tule Lake Camp and the resistance that occurred there. By analyzing the vital records and comparing them to what is already public knowledge about Tule Lake, we can tell a more complete story of the experiences of the people incarcerated there. More specifically, we datafied and integrated records relating to death, casualty, and disease, as they provide a direct link to specific events at Tule Lake. The following is a list of the records most relevant to our purpose:

- The record that compares the death rates by various causes against the rates of US residents as a whole.
- A list of demands from the Tule Lake prisoners illustrating poor sanitation, hospital conditions, and food scarcity.
- Number codes used to identify diseases and causes of death that can be used to cross reference other vital records.
- Transcripts of conversations between Tule Lake authorities and prisoner representatives relating to camp conditions, requests for funeral rites and delivery of ashes of family members outside of Tule Lake.
- Camp authority's official report on the events leading up to acts of resistance, and the brutal crackdown from government authorities.
- Morgue, Cremation, Retention of Ashes, FAR, and Incident Card records were datafied, compared, and integrated into a unified dataset. A pattern emerged whereby between June 1942 and September 1943 death records "fall between the cracks".

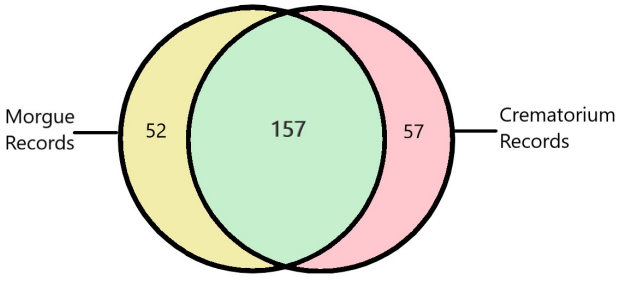


Figure 6: Integrating Morgue and Crematorium records.

C. Computational Thinking Patterns

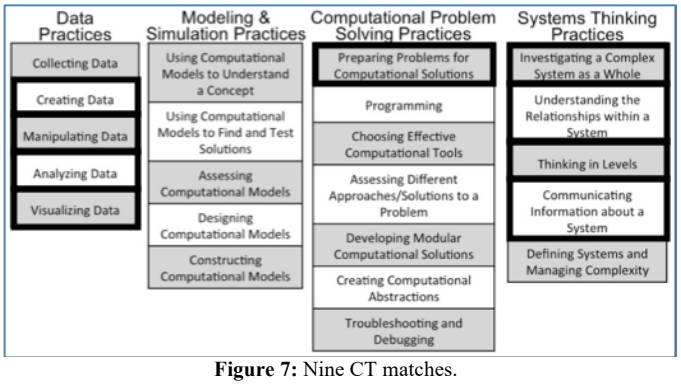


Figure 7: Nine CT matches.

The nine relevant CT Practices include:

- *Four **Data Practices***: (1) *Creating Data:* by generating new integrated datasets; (2) *Manipulating Data* by cleaning datasets in *Open Refine*; (3) *Analyzing Data* by looking at patterns and signals across the Death records; and (4) *Visualizing Data* through charts and graphs.
- *One **Computational Problem Solving Practices***: (1) *Preparing Problems for Computational Solutions* datafying records in preparation for computational treatments.
- *Four **Systems Thinking Practice:*** (1) *Investigating a Complex System as a Whole* in order to build a model that captures casualties in the Camps; (2) *Understanding the Relationships within a System* in order to integrate components; (3) *Thinking in Levels* in order to break down casualty patterns across the Camps; and (4) *Communicating* Information about a System to document the complexities.

VIII. Designing Controlled Vocabularies

Student team: Tahura TURABI (Info.Sci & Japanese), Margaret Rose HUNT (MLIS) [team leader], Hannah FRISCH (MLIS), and Hilary Szu Yin SHIUE (MLIS).

A. Digital Curation Significance

In the 2013 Getty publication *"Introduction to Controlled Vocabularies"* [7], a controlled vocabulary is defined as "an organized arrangement of words and phrases used to index content and/or to retrieve content through browsing or searching. It typically includes preferred and variant terms and has a defined scope or describes a specific domain." Types of controlled vocabularies include: (1) relationships, (2) subject heading lists, (3) controlled lists, (4) synonym ring lists, (5) authority files, (6) taxonomies, (7) alphanumeric classification schemes, (8), thesauri, (9) ontologies, and (10) folksonomies.

While controlled vocabularies need not be hierarchical, taxonomies are often organized into a hierarchical structure, with sub-categories of the search term. Ontologies are a more "formal, machine-readable specification of a conceptual model in which concepts, properties, relationships, functions, constraints, and axioms are all explicitly defined". They support grammars and allow for the formulation of assertions.

B. Case Study Connections

The student team tackled the vexing problem of standardizing the incident classification category on the Incident Index Cards, through the design and testing of an overall multilevel controlled vocabulary or taxonomy. The team used *Open Refine* to cluster and clean, in order to harmonize related incident card category terms. Use of the R-scripting language also helped automate the splitting of the larger spreadsheet to facilitate group work. An agile methodology was pursued where students adopted and cleaned subsets of the collection in order to validate an initial design of the taxonomy and iterate over the design based on non-matching term concepts. The *Tableau* software was also used to visualize the size and balance in resulting categories and subcategories. The team is looking at automating the mapping from initial card categories to taxonomy categories once an initial taxonomy design has converged, particularly on the administrative subset of the data. This represents a novel contribution of computational methods on human-created data, towards creating machine-readable representations.

C. Computational Thinking Patterns

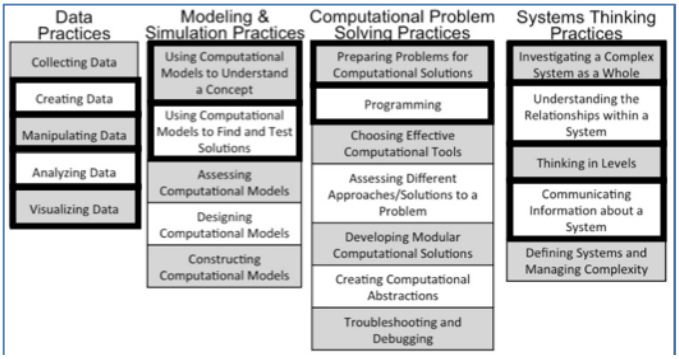


Figure 8: Twelve CT matches.

The twelve relevant CT Practices include:

- *Four **Data Practices***: (1) *Creating Data* by generating new taxonomy categories; (2) *Manipulating Data* by classifying and cleaning index card categories in *Open Refine*; (3) *Analyzing Data* through facets; and (4) *Visualizing Data* through the use of the *Tableau* software.
- *Two **Modeling & Simulations Practices***: (1) *Using Computational Models to Understand a Concept* by running clustering algorithms to gauge term similarity, and (2) *Using Computational Models to Find and Test Solutions*

by reducing the complexity of terms through algorithmic reduction.

- *Two* ***Computational Problem Solving Practices***: (1) *Preparing Problems for Computational Solutions by* datafying records in preparation for computational treatments, and (2) *Programming* using R scripts to split the data.
- *Four* ***Systems Thinking Practices****:* (1) *Investigating a Complex System as a Whole* in order to build a model that captures casualties in the Camps; (2) *Understanding the Relationships within a System* in order to integrate components; (3) *Thinking in Levels* in order to break down casualty patterns across the Camps; and (4) *Communicating Information about a System* to document the complexities.

IX. Mapping Events and People

Student team: Connor MULLANE (MLIS) [team leader], Britton SCHAMS (MLIS), Mirielle VASSELLI (MLIS), Chenxi LIU (HCIM), and Jiale XU (HCIM).

A. Digital Curation Significance

Spatial thinking and analysis are inherently computational in nature: "Spatial thinking involves developing concepts of space that involve relations and calculations of distance, using tools of representation like maps and graphs, engaging in the process of reasoning to organize and solve problems, and thinking in levels of detail." [8] In 2007, a group of historians participated in a two-week workshop on Geographies of the Holocaust to determine: the potential benefits of applying geographic methods, such as spatial analysis and visualization to the study of concentration camps, and the extent to which debates in human and cultural geography about 'placing the past' could be applied to the camps.[3] These efforts represent an important area of digital curation that deserves to be incorporated into archival science.

B. Case Study Connections

Understanding the spatial dynamics and relationships of location and proximity in the Camps can offer unique insights into the nature of oppression and resistance. Using open source tools like *Open Refine* and mapping tools such as *QGIS*, the team was able to research and map significant narratives of resistance:

- A *Negotiating Committee* of 17 men attended a meeting with Tule Lake Camp Director, Raymond Best, and WRA Director, Dillon S. Myer, on November 1st, 1943, in order to try to find solutions to problems at Tule Lake, including the firing of farm workers without warning, the death of a truck driver, and the ensuing tension at his public funeral. Our initial investigations map the locations of the Committee members in order to spatially visualize the extent of the leadership network in the camp, and assess the geographical footprint of this committee.

Figure 9: Nov. 1, 1943 Negotiating Committee building locations at Tule Lake.

Other examples of spatial events under consideration include:

- *"6 women staging a sit-down strike at Gate #3 demanding to see their husbands in stockade"* at Tule Lake. Who were these women? Were they related to resistance leaders? These are some of the questions we are hoping to uncover by mapping their locations of protest as well as their building within the camp. Exploring spatial patterns of resistance in Tule Lake is of interest.
- *"7 men attempting to escape from the camp by crawling under the perimeter fence by the Canal between Towers 12 and 13", o*n September 8, 1945.

[3] Geographies of the Holocaust, see: https://www.ushmm.org/learn/mapping-initiatives/geographies-of-the-holocaust/

C. Computational Thinking Patterns

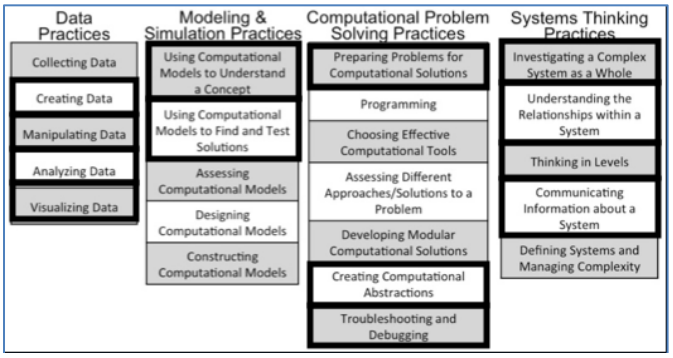


Figure 10: Thirteen CT matches.

The thirteen relevant CT Practices include:

- *Four **Data Practices***: (1) *Creating Data:* by creating an interactive clickable map of the Tule Lake Camp with every structure listed, harvest spatial data from RG210 records in the form of names of locations in the camp, and generate the corresponding lat/lon coordinates (this involves geospatial techniques of georeferencing, vectorization, geolocation, and the use of computational geoprocessing tools); (2) *Manipulating Data* by running Spatial Querying filtering queries; (3) *Analyzing Data* by testing geospatial hypotheses; and (4) *Visualizing Data* through the use of mapping software.
- *Two **Modeling & Simulations Practices***: (1) *Using Computational Models to Understand a Concept* by formulating spatial questions, and (2) *Using Computational Models to Find and Test Solutions* by using a GIS System to map ideas.
- *Three **Computational Problem Solving Practices***: (1) *Preparing Problems for Computational Solutions* by using *Open Refine* to associate lat/lon coordinates with named camp locations (in a GeoGazeteer fashion), (2) *Creating Computational Abstractions* by generating name objects with computed coordinates, and (3) *Troubleshooting and Debugging* to incrementally test spatial hypotheses.
- *Four **Systems Thinking Practices**:* (1) *Investigating a Complex System as a Whole* to speculate about the spatial dynamics of a Camp, (2) *Understanding the Relationships within a System* between people, events, dates, and places, (3) *Thinking in Levels* by decomposing complex spatial dynamics into smaller mappable problem, and (4) *Communicating* the information with stakeholders.

X. Connecting Events and People through Networks

Student team: Sarah AGARRAT (Undergrad CS/Statistics) [team leader], Hannah KRAUSS (MLIS), Danish MIR (MIM), Mayhah SURI (MIM), and Debashish PRADHAN (HCIM).

A. Digital Curation Significance

Graphs and networks are mechanisms to represent connections between people, places, objects, and events. These are human social networks. Graph databases are emerging as part of the NoSQL highly-scalable distributed landscape, with examples such as Neo4j. "Graph databases offer a new approach that supports deep and rich investigation of data, and they seem a natural fit to research-led archival integration." [9] They are used as part of the EHRI project (European Holocaust Research Infrastructure).

B. Case Study Connections

Our initial focus has been on identifying "networks of resistance" in Tule Lake. The team identified human social networks such as "*6 women staging a sit-down strike at Gate #3", "Negotiating Committee of 17 men on Nov. 1, 1943*", "*7 men attempting to escape from the camp"* (referenced in the previous section) and explored the use of graph database tools to datafy and then visualize these networks. Other examples of social networks under study include:

- Co-signers of a petition to release people held in the stockade.
- Transports out of Tule Lake to Santa Fe and Bismarck N.D.
- Deportations to Japan and Hawaii.

The following graph created in Neo4j connects people nodes (in red) incarcerated at Tule Lake (orange node) based on their participation in acts of resistance such as being "apprehended for marching and wearing insignia on June 25, 1945" (in grey) and being subsequently transferred to Bismarck, N.D. on July 3, 1945. As shown these graphs begin to connect people into webs of events that show important relationships and allow filtering and querying through powerful graph query languages.

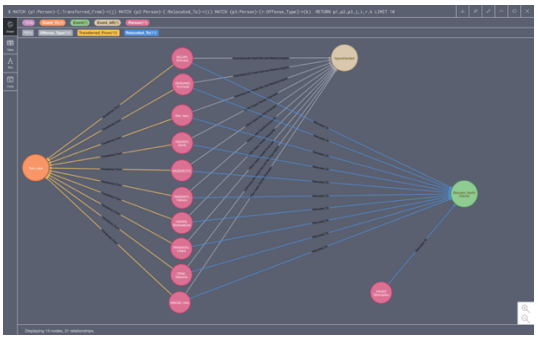


Figure 11: Human social network of people arrested in the Tule Lake Camp for protesting and subsequently transferred to Bismarck N.D.

C. Computational Thinking Patterns

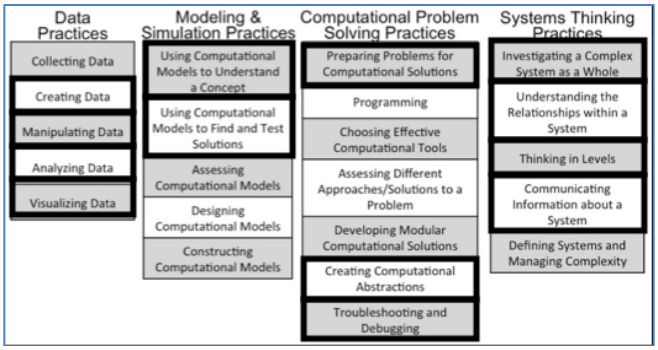


Figure 12: Thirteen CT matches.

In the same fashion as the *Mapping People and Events* team, we identified thirteen relevant CT Practices include:

- *Four **Data Practices***: (1) *Creating Data:* by creating property graphs consisting of nodes, edges, and properties; (2) *Manipulating Data* Graph Queries; (3) *Analyzing Data* by running Graph Queries; and (4) *Visualizing Data* through the Graph Visualizing toolkits.
- *Two **Modeling & Simulations Practices***: (1) *Using Computational Models to Understand a Concept* by formulating graph questions, and (2) *Using Computational Models to Find and Test Solutions* the use of Neo4j.
- *Three **Computational Problem Solving Practices***: (1) *Preparing Problems for Computational Solutions* by loading spreadsheets in bulk into Neo4j, (2) *Creating Computational Abstractions* by generating graph networks, and (3) *Troubleshooting and Debugging* to incrementally test graph hypotheses.
- *Four **Systems Thinking Practices**:* (1) *Investigating a Complex System as a Whole* to speculate about how people and events are connected in the Camp, (2) *Understanding the Relationships within a System* between people, events, dates, and places, (3) *Thinking in Levels* by decomposing complex networks into smaller ones, and (4) *Communicating* the information with stakeholders.

XI. Conclusions and Future Work

This wide-ranging digital archiving project is an important step in mapping established computational thinking practices to archival science education and research. The project has utilized a multi-pronged approach, creating opportunities to develop datafied archival records including taxonomies and definitions, discover narratives and connections, and visualize and map places, people and events. These approaches incorporate data collection, modeling and simulation, computational problem-solving, and systems thinking. This project serves as a model for students and researchers designing digital curation projects from primary, non-digitized records.

Future work includes: (1) *Name Registries* – developing deeper matching strategies; (2) *Integrating Vital Records* – resolving discrepancies in the way death is reported; (3) *Controlled Vocabularies* – automating the classification of the Incident Card categories into the proposed taxonomy; (4) *Mapping Events and People* – developing a spatio-temporal model; and (5) *Connecting Events and People through Networks* – exploring network analysis using graph algorithms.

References

[1] R. Marciano, V. Lemieux, M. Hedges, M. Esteva, W. Underwood, M. Kurtz, and M. Conrad (2018). 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* (Advances in Librarianship, Volume 44B, pp.179-199) . Eds: J. Percell, L. C. Sarin, P. T. Jaeger, J. C. Bertot. Emerald Publishing Limited. See:http://dcicblog.umd.edu/cas/wp-content/uploads/sites/13/2017/06/Marciano-et-al-Archival-Records-and-Training-in-the-Age-of-Big-Data-final.pdf

[2] D. Weintrop, E. Beheshti, M. Horn, K. Orton, K. Jona, L. Trouille, & U. Wilensky (2016). Defining Computational Thinking for Mathematics and Science Classrooms. Journal of Science Education and Technology, 25(1), 127–147. See: https://www.terpconnect.umd.edu/~weintrop/papers/WeintropEtAl_2015_DefiningCT.pdf

[3] J. Wing (2006). Computational thinking. *Communications of the ACM*, *49*(3), 33–35. Retrieved from https://www.cs.cmu.edu/~15110-s13/Wing06-ct.pdf

[4] W. Underwood, D. Weintrop, M. Kurtz, and R. Marciano, Introducing Computational Thinking in Archival Science Education, IEEE Big Data 2018, Computational Archival Science (CAS) workshop #3, Seattle, Dec. 12, 2018. See: https://dcicblog.umd.edu/cas/wp-content/uploads/sites/13/2018/12/1.Underwood.pdf.

[5] R Marciano, et al. Automating the Detection of Personally Identifiable Information (PII) in Japanese-American WWII Incarceration Camps. Computational Archival Science Workshop #3 , IEEE International Conference on Big Data 2018, Dec. 12, 2018. See: https://dcicblog.umd.edu/cas/wp-content/uploads/sites/13/2018/12/2.Marciano.pdf

[6] W. Underwood, R. Marciano, Computational Thinking in Archival Science Research and Education, IEEE Big Data 2019, Computational Archival Science (CAS) workshop #4, Los Angeles, Dec. 11, 2019.

[7] Patricia Harpring, Series edited by Murtha Baca, Introduction to Controlled Vocabularies: Terminology for Art, Architecture, and Other Cultural Works, 2013 Getty Publications. See: https://www.getty.edu/research/publications/electronic_publications/intro_controlled_vocab.

[8] S. Rebich Hespanha, F. Goodchild & D. Janelle (2009) 'Spatial Thinking and Technologies in the Undergraduate Social Science Classroom', Journal of Geography in Higher Education, 33: 1, S17-S27. See: https://www.nceas.ucsb.edu/~hespanha/srh/Publications_files/2009RebichHespanha_Goodchild%26Janelle.pdf

[9] T. Blanke, M. Bryant, R. Speck, Developing the Collection Graph, Library Hi Tech, 2015.