Computational Treatments to Recover Erased Heritage: A Legacy of Slavery Case Study (CT-LoS)

Lori A. Perine
College of Information Studies
University of Maryland
College Park, USA
lperine@umd.edu

Rajesh Kumar Gnanasekaran
College of Information Studies
University of Maryland
College Park, USA
rgnanase@umd.edu

Phillip Nicholas
College of Information Studies
University of Maryland
College Park, USA
pnichol2@terpmail.umd.edu

Alexis Hill
College of Information Studies
University of Maryland
College Park, USA
ahill126@terpmail.umd.edu

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

Abstract—Graduate students at the University of Maryland’s College of Information Studies (UMD iSchool) collaborated in interdisciplinary teams on a case study to explore application of computational methodologies to datafied collections related to slavery in the Maryland State Archives (MSA). Two research questions were examined: (1) What are the opportunities and limitations for using computational methods and open source tools to characterize data encoded within records of enslavement and to discover new patterns and relationships in that data? (2) How does knowledge of social and cultural systems impact those opportunities and limitations? Computational methods and tools were most effectively used when socio-cultural contextualization and technology’s role as a mediator of representation were taken into account. Three additional technical research areas are identified to enhance recovery of heritage hidden in records of enslavement: visualization, graph databases, and ontologies and metadata.

Keywords—Computational Thinking, Digital Curation, Computational Archival Science (CAS), Legacy of Slavery.

I. INTRODUCTION

Graduate students from the Master of Library and Information Science (MLIS) and doctoral programs of the University of Maryland’s College of Information Studies (UMD iSchool) collaborated in interdisciplinary teams on a project to explore application of computational methodologies to datafied collections in the Maryland State Archives (MSA). The project teams, which included the authors of this paper, had diverse backgrounds in IT, information policy, information management, history, art history, and archives.

The collections that we used were selected from the MSA’s Legacy of Slavery (LoS) collections. The LoS collections were created during a major initiative launched in 2001 to document the lives and experiences of enslaved and free Black populations in Maryland from 1634 to 1867 [1]. MSA staff members and volunteers examined various records primarily located in MSA’s repository: case studies, runaway slave advertisements, census data, court records, and published materials from pre-Civil War Maryland relating to slavery. The MSA team identified approximately 420,000 individuals from those records. As the initiative progressed, the original resources from chattel records, census records, slave statistics, manumissions, and certificates of freedom were partially transcribed, mostly by volunteers. They digitized, extracted and assembled the transcribed information into 16 separate, unconnected SQL databases [2].

In undertaking this process, the MSA team explicitly intended to repurpose original source records, so that information about the complexity and vitality of ante-bellum Black populations in Maryland could be better understood. A long term objective is to make these collections openly accessible to the general public.

Our collaboration advances the MSA objectives to tell the story of enslavement and freedom and to promote the contributions, influences, and triumphs of Black populations in Maryland. It provides foundational understanding about (1) how computational methods could enhance engagement with information contained in the LoS archival records; and (2) how computational methods might help recover erased heritage contained in the relationships and stories that heretofore have been hidden in the records.

II. RESEARCH PROBLEM STATEMENT

This paper presents a student-driven, exploratory case study to apply computational methods associated with “big data” to information contained within text-based records from selected LoS collections held by the MSA. The case study explores two primary questions:


2 In this paper, we use the term Black to refer to a person of Sub-Saharan African descent, regardless of classification in primary source documents. As will be described below, a person of African descent could be classified as slave, free, born free, and several other variations.
1. What are the opportunities and limitations for using computational methods and open source tools to characterize data encoded within records of enslavement and to discover new patterns and relationships in that data?

2. How does knowledge of social and cultural systems impact those opportunities and limitations?

The computational thinking taxonomy in Fig. 1 is used as a framework for inquiry. Our computational explorations used open source tools, many of which can be applied with minimum knowledge of programming. We also consider how innovative technologies might enable representations of source materials that mitigate erasures in archival records. In this case, where we are using datasets related to enslavement, an explicit objective is to explore how to make more visible the heritage of individuals who often are not represented in the historical record.

The multidisciplinary nature of our team was a key enabler of our exploration. While not the focus of this paper, we note its importance, which is examined by members of our team in a forthcoming paper for the Society of American Archivists (SAA) [3].

III. LITERATURE SURVEY

Computational thinking (CT), the analytic framework for this case study, is an approach to thinking about and engaging with intellectually challenging problems, at multiple levels of abstraction. It is based on a set of competencies originally articulated by Wing [4] and comprises use of modeling, decomposition, pattern recognition, abstraction, algorithm design, and scale. Weintrop et al. [5] developed a formal taxonomy for CT, with application to mathematics and science education (Fig. 1). The taxonomy disaggregates CT into a set of computational practices organized along four themes: (1) data, (2) modeling and simulation, (3) computational problem solving, and (4) systems thinking.

CT is a core element of the emerging, interdisciplinary field of computational archival science (CAS). CAS explores the application of computationally based methods and technologies associated with “big data” to archival and record management practices. Multi-disciplinary collaborations partner researchers and practitioners of artificial intelligence (AI), machine learning, and visualization with archivists, historians, and information management professionals. Technologies are leveraged to discover and increase accessibility to information that is buried within massive stores of records [6]. Proponents active in the development of CAS note that emerging technologies already “have profoundly altered the nature of archives, by disrupting how information is created, recorded, captured, encoded, curated, shared, made available and used” [7]. The application of CT along with specific computational methods opens new possibilities for research, discovery, and knowledge creation with archival records.

Marciano et. al [8] propose the following working definition for CAS:

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. The intent is to engage and undertake research with archival materials, as well as apply the collective knowledge of computer and archival science to understand the ways that new technologies change the generation, use, storage and preservation of records, and the implications of these changes for archival functions and the societal and organizational use and preservation of authentic digital records. The suggests that computational archival science is a blend of computational thinking and archival thinking.

Researchers from the Advanced Information Collaboratory (AIC) have adapted the CT taxonomy for graduate-level curriculum in archival sciences [9, 10]. The adaptation maps CT’s computational practices to the fundamental concepts and techniques specified by the Society of American Archivists’ Guidelines for a Graduate Program in Archival Studies (GPAS) Curriculum [11]. For example, data practices in the CT taxonomy such as visualization and data analytics are mapped to the knowledge area of Social and Cultural Systems [2]. This knowledge area promotes skills and techniques that help students contextualize archival artifacts, through understanding the social and cultural systems, public and private institutions, as well as political, social and economic dynamics in which records are created, maintained, and used [11].

Although curriculum is not the focus of our paper, these adaptations of CT provide important conceptual scaffolding and domain-specific references for our case study. CT themes and their elements are conceptualized as discrete in the Weintrop taxonomy, whereas in practice we often found ourselves applying practices simultaneously and somewhat iteratively. We make note of this in our discussion, but leave to further exploration whether and how the domain of application...
influences the sequencing of CT elements. Similarly, the Underwood adaptation to archival science curriculum is useful for the framing and terminology that it provides for CT in the domain of archival collections and practices. We reference some of these connections in our discussion, but use the Weintrop taxonomy to create general mappings of the computational methods that we deployed.

Given that we were working with collections related to the enslavement of a Black population in the United States, awareness of representation was an important consideration in our exploration. The text-based digitized records that we used from the LoS collections are derived from highly heterogeneous primary source documents that were created in a social and legal context where key participants in the events being documented (enslaved and free Black individuals) are obscured or inadequately identified.

Archivists are re-examining general practices to better represent persons and events that were marginalized or “erased” from historical records [12, 13, 14]. The use of digitization and related technologies can amplify marginalization or erasure, whether through limitations in the original source documents or limitations within the technologies. The latter may unintentionally obscure visual and written features, especially those related to race, gender, and/or class. Williams [15] offers an art historical perspective on this phenomenon, demonstrating that technical limitations can lead to erasure and distortion of archival records involving underrepresented and/or marginalized communities.

While Williams’ work focuses on visual representations in archives, her insights about technology as a representational tool have broad value within CAS. In particular, she brings attention to the opportunity for CAS to be conscious of its role as a “mediator” in bringing representations of source materials to the user. If used intentionally, this role situates CAS with a potential to reverse rather than amplify erasures. It is this potential towards which we take initial exploratory steps in this case study.

To our knowledge, there has been no prior work that directly addresses how computational methods might help uncover hidden knowledge and/or mitigate erasures within archival collections related to enslavement. Research tends to focus on applications for archival management, such as visualizations to assist with user access in retrieving archival records [16,17]. Specific to collections related to enslavement, there are several recent research initiatives focused on developing ontologies for collections related to the transatlantic slave trade, with the objective of integrating data from diverse sources [18, 19]. Adaptation or extension of such ontologies for the LoS collections are among our recommendations for future research.

IV. COMPUTATIONAL THINKING IN THE DATA WORKFLOW

The explorations in this case study were organized around two datasets from the MSA collections: Manumissions and Certificates of Freedom. A manumission is a legal document granting freedom to an enslaved person. A Certificate of Freedom (CoF) is a legal document issued from 1803 to 1865 in the state of Maryland to Black persons, who were required to record proof of their free or emancipated status in the county court. The details about these types of records and the type of data found in each are given in sections V and VI, respectively. There is a natural connection between the two collections, as persons who were manumitted and remained in the state would have been required to obtain a CoF. This was the basis of our cross-collection exploration in section VII.

The iSchool teams worked with data from the MSA’s SQL databases, extracted and subsetted into comma-separated values (csv) files for easy access. The data flow pipeline (Fig. 2) built on the MSA research team’s prior work of extracting information from original source documents and storing it in SQL databases [2]. The data engineering area, represented in blue produced a series of unconnected relational SQL databases associated with each series of archival records, and which can be queried via web-based interfaces on the MSA website [1]. The Data Analytics section in red highlights the computational explorations by our iSchool teams.

Fig. 2: Data Workflow for Computational Exploration

Our teams adopted a generic workflow common to computational exploration and analysis of datasets:

1. Data sourcing from the MSA’s SQL databases.
2. Data cleaning and wrangling.
3. Data filtering and extraction.
4. Exploratory analytics.
5. Data visualization.

Within that workflow, the teams made exploratory use of various open source computational tools for data wrangling, data visualization, and networking of events and people represented in the records.

This process engaged the application of three CT practice areas: data practices, computational problem solving practices, and systems thinking practices. The practices were applied in an integrated and iterative fashion throughout five basic steps of our data workflow.

Data Practices include skills for collecting, creating, analyzing, and sharing data are at the foundation of this data practices CT practice area [5]. These practices have archival science correspondences in many areas classified as Knowledge of Archival Materials and Functions in the SAA’s
GPAS [11]. This involves technology- and data-enable features of data (e.g. metadata), faceted search techniques, and use of forensic tools to assist appraisal. A more complete mapping is given in [9]. The data workflow for our project was the main instantiation of the CT data practices theme.

**Computational Problem Solving Practices** involve skills for assessing, choosing, preparing, and implementing computational tools to explore or solve a problem [5]. Our computational problem solving practices included developing the data workflow, assessing approaches for each step of the workflow, and selecting from among open-source computational tools for implementation. We provide more details of this approach in sections V and VI below.

**Systems Thinking Practices** focus on inclusive examination of the dynamics and relationships within the system and among its constituent parts [5]. In the context of SAA’s GPAS, this translates to skills and techniques that help students contextualize archival artifacts, through understanding the social and cultural systems, public and private institutions, as well as political, social and economic dynamics in which records are created, maintained, and used [11]. Engagement with systems thinking practices had both technical and socio-cultural dimensions for our project, as described below.

**Socio-cultural contextualization as a component of systems thinking** proved to be a core element of our approach. Almost from the start of the project, we recognized that many of our technical decisions would be dependent upon the context of the datafied MSA archival collections. That context had two components: first, the socio-cultural context in which the original source artifacts and second were created and collected; and second, the socio-technical context for converting those artifacts into digital formats. Both impacted almost every aspect of our data workflow. Thus, understanding the “biography” of the datasets we were using was integrated into our systems thinking.

The information available in original source documents for manumissions and CoF records were generated in specific cultural and legal contexts that codified many of the relationships and events involving Black people in pre-Civil War Maryland. As in other slaveholding states, Maryland’s laws of slavery were primarily laws of property [20]. Thus, those enslaved often, by law and/or custom, are not well identified in archival documents, even in those related to their own emancipation. One might argue that Black individuals, whether classified as enslaved or free, would be the primary subjects of these records. However, the structure of both law and social customs relegates these persons to a secondary and sometimes anonymous status in many original records. Records for both manumissions and CoFs are created from the perspective of the legal and social status of the slaveholder. Minor differences in custom and law from county to county added an additional layer of complexity.

These were important points to consider as we looked at how to apply computational tools and methods to make visible all the key individuals related to the events encoded in the datasets, especially the Black people who were “hidden” in the records. Whereas the provenance of the original documents is a key topic of archival science practice [9], we found that the legal, historical, and even geographic context of collecting and creating original source documents had implications in our applications of computational tools and methods. This is discussed in greater detail in sections V and VI.

Similarly, we found that the socio-cultural context of the original source documents influenced the digitization of the collections. Transcription errors due to visual translation of 18th and 19th century script, were common. As we encountered those errors, we also recognized that volunteer transcribers may not have been sufficiently familiar with the legal and historical context of the source materials, and thus were introducing additional errors of misinterpretation as they classified information that would be assigned to database fields.

As with the context of the original source documents, the context for datifying those documents had implications for how we applied computational tools, and how far we were able to take a frame of inquiry that included all key individuals. We recognized limitations in the technology and tools used to create the relational SQL databases that contained the digitized collection. On the one hand, technology has evolved over the two decades since the MSA initiated the LoS project to allow for more finely-tuned and better automated capture of information from documents. On the other hand, the decisions and/or assumptions made about defining data fields and about what information to “mine” from the source documents implicitly reflect the legal and social structure of the original documents. We often discovered copious information about other individuals and related events in the notes field of the datasets. This rich information could be used to further explore and characterize the roles of individuals involved in emancipation “events”, but was not captured in the main fields of the SQL databases from which we sourced our data.

The decisions that had been made by prior transcribers and digitizers about how to define data fields and what information to “mine” from the original documents reflect both the inherent limitations of relational databases, as well as archival practices for encoding information in the source documents.

Some key questions that arose in our deliberations included: *How well are meaning and relationships captured by the digitized encoding? What data elements are lost or ignored, and why? Who decides what is included and according to what criteria?* We were not necessarily able to answer these questions directly. However, reflection on the questions permitted us to consider computational solutions with a greater awareness of our “mediator” role in using technology to extract and represent information contained in the LoS collections.
V. COLLECTION 1: MANUMISSIONS

Team 1: Left to right: Greg JANSEN, Lori PERINE (Doctoral), Rongqian MA (Doctoral at U. Pittsburgh), Dan YANG (MIM), Anuja WANI (MIM), Dr. Lyneise WILLIAMS (UNC Chapel Hill & VERA Collaborative)

A. “Data Background” / Contextualization

A manumission document is the legal document emancipating an enslaved person. Original source documents for manumissions are highly heterogeneous and can be found in land, probate, and chattel records [21]. The documents typically contain information such as name and residence of the slaveholder, name and age of the enslaved individual, and the date the enslaved individual is to be freed. In some cases, the date a slave was to be freed is different than the date the manumission document was recorded. Additional information may be available, depending on the nature of the document.

In 1796, the Maryland General Assembly passed acts that provided more consistent pathways for liberating enslaved persons [22]. Thus documentation of manumissions became more regularized up through November 1, 1864, the date of universal emancipation in the state of Maryland. The format and source of documents were still varied, with somewhat standard formats appearing in a few counties.

A peculiarity found in manumissions documents which impacted our data practices was multiple manumissions dates. The multiple dates, some decades apart, were due to the custom of delayed manumissions. Slaveholders were able to file legal documents granting freedom to an enslaved person contingent upon a future event: the person had to reach a certain age, be conscripted into military service, emigrate to Africa in the recolonization program developed by major slave holders and ultimately funded by the state legislature [23], or remain in bondage for an arbitrarily determined number of years. When transcribed into the SQL databases, this resulted in records with multiple date fields.

Once cleaned, the manumissions dataset contained 7,399 records geographically associated with 10 of Maryland’s 24 counties. The records spanned the years 1770 to 1870. Individuals names in the data fields encoding information about enslaved persons were ages 0 (infant) to 80 years of age. The records were somewhat more balanced with respect to gender, in comparison with the predominantly male CoF records: about 48% male and 52% female.

B. Computational Tools and Methods

The manumissions team worked primarily with the open source R programming software [24] to wrangle, filter and extract, explore, and visualize the dataset. This made it easier to use customized programming to handle data inconsistencies, such as the multiple date fields. We were able to automate, document, and save different analytic strategies for reuse by creating R functions and scripts for key tasks.

The primary challenges with the manumissions data arose from its lack of standardization. The original records came from many resources and with inconsistent, non-standardized information captured in those records. As the information was transcribed, decisions were made to fit into SQL database fields that sometimes seem redundant, and other times are incomplete.

Thus we began by extracting and modifying the classes of data features that would allow us to characterize the dataset, seek information that were most likely to be common to other collections (for cross-collection matching), and those which might provide key information of interest for identifying both the individuals involved in the manumissions event (such as names of the slaveholder and enslaved persons, age, sex) and the key attributes of the event itself (county, date).

We then sought to standardize information that appeared erroneous, redundant, or impossible to transform into a standard format. For example, a summary of the Age field yielded a maximum value that clearly was older than a human possibly could be. So we looked for the outlier, which was most likely a transcription error, and changed the Age to NA. Some data features that had been captured for the purposes of archival management were excluded from this particular exploration.

The most challenging item concerned the multiple date fields. The practice of delayed manumissions meant that some records included up to three dates. We had to determine which date to choose before we could perform any temporal data analysis. In some records, there were differences in when a private document may have been signed, versus when it was formally recognized by a legal entity, such as a court or register of wills, while the actual act of manumission would be delayed to a later date. Some dates were recorded only as years, some dates included month and day in addition to year, in various formats.

Sometimes the later date was after 1864, when Maryland abolished slavery. We found an explanation for this apparent discrepancy by examining the notes fields of the records. A slaveholder might authorize freedom at age 30 for a newborn in 1840, for example, resulting in the year 1870 appearing as the emancipation date in the digital transcription of the document. This pattern of emancipation delayed by several decades appeared in several variations throughout the records.

In order to be able to do temporal analysis and make comparisons across the entire dataset, we decided to take the earliest date on the record as the key identifying year for the manumissions event. Although that failed to capture the actual date of emancipation for the enslaved person in many cases, it did allow us to follow trends in the slaveholding population. To implement this decision, we programmed a function to extract the year as characters from each date field, compare the
extracted years for each record, then chose the earliest date. That then became the year used for analysis.

The ggplot 2 package in R was used to create some basic visualizations of the data. We focused on data features that would allow us to represent the stories of the people receiving emancipation.

An example visualization is presented in Fig. 3. We note that while visualization shows that the majority of people manumitted in Maryland’s counties were under the age of 30, that does not necessarily mean that these people saw freedom at an early age. As noted in the discussion above, the practice of delayed manumissions was common. Thus, someone captured in a manumissions document may not physically be freed until years in the future. The visualization captures the age of an enslaved person when the manumission was recorded, but not necessarily when they were freed.

![Age Distribution of Manumitted Persons, by County in the State of Maryland, 1770-1870.](image1)

**Fig. 3:** Age Distribution of Manumitted Persons, by County in the State of Maryland, 1770-1870.

### C. Computational Thinking Practices Identified

The computational activities performed on this collection were matched to descriptions of the 22 computational thinking practices. It was found that the activities correspond to eleven of the computational thinking Practices. See Fig. 4.

The eleven relevant CT Practices include:

1. **Three Data Practices:** (1) Manipulating Data; (2) Analyzing Data; and (3) Visualizing Data. The R

programming software was used to wrangle, filter and extract, explore, and visualize the dataset.

1. **Five Computational Problem Solving Practices:** (1) Preparing Problems for Computational Solutions; (2) Programming; (3) Choosing Effective Computational Tools, (4) Assessing Different Approaches/Solutions to a Problem; and (5) Troubleshooting and Debugging. The raw data was exported as a csv file from the Maryland State Archives data repository and used with R in ways similar to the Certificates of Freedom. In addition, an R Programming model was developed and troubleshooting and debugging were involved.

1. **Three Systems Thinking Practices:** (1) Understanding the Relationships within a System as Complex System as a Whole to understand the processes/mechanisms that have been involved with the issuance of the documents, to understand the thought process behind manually reading and transcribing them into a finite set of digital features,

## VI. COLLECTION 2: CERTIFICATES OF FREEDOM

### A. “Data Background” Contextualization

In the State of Maryland, a “Certificate of Freedom is a legal document that was issued to Black people who were required to record proof of their freedom in the county court” [1]. Without this legal document, an originally free Black person would have been wary of being perceived as an enslaved person. They had to carry this legal document as proof of their freedom, to avoid being misrepresented as an enslaved person. The certificate was issued based on information documented and provided by the former slaveholder and a witness. The court or register of wills would also use manumitting documents for verification.

The Certificates of Freedom (CoF) were handwritten documents containing general biographic, demographic, and descriptive information about a freed person. In most cases, the data captured or documented for the CoF by the courts or court clerk followed a set of standards for documents across different counties which consisted of data features like county issuing the document, slave owner’s first and last name, enslaved person’s first and last name, gender, age, height, complexion, scar (for identification purposes), alias name, date of issue, witness name, prior status of the enslaved person, and a notes feature to enter comments/remarks.

### Team 2:

Left to right: Phillip NICHOLAS (MLIS), Alexis HILL (MLIS), Mark CONRAD, Geneva BLACKWELL (InfoSci) Rajesh Kumar GNANASEKARAN (Doctoral), Sonia PASCUA (Doctoral at Drexel U.)
This rich set of features were captured in an LoS database, which greatly enabled our exploration and visualization within the CoF collection, as well as exploration of connections with other collections of the LoS data repository. Through temporal analysis of comparing certain features like the year of issue of documents, and the time of certain historical events, valuable insights were gleaned.

There are 23,655 records in the CoF dataset from the LoS collection. Some are potential duplicates. The dates of issuance range from 1806 to 1864, covering sixteen Maryland counties and one Maryland city. Freed Black people referenced in the certificates are age 3 months to 82 years. The certificates were issued primarily to males (93%).

B. Computational Tools and Methods

The CoF team used the open-source tools Open Refine [25] and Tableau [26] to explore, clean, extract and visualize the dataset. The dataset was initially uploaded to the Open Refine tool. Using the Facet feature of this tool, individual features were explored for its text, and numerical facets and gleaned important information like the number of groups or unique values per feature if its text based, the distribution range if its number based. This exploration was also useful to identify if any numerical data was captured as a text indicating the need for type conversion.

We immediately identified transcription errors for some features (e.g. age, height) and inconsistent labelling for other features, such as complexion.

For example, the “height” feature was originally transcribed as a text field although it is interpreted as a numeric feature. The Open Refine tool allowed us to create custom expressions to transform a text field into a numeric field through a formula which was used to transform the height text into feet, and inches numeric features.

Working with the labelling in the complexion feature was more complicated. Different labels were applied for the same classification, due to subjective, culturally-based use of language used by white slaveholders and witnesses to describe the complexion of Black people. Each former slaveholder and witness characterized the freed person’s complexion based on their own perceptions, using descriptors such as “dark”, “mulatto”, or “brown”. In all, we noted 42 different labels for their own perceptions, using the language used by white slaveholders and witnesses to describe the complexion of Black people. Each former slaveholder and witness characterized the freed person’s complexion based on their own perceptions, using descriptors such as “dark”, “mulatto”, or “brown”. In all, we noted 42 different labels for the complexion captured in the transcribed data, which were then grouped using Open Refine into 7 categories from Bright (White-passing) to Black (extremely dark skin). We understood that the complexion feature is very subjective, as some of them were compared to materials (wood, copper, etc.) most likely due to the differences in location in terms of county, and there were spelling inconsistencies and some also had long descriptions.

In addition to the traditional options in a toolset like filter and sort, the Open Refine tool also provided additional features which facilitated the data cleaning process. For instance, a feature named “Prior Status”, which indicates the status of the beneficiary of the Certificate of Freedom prior to the issuance of the document, transcribed as a text had many variations of similar status recorded as could be seen below like ‘Free born’, ‘Born free’, with different capitalizations etc. The Open Refine tool’s cluster option helped to group these entries into closely similar groups and allowed us to choose a value to be stored for these similar groups. The groups as identified by the Open Refine is shown below:

![Cluster feature in Open Refine for ‘Prior Status’](image)

Having noted and cleaned the previous transcription errors and inconsistencies, we used our understanding of the socio-cultural context of the data to explore other features of the dataset for similar errors. For example, the “Age” field’s observations were transcribed with the number of months as an incorrect fraction of a year, for example, 3 months was entered as 0.03 of a year which should be ¼ of a year, 0.18 was recorded for 18 months which should be 1.5 years.

Finally, we used Tableau as the tool to visualize the cleansed data. A visualization of the frequency of distribution for the CoF collection is featured in the cross-collection discussion in section VII.

C. Computational Thinking Practices Identified

The computational activities performed on this collection were matched to descriptions of the 22 computational thinking practices. It was found that the activities correspond to ten of the computational thinking practices. See Fig. 6.

The ten relevant CT Practices include:

- **Three Data Practices**: (1) **Manipulating Data** – *Open Refine* was used to clean and normalize the data; (2) **Analyzing Data** in which *Open Refine* was used to analyze the digital features of the data and group them into clusters; and (3) **Visualizing Data** – The *Tableau* visualization tool was used to create to visualize and understand the results of the analysis.
Four Computational Problem Solving Practices: (1) Preparing Problems for Computational Solutions -- The raw data was exported as a csv file from the Maryland State Archives data repository and used with tools like MS Spreadsheet and Open Refine, (2) Choosing effective Computational Tools -- New features were created using the Open Refine tool out-of-the-box through natural language cluster grouping methods, the Tableau was used to perform date functions prior to 1800 appropriately which could not be done using MS Spreadsheet in a feasible manner; (3) Assessing Different Approaches/Solutions to a Problem -- Different computational software tools like MS Spreadsheet, Open Refine and Tableau were assessed for ease-of-use and limitations and (4) Creating Computational Abstractions -- Open Refine and Tableau tools were used to create charts and graphs on the temporal data related to the date of birth, the year of issue of the Certificates of Freedom document for focussed analysis.

Three Systems Thinking Practices: (1) Investigating a Complex System as a Whole to understand the processes/mechanisms that have been involved with the issuance of the documents, to understand the thought process behind manually reading and transcribing them into a finite set of digital features, (2) Understanding the Relationships within a System between the enslaved and owner, events, dates, and places and linking common elements together across collections to find networks, and (3) Communicating the information with stakeholders.

VII. CROSS COLLECTION EXPLORATION

Comparing visualizations across collections, we looked for patterns that might indicate significant historical events in the lives of Maryland’s Black population. For example, in Fig. 7 we noted that basic visualization of frequencies for both the CoFs (top, created using Tableau) and the manumissions (created using R) showed a rather abrupt change in the early 1830s. When we consulted timelines related to slavery in Maryland, we noted both the formation of the Maryland Colonization Society in 1831 and Nat Turner’s Revolt in neighboring Virginia in 1832 [27]. The former tracks with the spike in both CoFs and manumissions, as enslaved people were manumitted on the condition of emigration to Africa. The latter was cause for the introduction of a period of renewed restrictions on the free Black population and reduced manumissions.

The teams also explored rudimentary linkages of key individuals across the manumissions and CoF datasets. Since people manumitted after 1803 would be required by law to carry a CoF, we hypothesized that finding those matches would be relatively simple.

Identifying matching data fields across collections proved to be less straightforward, however. Demographic information about age and even gender (sex) contained in the two collections could not be used, as many times such information was estimated (age) or missing. Using dates also was challenging, due to the multiple dates found in manumissions documents. Matching with the use of names also was not reliable: enslaved people often were only identified by a first name or nickname. We eventually settled on using county, owner first and last names, and the first name only of an enslaved person (in manumissions) and freed person (in CoF) to identify matches.

Both R and a graph database tool, Neo4J [28], were used for the cross-collection analysis. Our rudimentary matching algorithm in R was tested on records from one county, Anne Arundel, The Anne Arundel county records tended to be more complete and consistent for both collections, with a richer set of features. The test yielded only 70 matches out of approximately 3400 records.

With Neo4J, we created a mock-up for visualizing the connections between people and events found across both data collections (Fig. 8). Generating a broader set of matches and reshaping the data will be a next step in exploring how this tool might be used to illuminate the connections between persons and events across collections.
contextualization, we found that it was important to all phases of the data workflow: cleaning and contextualization of the data based on socio-cultural factors became an essential component of the systems thinking we engaged to implement data practices and computational problem solving practices. We found that knowledge of the legal, historical, and even geographic context of collecting and creating original source documents had implications for our applications of computational tools and methods.

Since we used previously-digitized records, we were somewhat limited by the prior decisions made by transcribers and digitizers. From an analytic perspective, this meant that we did not have easy access to data rich information that was placed in notes fields. Further, key individuals who would be the primary subjects of the records, enslaved and free Black people, continued to be relegated to a secondary and sometimes anonymous status in digitized records, just as they were in the original source documents. Thus, although we sought to act as technology-enabled mediators to make those key individuals more visible, we were able to do so only in a limited fashion with aggregated representations. Our ability to track Black individuals across collections met with meager results.

Nevertheless, we saw clear potential for using the open source computational tools to recover more of the heritage of the Black population referenced in the datasets. Contextualization of the data based on socio-cultural factors was important to all phases of the data workflow: cleaning and wrangling, exploratory analysis, and visualization. With that contextualization, we found that we could produce meaningful summaries and visualizations that held possibilities for viewing the historical record through a different lens. We saw the possibility of using patterns in the data to provide new insights into the impact of historical events.

On the basis of this exploration, we identify three areas where additional research in computational methods could enhance discovery and knowledge, particularly in collections related to slavery:

1. **Visualization.** Our exploration of visualization methods was quite rudimentary, given current state of the art for using visualization as a discovery, informational, and access tool. Allen [29] outlines five possible approaches for using visualization as an interface to improve access to archival records: timelines, maps, networks, and tools to explore and articulate institutional hierarchies. More sophisticated visualization, using both static and dynamic functionality, would improve the ability to characterize the information in the datasets across time and place, and would enable additional comparisons of information across collections. Visualization might also be used in connection with archival finding aids [16], to enhance user exploration and research of these and similar archival collections.

2. **Graph databases and textual data mining:** Transforming data from existing datasets for use in graph databases opens new possibilities for examining both individual and aggregated relationships between the people and events encoded in the LoS and similar datasets. Textual mining of data now relegated to notes fields would yield richer information that may be more easily stored and analyzed in a graph database versus relational database. More broadly, the use of graph versus relational databases may accelerate use of artificial intelligence and machine learning for analytic insights [30].

3. **Ontologies and metadata:** Use of ontologies related to enslavement are being developed to enable connection of information from disparate sources. Recent research focuses on metadata record types and core fields that frequently occur in various data repositories related to the slave trade [18]. If developed with an attention to the mediator role of technology, these ontologies could evolve to enable representation from multiple perspectives, including the currently hidden perspective of the enslaved and free Blacks. Metadata extensions or “retrofits” can be employed with existing datasets. Ontologies can enable use of graph databases applications and provide a framework for managing further digitization of primary source materials.

**ACKNOWLEDGMENTS**

We wish to acknowledge two funded grants that have contributed to this work: (1) the 2019-2020 AHRC Research Networking for UK-US Collaborations in Digital Scholarship in Cultural Institutions called “International Research Collaboration Network in Computational Archival Science (iRCN-CAS)”, with KCL, TNA and MSA, and (3) the 2020-2022 IMLS Laura Bush 21st Century Librarian Program: “Piloting an Online Collaborative Network for Integrating Computational Thinking into Library and Archival Education and Practice.” In addition, a deep appreciation for the continued support and collaboration of the staff from the
Maryland State Archives who have created these incredible historical resources, Christopher Haley (Director of the Study of Legacy of Slavery in Maryland) and Maya Davis (Research Archivist and Legislative Liaison).

REFERENCES


