Establishing a Research Agenda for Computational Archival Science through Interdisciplinary Collaborations between Archivists and Technologists

PHILIP NICHOLAS, RAJESH KUMAR GNANASEKARAN, LORI PERINE, ALEXIS HILL, and RICHARD MARCIANO
University of Maryland, College Park

Abstract: This paper explores three research questions related to fostering radical collaboration in bringing computational thinking to archival science. The work is carried out in the context of a case study exploring the Maryland State Archive's (MSA) Legacy of Slavery (LoS) "Certificates of Freedom" collection. We use this case study to introduce a 6-part tool called "Template for Documentation" designed by The National Archives/King's College London (TNA/KCL) which aims at eliciting the fundamental collaborative and decision-making motivations that drive interdisciplinary teamwork including the social obstacles and issues encountered.

Introduction

On April 15, 2021, the US National Archives (NARA) announced the release of a number of open datasets for Artificial Intelligence (AI) and Machine Learning (ML) processing experimentation (including both born digital and digitized content). Computational techniques such as AI/ML applied to records and archives, while not mainstream yet, are likely to be adopted as digital collections grow in scale. Similar initiatives are beginning to emerge across many other cultural institutions. In addition, NARA has set a date to stop accepting paper-based records by the end of 2022 and only obtain permanent records in electronic format starting January 1, 2023 (Chappellet-Lanier 2019, 1). To support this new mission statement, NARA has indicated that over the next few years they will need to seriously embrace the use of automation and computational practices like AI. These emerging technologies are profoundly changing and disrupting the theories, practices, and methodologies that archives use to record, capture, process, curate, encode, share, and provide access to information, as illustrated on research projects conducted at The National Archives (TNA, UK).[1]

We advocate for broad cross-disciplinary collaborations. In an effort to evaluate these collaborative approaches, the Arts and Humanities Research Council (AHRC)-funded International Research Collaboration Network in Computational Archival Science (IRCN - CAS) organized a series of collaborative events. These events consisted of two research symposia and two data exploration workshops. During these events, the IRCN - CAS was instrumental in bringing together key national and international experts from various disciplines including computer science, archives, art history, digital humanities, and cultural heritage institutions. The events spanned February 2019 to February 2020 (Goudarouli et al. 2020, 4).

From these events, spaces for interdisciplinary conversations led by student experimentations were initiated, where scholars collaborated on data explorations of digitized collections from The National Archives, UK (TNA) and the Maryland State Archives (MSA). These collaborative events facilitated the process for interdisciplinary practices and cross fertilization of ideas, as well the exploration of opportunities for automation to re-contextualize records. While these types of exchanges are growing in professional circles, they surprisingly tend to be less common with students. This particular project engaged diverse student teams. Disciplinary diversity was represented by students across archival studies, history,

[1] See: https://blog.nationalarchives.gov.uk/tag/artificial-intelligence

Published by Society of American Archivists, 2020.

computer science, human computer interface design (HCI), and information science. Academic diversity was represented by mixing students across degrees including undergraduates, various Master's degree students, and doctoral students. We then augmented these student teams with professionals including data scientists, archivists, software developers, art historians, and the curators of the archival collections themselves). Three of these hybrid teams are depicted in Fig. 1. This approach enables interdisciplinary participants the opportunity to "explore new intellectual terrain, learn the lexicon of other fields, and author works for new audiences" (Poole 2018, 185).

Fig, 1: Three of these hybrid student teams in action at the Oct. 2019 MSA Datathon: (1) Manumission Records (top left), (2) Certificates of Freedom (top right), and (3) Runaway Slave Ads (bottom)

Problem Statement

In this paper, we explore fostering radical collaboration of a community of masters and doctoral students from interdisciplinary backgrounds. In the UK digital scholarship project, "Living With Machines," radical collaboration is defined as an approach which "aims to generate and exploit cross-disciplinary synergies between the team and a wide network of experts, including experts from various academic disciplines" (Ahnert et al. 2020).

This paper explores the following three questions:

1. What happens when prospective and expert archivists / librarians, historians, computer scientists / IT specialists, and faculty members apply Computational Thinking (CT) and explore archival digital collections collaboratively? *[CT is defined as a form of problem solving that uses modeling, decomposition, pattern recognition, abstraction, algorithm design, and scale (Wing 2006)]*

2. What are the challenges and limitations of contextualizing historical information through a interdisciplinary collaborative framework? *[the context of a record is key for understanding its value as historical evidence – thus we are concerned with the impact of contextualization through computational enhancements such as metadata extraction, AI/ML, linking, etc.]*
3. How does applying CT through an interdisciplinary approach enable the contextualizing historical information and further new research, access, and discovery? *[this work was started and documented in (Underwood 2019)]*

Methodology

The Advanced Information Collaboratory (AIC)[2] at the University of Maryland (UMD) has partnered with The National Archives (TNA) and King's College London (KCL) and the Maryland State Archives (MSA) in an international collaboration to explore the application of computational techniques to digital heritage collections (Goudarouli et al. 2020, 2). Within the partnership, an important contribution by the AIC is training current and future information professionals to think computationally and understand the potential of utilizing new technologies through collaborative interdisciplinary experiences (Marciano 2020-2, 1-2). Consistent with that focus, the AIC recruited interdisciplinary teams of students, among them the authors of this paper (represented in Fig. 1, top two photos), to collaborate on computational exploration of archival collections.

In this paper, an interdisciplinary team with backgrounds in history, archival science, and information technology, convened through the AIC to collaborate on a case study exploring the MSA's Legacy of Slavery (LoS) collection. The MSA embarked upon a major initiative in 2001 to create the LoS collection, in order to document the stories and experiences of the enslaved people and free Black[3] populations in Maryland from 1634 to 1867 (Haley and Davis 2020). Over the course of the initiative, MSA staff members and volunteers transcribed and digitized records related to approximately 420,000 individual names and organized them into 16 separate and un-networked SQL databases (Perine et al. 2020, 1).

In this paper, we focus primarily on the interdisciplinary collaborative explorations and cross-collection examinations of the Certificates of Freedom (CoF) records. A CoF was a legal document issued from 1806 to 1864 by a Maryland county clerk to freed Black individuals, who were required to show legal proof of their emancipated status.

To explore questions about the nature of the collaborative interactions, the project team used a six-part "Template for Documentation" (Template) developed by TNA and KCL. The Template was designed by The National Archives/King's College London (TNA/KCL) specifically for this event. It is a six-part tool which aims at eliciting the fundamental collaborative and decision-making motivations that drive interdisciplinary teamwork including the social obstacles and issues encountered: (1) general questions, (2) approach and decision-making, (3) obstacles and issues, (4) working in interdisciplinary teams, (5) moving forward, and (6) out-of-the-box (details in Appendix 1). One of the major outcomes of this paper, is the first utilization and testing of this Template tool using the the CoF collaborative work (details in Appendix 2).

Finally, we situate the computational contributions by using a processing pipeline (so-called workflow) (Fig. 2). The area outlined in blue shows the MSA's prior work creating digitized records from original source documents. The data pipeline highlighted in red shows the computational exploration that the students conducted collaboratively. As the team explored the LoS datasets using open-source computational tools for data cleaning, metadata extraction, and visualization, the technical decisions were informed by the social and cultural systems expertise of team members.

[2] Advanced Information Collaboratory, see: https://ai-collaboratory.net

[3] For this paper, we will refer to the descendants of Sub-Saharan Africans in the state of Maryland as Black people.

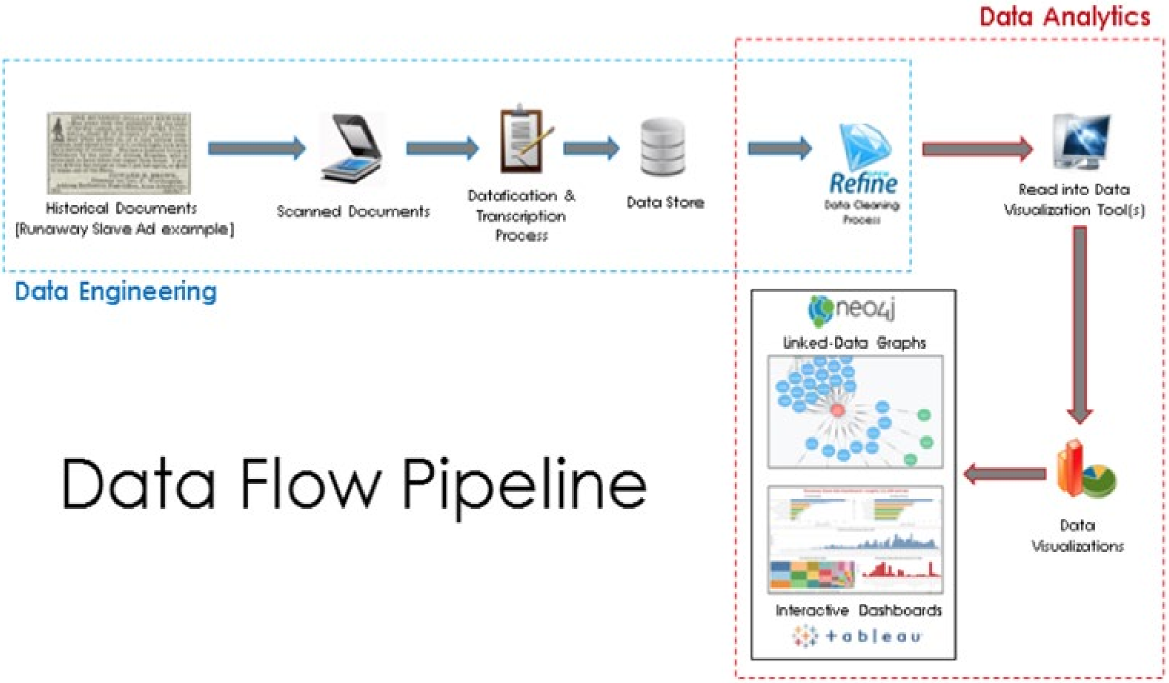
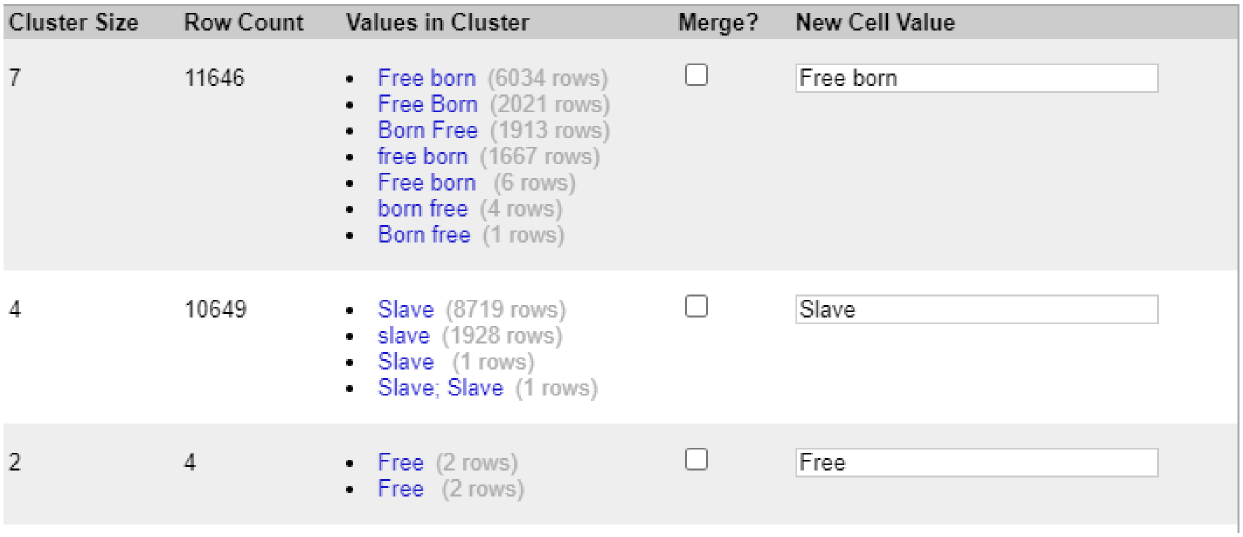


Fig. 2: Data Workflow for Computational Exploration

Detailed Description of Case

We primarily report on work conducted with the Certificates of Freedom (CoF) dataset collection, consisting of 23,655 records that were transcribed by the MSA from the handwritten documents which contained general demographic, biographic and descriptive information about a freed person at the time of issue. From these physical documents, the MSA captured a set of features comprised of namely the county of issue, slave owner’s first and last name, enslaved person’s first and last name, gender, age, height, complexion, scar (to identify the person), date of issue, witness name, prior status of the enslaved person, and a note feature capturing remarks from the clerk (Perine 2020, 7). The dataset could be understood to contain a rich set of historical and contextual features which needed a multi-disciplinary approach to analyze. The IT specialists on the team downloaded information from the MSA’s database as a .csv file. For data features such as date of issue, age, complexion, and prior status, among others, the historians and archivists on the project team provided knowledge about the socio-cultural and legal practices of Ante-bellum Maryland, so that we could understand how to prepare the data for analysis. Together, the interdisciplinary team reviewed the types of entries found in each field using open-source tools such as Open Refine, for data exploration and cleaning, and Tableau, for visualization. Thus, we were able to augment standard data analysis tasks, such as treating outliers or duplicate entries, with greater understanding of how to spot inconsistencies and transcription errors. In addition, we were able to cluster subjective data entries that referred to similar phenomena. The dataset that we ultimately used contained records from 16 of Maryland’s 22 counties with the date of CoF issue between 1806 to 1864. The documented age of the CoF recipients ranged from 3 months to 82 years. We also noted that the CoFs were issued almost exclusively (93%) to males.

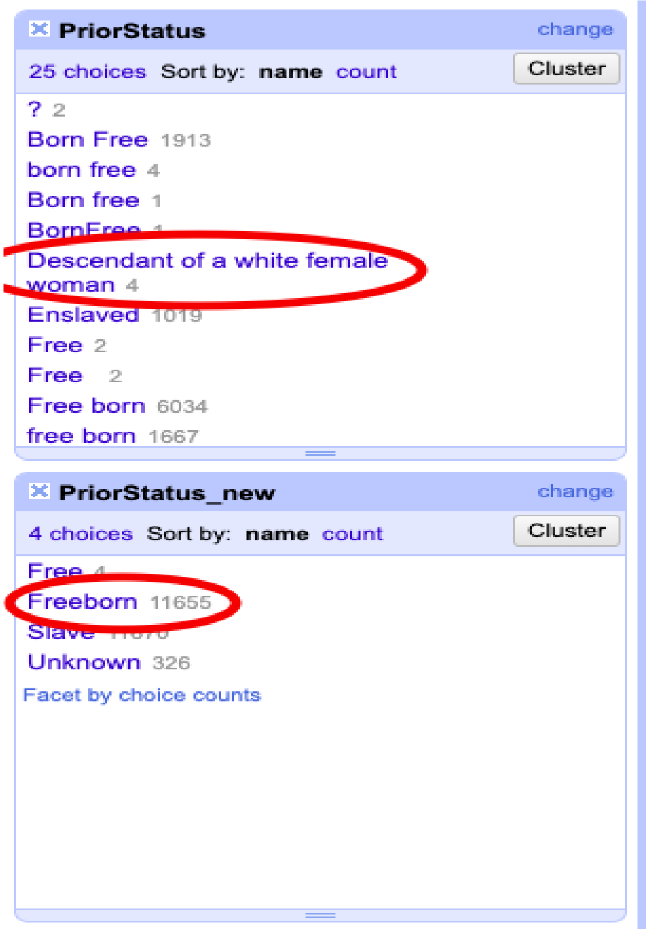
IT specialists and computer scientists leveraged Open Refine tool’s unique options like Facet feature, feature transformations, and clustering of records for exploration and cleaning purposes. The Facet feature was used to explore each feature as numeric or text data. For numeric data, the tool distributed the data as a range from lowest to highest set of values. This was instrumental for identifying inconsistencies, such as in the case of the 'Height' feature, where a record documented the height as 9 feet 7.5 inches. When using Facet to choose data as text, the tool groups the data into unique values as categories. Thus, computer scientists were able to transform features that were originally documented as text to numeric type for effective exploration and visualization. In order to analyze features such as complexion and prior status, the team historians and archivists helped the computer scientists and IT specialists to understand the contextual nature of the underlying data. The use of text facets, along with clustering, allowed us to knowledgeably group the features. For instance, the feature *prior status* contains free form text which describes whether a person was formerly enslaved then manumitted ("freed"), born outside of enslavement, or otherwise not subject to enslavement. The text values in this field were captured in many variations, which the team was able to identify as belonging to similar groups using the Open Refine clustering option. Fig. 3 shows how text notations for being born outside of enslavement were documented by the transcribers in different variations (*Free born, born free, free born*, etc.). The tool clustered them into a single group then assigned a unique value for the group.



Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
7	11646	• Free born (6034 rows) • Free Born (2021 rows) • Born Free (1913 rows) • free born (1667 rows) • Free born (6 rows) • born free (4 rows) • Born free (1 rows)	□	Free born
4	10649	• Slave (8719 rows) • slave (1928 rows) • Slave (1 rows) • Slave; Slave (1 rows)	□	Slave
2	4	• Free (2 rows) • Free (2 rows)	□	Free

Fig. 3: Cluster feature in Open Refine tool for 'Prior Status'

The *prior status* feature also had a set of values that were identified as "Descendant of a white female woman". To determine to which group these values really belonged, the historians played a vital role. Digging through historical texts, they were able to provide the valuable insight that, during the era when the CoF's were issued, a person born to a white female was classified as being born outside of enslavement. This historical context allowed us to recode the values shown in Fig. 4 below as "Free born" in the dataset.



PriorStatus change

25 choices Sort by: name count Cluster

? 2
Born Free 1913
born free 4
Born free 1
BornFree 1
Descendant of a white female woman 4
Enslaved 1019
Free 2
Free 2
Free born 6034
free born 1667

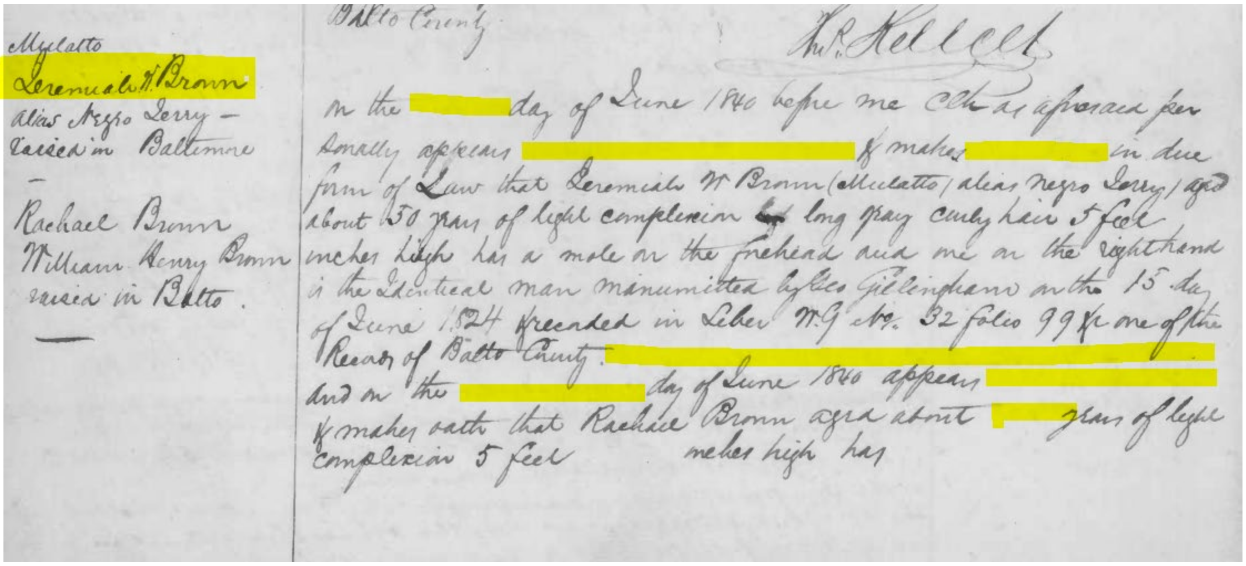
PriorStatus_new change

4 choices Sort by: name count Cluster

Free 4
Freeborn 11655
Slave 11070
Unknown 326
Facet by choice counts

Fig. 4: Cluster feature in Open Refine tool for 'Prior Status'

As the computer scientists and IT specialists explored the dataset for the feature *date of issue*, there were values which simply did not fit in a range of numeric data as they were improperly documented. One such example is shown in the archivists' help was needed to determine the record did not have a specific day of issue. As could be seen in Fig. 5, the archivists found a document match with the First and Last name of the freed person for identification as the original document itself did not have the day of issue. However, it did have the month and year of issue. With this information, after consensus among the team members, it was assumed that the day of issue was not documented because it was missing in the original text. Hence, we agreed to assign an arbitrary day to include this observation into the numeric range. This "smoothing" of the date was done to not lose other valuable information that might have arisen by including only the month and year of the date of issue from this record.



Mulatto
Jeremiah W. Brown
alias Negro Jerry —
raised in Baltimore
—
Rachael Brown
William Henry Brown
raised in Balto.

Balto County

Kell Clk

On the day of June 1840 before me Clk as aforesaid per
sonally appears & makes in due
form of Law that Jeremiah W Brown (Mulatto) alias Negro Jerry) aged
about 50 years of light complexion long gray curly hair 5 feet
inches high has a mole on the forehead and one on the right hand
is the Identical man manumitted by Geo Gillingham on the 15 day
of June 1824 & recorded in Liber W.G No. 32 folio 99 &c one of the
Records of Balto County.
And on the day of June 1840 appears
& makes oath that Rachael Brown aged about years of light
complexion 5 feet inches high has

Fig. 5: Date feature highlighted with missing Day of issue

Throughout the process, it should be noted that the team, especially with inputs from the archivists and historians, retained the original data values so as to not lose the provenance of these data features. Whenever we modified an existing feature, a new data field was created, and our changes were saved in the new column.

After cleaning the dataset with OpenRefine, we uploaded it to the Tableau tool, which is mostly used for visualization. With this tool, the computer scientists and IT specialists created several charts and graphs of this unique dataset. Of particular note is the time series visualization of the *date of issue* feature showing the year when male and female freed people were issued their CoF. The historians and archivists made a stunning discovery that no females were issued CoFs until 1833 whereas male had been issued CoFs since 1806 (see Fig. 6). They also noted that this particular revelation was made possible mostly because of the idea of visualizing the digital archival data using computational tools, in collaboration with the IT and computer science team members.

Findings

This interdisciplinary approach brought together researchers, students, professors, and professionals from varied backgrounds for a collaborative workshop on the Legacy of Slavery data collections. The collaborative work enabled the team members to learn about the collections from historical perspectives and at the same time leverage readily available software tools to explore, clean and glean insights from the data.

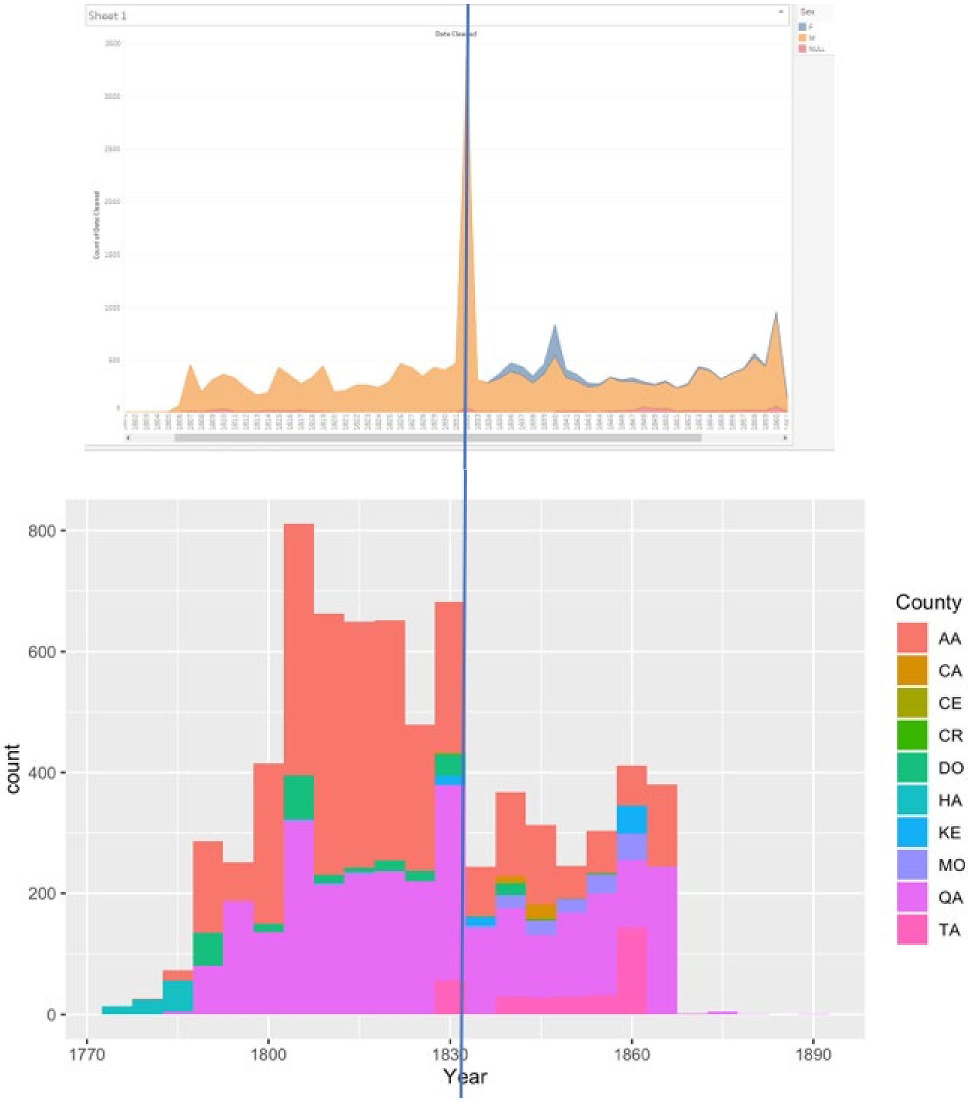


Fig. 6: Annual Certificates of Freedom (by gender) and Manumissions (by county)

With inputs from experts on both the teams, the discussions around historical context and data exploration culminated in some unexpected conversations and findings. For example, Fig. 6 above shows one such significant insight in which computer scientists aligned a time-series visual representation of the *date of issue* feature vs gender of the CoF collection with a similar time-series visual representation of the number of recorded manumissions by county from the separate dataset created from the Manumissions collection. Using the resulting visualization, the historians and archivists noted that there was an abrupt change in the issuance of both the CoF and Manumissions around the early 1830's, prompting them to investigate historical events around that time period relevant to the Black population. They were able to identify two major historic events that took place around this time period: the formation of the Maryland Colonization Society in 1831 and Nat Turner's Revolt in neighboring Virginia in 1832 (The Maryland State Archives 2007, 29). The former event could be a reason as to why there was an abrupt change in the counts of CoF and Manumissions issued as enslaved people were issued Manumissions contingent upon the condition of emigration to Africa (Perine 2020, 8).

Efforts to link the two dataset collections through common features like the enslaved person's first and last name, slave owner's first and last name, county issuing the document were also attempted. The idea was based on the historical assumption that when an enslaved person is issued a Manumission record it could eventually be used to receive a Certificate of Freedom by producing that document to a county clerk. However, the outcome of this data link process was not encouraging as there were non-unique record matches between the two collections. Later, the historians pointed out that the duplicates were prone to happen as the enslaved people were often identified by their first name or nickname and rarely by their last name, which was evident from the results as most of the duplicates were because of multiple people matches with the first name and last name as 'NULL'.

From this experiment, we understand that the historical, cultural, contextual knowledge of the data analyzed had a pronounced impact on the results and findings. In addition, the IT specialists and computer scientists benefitted appreciably from the historians' and archivists' perspectives as an introduction to leveraging open-source tools like Open Refine, and visualization tools like Tableau for exploring digitally archived collections.

Finally, since this was the first time the six-part "Template for Documentation" (Template) was put to use, we offer some reflections on the tool, focusing on two section that are probably the most important with respect to eliciting the fundamental collaborative and decision-making motivations that drive interdisciplinary teamwork:

- Section "3. Obstacles and issues": The answers do not bring up organizational or interdisciplinary issues (most likely because the work conducted over two days and in person, with ample discussions and interactions). Issues arising from data were categorized in terms of: data entry, ambiguity or missing data in the original records, and technical limitations of the tools themselves.
- Section 4. Working in interdisciplinary teams": The responses focus on the value of multidisciplinary insights, understanding different perspectives, historically contextualizing findings, and sharing ideas about archival provenance and data analysis.

Appendix 2: captures the broader responses. Feedback from participants indicated that using this Template tool was useful, but that it would need to be tested across additional collaborations.

Conclusion

With large scale digitization projects of analog archives and the emergence of born-digital archives occurring, the emergent field of computational archival science provides a context to enable interdisciplinary, radical collaboration within archival research. As archivists, we provide access and retrieve information for public use, and as technologists we provide computational treatment to born-digital collections. Through this collaborative exploratory case study, we explored the use of computational tools to better understand the archival record, thus expanding possibilities for research, discovery, and new knowledge. Although library and archival science degree students are gaining the necessary theoretical skills to understand the contextualization of archival records, they are still not receiving the adequate practical skills in order to be competitive in a changing field starting to rely on automation and computational practices. However, if this radical collaborative approach could be replicated and implemented within the teaching curriculum of LIS educational institutions, future archivists will gain the skills necessary to adapt to the changing digital nature of archives. In addition, we see space for current archivists and technologists in archives to utilize this same practice, which will develop and prepare these practitioners with the needed skills and tools to succeed in a changing industry. By learning new skills through radical collaboration, students and practitioners will be prepared to further enhance digital access to archival collections and increase research and discovery for user communities. Documenting aspects of the collaboration were enabled through the use of the Template.

Resources

Advanced Information Collaboratory. "Oct. 28-29, 2019: Student-Led 'Datathon' at the Maryland State Archives." https://ai-collaboratory.net/projects/ct-los/student-led-datathon-at-the-maryland-state-archives/

Ahnert, Ruth, David Beavan, Emma Griffin, James Hetherington, Jon Lawrence, Maja Maricevic, Barbara McGillivray, Mia Ridge, and Alan Wilson. "Living With Machines." Accessed October 15, 2020. https://livingwithmachines.ac.uk

Chappellet-Lanier, Tajha. "OMB Issues Guidance on NARA's Transition to Electronic Record Keeping." *Fedscoop*, July 1, 2019. https://www.fedscoop.com/nara-electronic-records-omb-guidance/

Goudarouli, Eirini, Anna Sexton, and John Sheridan. "The Challenge of the Digital and the Future Archive: Through the Lens of The National Archives UK." *Philos. Technol.* 32, no. 1 (2019): 173-183. https://doi.org/10.1007/s13347-018-0333-3

Goudarouli, Eirini, Mark Hedges, Richard Marciano, David Beavan. "Computational Archival Science (CAS): An international research collaboration network," at the 2020 Annual TNA Digital Lecture and Staff Research Poster Exhibition. See:https://www.nationalarchives.gov.uk/about/our-research-and-academic-collaboration/our-research-and-people/2020-annual-digital-lecture-staff-research-poster-exhibition/#computational-archival-science

Haley, Christopher, and Maya Davis. "Legacy of Slavery in Maryland, The Maryland State Archives." Accessed October 15, 2020. https://slavery.msa.maryland.gov/

The Maryland State Archives, and The University of Maryland, College Park. *A Guide to the History of Slavery in Maryland*. Maryland: Maryland State Archives, 2007. https://msa.maryland.gov/msa/intromsa/pdf/slavery_pamphlet.pdf

Marciano, Richard, 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: https://ai-collaboratory.net/wp-content/uploads/2020/10/Marciano-et-al-Archival-Records-and-Training-in-the-Age-of-Big-Data-final.pdf

Marciano, Richard, Greg Jansen, and William Underwood. "Developing a Framework to Enable Collaboration in Computational Archival Science Education." (Chicago: Society of American Archivists, 2020), 1-10. https://www2.archivists.org/sites/all/files/Research_Forum%20_2019_Marciano_final.pdf

Marciano, Richard, Greg Jansen, William Underwood, Mark Conrad, Michael Kurtz, Lyneise Williams, Jane Greenberg, Victoria Lemeiex, Mark Hedges, and Eirini Goudarouli. "Welcome to the Advanced Information Collaboratory (AIC)." Accessed October 15, 2020. https://ai-collaboratory.net

Perine, Lori A., Rajesh Kumar Gnanasekaran, Phillip Nicholas, Alexis Hill, and Richard Marciano. "Computational Treatments to Recover Erased Heritage: A Legacy of Slavery Case Study (CT-LoS)." *IEEE Big Data 2020, Computational Archival Science (CAS) workshop #5, Virtually* (2020): 1-10.

Poole, Alex H., and Deborah A. Garwood. "Interdisciplinary Scholarly Collaboration in Data Intensive, Public-Funded, International Digital Humanities Project Work." *Library and Information Science Research* 40, no. 3-4 (2018): 184-93. https://doi.org/10.1016/j.lisr.2018.08.003

Underwood, William, and Richard Marciano. "Computational Thinking in Archival Science Research and Education." *IEEE Big Data 2019, Computational Archival Science (CAS) workshop #4, Los Angeles* (2019): 1-7. https://ai-collaboratory.net/wp-content/uploads/2020/02/Underwood.pdf

Wing, Jeannette, 2006, " Computational Thinking". Communications of the ACM. 49(3), 33-35. Retrieved from https://www.cs.cmu.edu/~15110-s13/Wing06-ct.pdf

Appendix 1: **TNA-KCL Template for Documentation**

(Mark Hedges @ KCL & Eirini Goudarouli @ TNA)

1. **General Questions**
   a. What is the overall topic/area being addressed by the group?
   b. What are the specific challenges raised by this topic?
   c. Which challenges are you addressing?

2. **Approach and decision making**
   a. What is the approach being taken?
      i. Consider methodological aspects, technical aspects, etc.
      ii. In particular, what technical questions/tasks are being addressed in your group?
   b. How was the approach selected? Were there alternatives that were considered but not followed up?
   c. Document each stage of the decision process:
      i. What decisions are taken?
      ii. What are the options?
      iii. On what grounds is a particular decision made (evidence, criteria, ...)
      iv. What specific steps is your team taking?
   d. What was your experience of combining methodologies from different disciplines?
      i. Did you note any incompatibilities?
      ii. Did you develop any new methodological pathways?
   e. How did you divide up the work within your group? Was this division related to disciplinary background?
3. **Obstacles and issues**
   a. What obstacles/issues are you encountering?
      i. Organizational issues
      ii. Interdisciplinary issues
      iii. Issues arising from the data
      iv. Technical or tool issues
4. **Working in interdisciplinary teams**
   a. What disciplines are represented in your working group?
   b. What benefits/opportunities arising from interdisciplinarity have you noticed in your group?
   c. What challenges arising from interdisciplinarity have you noticed in your group?
      i. For example, in terms of communicating across disciplines.
   d. Are there any terminological or other confusions arising from working in interdisciplinary teams?
5. **Moving forward**
   a. Do you see specific possibilities arising from your investigations?
   These can include:
   i. ideas for next steps
   ii. things you might want to try but haven't
   iii. etc.
6. **Out-of-the-box**
   a. Please capture any ideas or discussions that arose from your group that do not fit into any of the previous sections.

Appendix 2: **Applying the Template for Documentation to the Certificate of Freedom (CoF) Collection**

1. General Questions

a. What is the overall topic/area being addressed by the group?

We extracted, analyzed, cleaned and created visualizations from the Certificates of Freedom dataset obtained from the Maryland State Archives' Legacy of Slavery Database.

b. What are the specific challenges raised by this topic?

We anticipated errors and misinterpretation of names, numbers, etc. since this database was mostly transcribed manually by hand from the physical or scanned copies of the Certificates of Freedom.

c. Which challenges are you addressing?

We looked at the dataset holistically at first, identifying features that allowed us to generate meaningful stories or visualizations. Upon confirmation of the features list, we would analyze each of them in detail to document bad data and eliminate them if possible, modify data types, exclude them from the final visualizations if found to be invalid, etc.

2. Approach and decision making

- **What is the approach being taken?**
  - **Consider methodological aspects, technical aspects, etc.**

    We followed a case study methodology for this project to achieve the objective of exploring, analyzing and visualizing the dataset collections downloaded from the Maryland State of Archives database.

  - **In particular, what technical questions/tasks are being addressed in your group?**

    As the dataset collections were available as downloadable csv files, the technical tasks addressed by our group were to identify the right tools that could be used to consume the csv files for exploratory analysis, cleaning and visualization purposes. The group decided to use the open source tool Open Refine based on previous success with it for exploration, cleaning, and use tableau software for visualization purposes.

- **How was the approach selected? Were there alternatives that were considered but not followed up?**

  Our approach was to individually clean the data column-by-column utilizing the text and numerical facet features in OpenRefine, then (2) combine the cleaned columns in GoogleSheets, and (3) finally visualize the cleaned data file utilizing Tableau software. Our original approach for Step 2 was to save all of our cleaned files in UMD Box Service, then assign one person to combine columns from the multiple cleaned files at once. However, GoogleSheets offered a more collaborative space that allowed all group members to work on simultaneously.

- **Document each stage of the decision process:**
  - **What decisions are taken?**

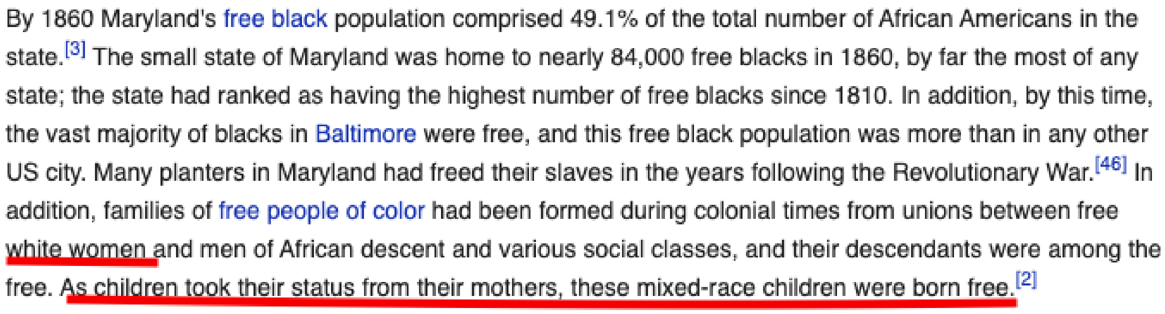
  As the dataset collections contained features that could be individually worked upon, based on the expertise of each team member, the features were divided among each of us to research, explore, clean and report on the findings from the assigned modules. Later as a group, brainstorming from individual results, providing feedback, and to glean insights from the findings.

- **What are the options?**

As the team members were from a diverse group of technology, historical, and archivist background, there were options to work individually all along or to work in groups all along, but we decided to do a hybrid setup of analyzing alone and reporting the results back to the group for discussion.

- **On what grounds is a particular decision made (evidence, criteria, ...)**

With respect to the analysis performed on the dataset, decisions were data-driven or historical facts driven. For instance, to address the feature in CoF dataset - Prior Status Column: Research was conducted to determine the prior status of those who were categorized as a "Descendant of a white female woman". Source: Wikipedia - History of slavery in Maryland. This research was beneficial in identifying what group certain observations belong to.



By 1860 Maryland's free black population comprised 49.1% of the total number of African Americans in the state.[3] The small state of Maryland was home to nearly 84,000 free blacks in 1860, by far the most of any state; the state had ranked as having the highest number of free blacks since 1810. In addition, by this time, the vast majority of blacks in Baltimore were free, and this free black population was more than in any other US city. Many planters in Maryland had freed their slaves in the years following the Revolutionary War.[46] In addition, families of free people of color had been formed during colonial times from unions between free white women and men of African descent and various social classes, and their descendants were among the free. As children took their status from their mothers, these mixed-race children were born free.[2]

- **What specific steps is your team taking?**

Through researching the literature, conversations with historians and experts in the field, discussions with archivists from the Maryland State Archives, the team members followed a set of steps where certain unique characteristics of a particular feature for instance were identified and shared with the entire group for their inputs before finalizing the results.

d. What was your experience of combining methodologies from different disciplines?

- **Did you note any incompatibilities?**

We did not note any incompatibilities between team members although there were healthy discussions on what-if scenarios as most of the data were historical and we were bringing each of our expertise into the conversations.

- **Did you develop any new methodological pathways?**

No.

e. How did you divide up the work within your group? Was this division related to the disciplinary background?

Every group member worked individually on cleaning one to three columns, however, some tasks were allocated to each group member based on disciplinary background, which is outlined below:

- Rajesh: Height, Date, Age, (Open Refine, Tableau and Jupyter Notebooks)
- Jeneva: Complexion and Experienced with JupyterNotebooks, OpenRefine, & Tableau
- Philip: First & Last Names of former enslaved & owners and provided historical knowledge and context of prior status and complexion (Open Refine)
- Alexis: Prior Status (Open Refine)

3. Obstacles and issues

a. What obstacles/issues are you encountering?

- **Organizational issues**

We mostly collaborated in-person for this project at the Maryland State Archives facility which immensely supported this activity and allowed us to access their data for exploration and analysis. As such there were no organizational issues with this Project.

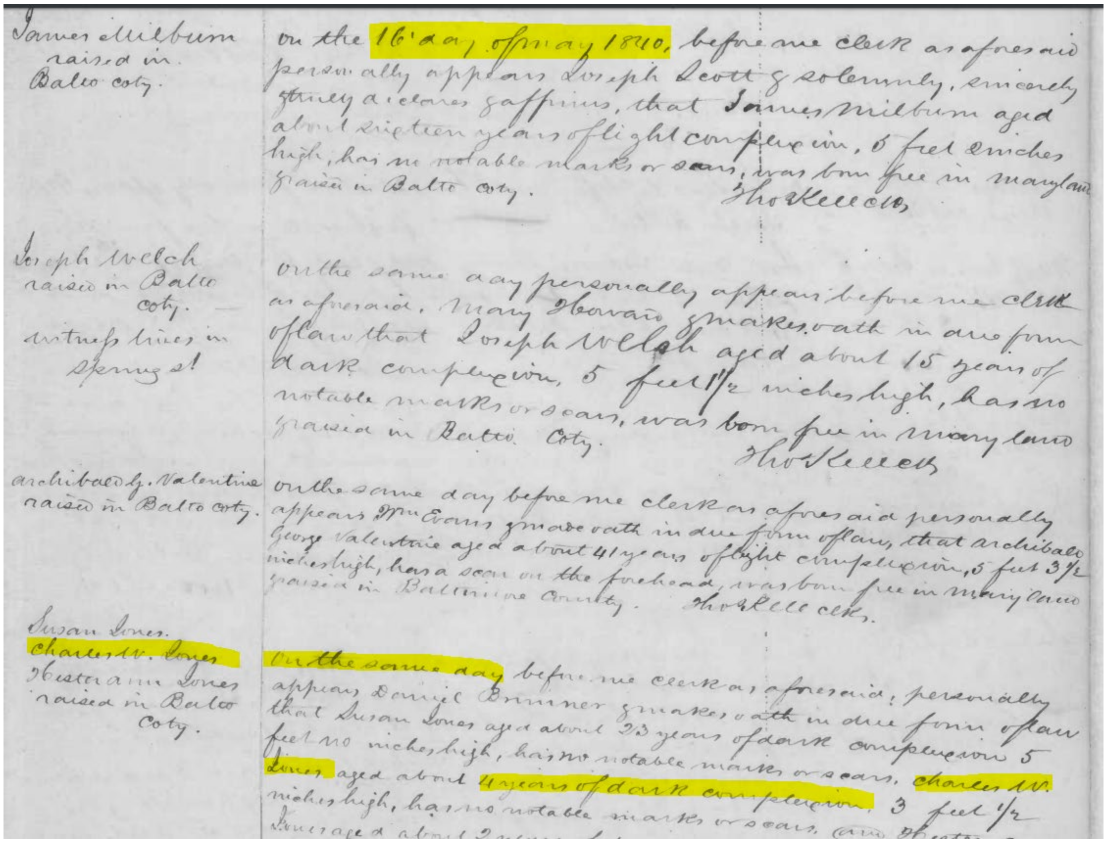
- **Interdisciplinary issues**

As there were members from different backgrounds, there were discussions on how data should be presented, collected, and analyzed without impacting the sensitivity of the people involved, especially since this set of collection was unique.

- **Issues arising from the data**

Date Feature -- This field is to indicate the date when the certificate of freedom was prepared and signed. There were a number of issues with this date field in the original dataset.

- Different date formats -- There were around 600 records with NULL value, a bunch of them with just YYYYMM format, most of them in the format YYYY-MM-DD and YYYYMMDD format.
- Another instance of data entry error was for c290 page 185 Charles W Jones as shown below with the date captured as 1840516 instead of 18400516



James Milburn
raised in
Balto coty.

On the 16'day of may 1840, before me clerk as aforesaid
personally appears Joseph Scott & solemnly, sincerely
& truly declares & affirms, that James Milburn aged
about sixteen years of light complexion, 5 feet 8 inches
high, has no notable marks or scars, was born free in Maryland
& raised in Balto coty.
Thos Kell clk

Joseph Welch
raised in Balto
coty.
witness lives in
Spring st

On the same day personally appears before me clerk
as aforesaid, Mary Howard & makes oath in due form
of law that Joseph Welch aged about 15 years of
dark complexion, 5 feet 1/2 inches high, has no
notable marks or scars, was born free in Maryland
& raised in Balto. Coty.
Thos Kell clk

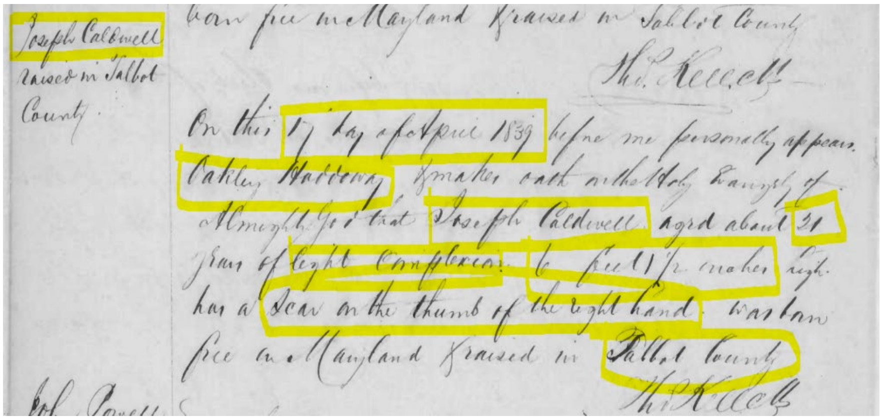
Archibald G. Valentine
raised in Balto coty.

On the same day before me clerk as aforesaid personally
appears Wm Evans & made oath in due form of law, that Archibald
George Valentine aged about 41 years of light complexion, 5 feet 3 1/2
inches high, has a scar on the forehead, was born free in Maryland
& raised in Baltimore County. Thos Kell clk.

Susan Jones.
Charles W. Jones
Hester Ann Jones
raised in Balto
Coty.

On the same day before me clerk as aforesaid, personally
appears Daniel Brunner & makes oath in due form of law
that Susan Jones aged about 23 years of dark complexion 5
feet no inches high, has no notable marks or scars, Charles W.
Jones aged about 4 years of dark complexion, 3 feet 1/2
inches high, has no notable marks or scars, and Hester
Jones aged about 2

- There was this unique case where it seemed like a clear data entry error as shown below: The date was captured as 189390417 but the actual date is 18390417 identified by looking at the scanned copy of the CoF for c290 page 130 for Joseph Caldwell, the county is found to be Talbot from the original ad but the data was entered as Baltimore County only for CoF but Census was captured correctly as Talbot county



Joseph Caldwell
raised in Talbot
County

born free in Maryland & raised in Talbot County

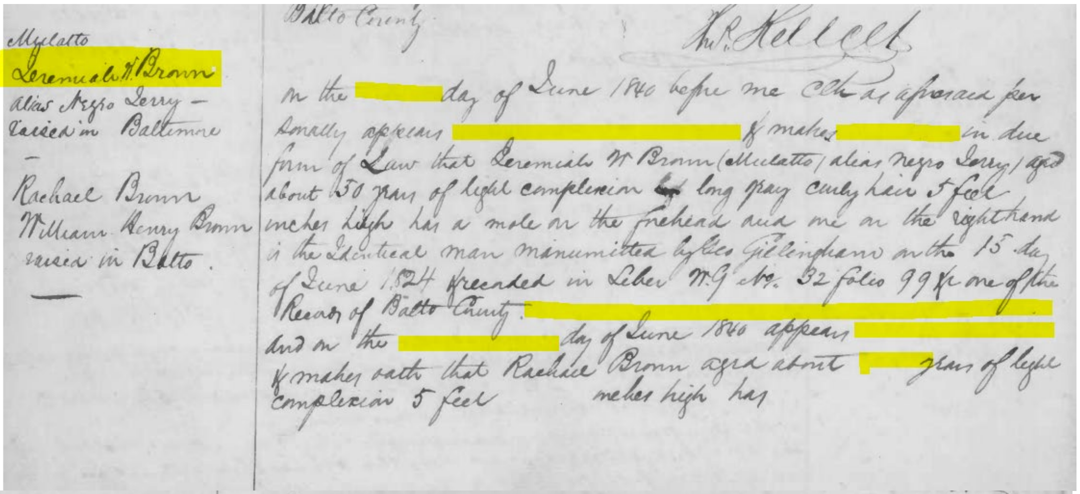
Thos. Kellett

On this 17 day of April 1839 before me personally appears Oakley Haddaway & makes oath on the Holy Evangely of Almighty God that Joseph Caldwell aged about 21 years of light Complexion 6 feet 1 1/2 inches high. has a Scar on the thumb of the right hand. was born free in Maryland & raised in Talbot County

Thos Kellett

Jol. Powell

- There were other instances one of which is shown below where there was no date but only month and year captured on the original CoF itself for c290 page 224 - Jeremiah Brown



Mulatto
Jeremiah W. Brown
alias Negro Jerry –
raised in Baltimore

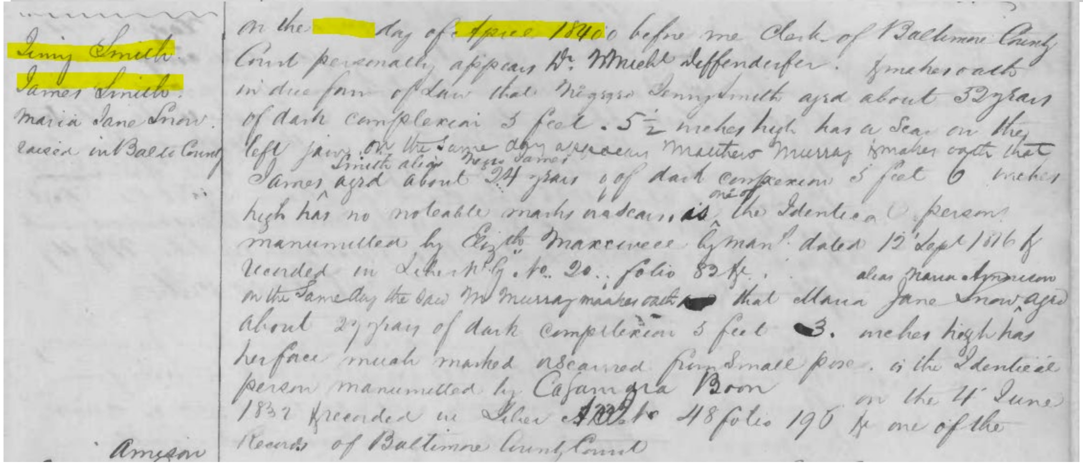
Rachael Brown
William Henry Brown
raised in Balto.

Balto County

Thos. Kellett

On the day of June 1840 before me Clk as aforesaid personally appears & makes in due form of Law that Jeremiah W Brown (Mulatto) alias Negro Jerry) aged about 50 years of light complexion ~~&~~ long gray curly hair 5 feet inches high has a mole on the forehead and one on the right hand is the Identical man manumitted by Geo Gillingham on the 15 day of June 1824 & recorded in Liber W.G. No. 32 folio 99 & one of the Records of Balto County.

And on the day of June 1840 appears & makes oath that Rachael Brown aged about years of light complexion 5 feet inches high has

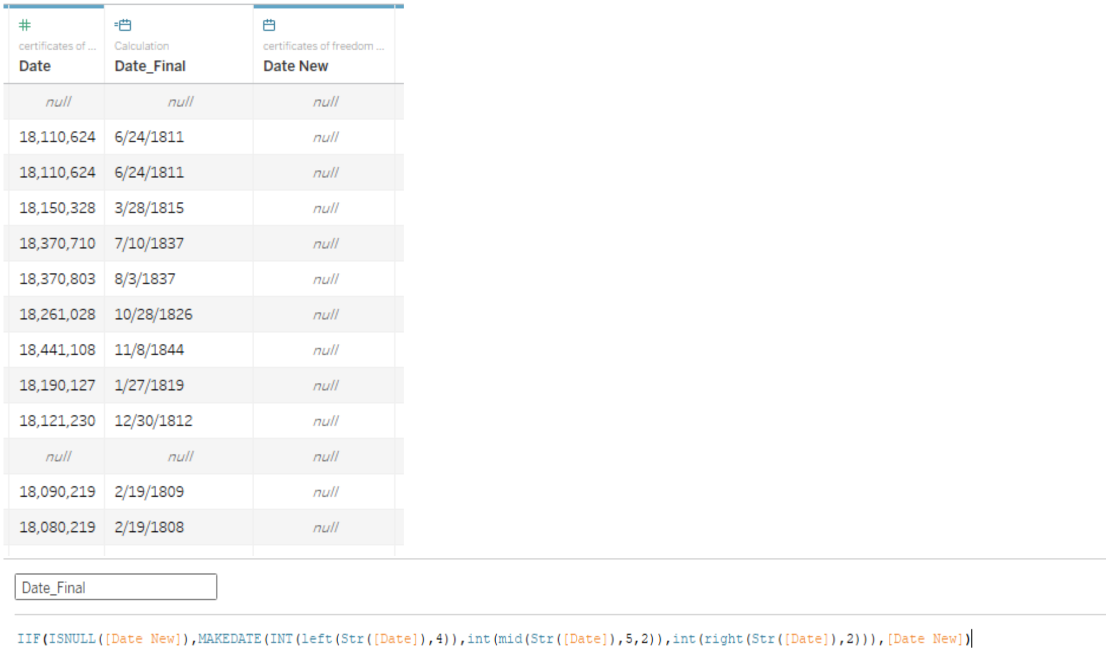


Jenny Smith.
James Smith.
Maria Jane Snow.
raised in Balto County

Ameson

On the day of April 18400 before me Clerk of Baltimore County Court personally appears Dr. Michl Deffenderfer. & makes oath in due form of Law that Negro Jenny Smith aged about 32 years of dark complexion 5 feet. 5 1/2 inches high has a Scar on the left jaw, on the same day appears Matthew Murray & makes oath that James Smith alias Negro James aged about 24 years of dark complexion 5 feet 6 inches high has no noteable marks or scars, is one of the Identical persons manumitted by Elizth Maxwell by manl. dated 12 Sept 1816 & recorded in Liber W.G. No. 20. folio 83 &c. on the same day the said Mr Murray makes oath that Maria Jane Snow alias Maria Ayrson aged about 27 years of dark complexion 5 feet 3. inches high has her face much marked & scarred from Small pox. is the Identical person manumitted by Casamgra Boon on the 4 June 1832 & recorded in Liber A.W. No. 48 folio 196 & one of the Records of Baltimore County Court

- One of the important limitations while working with excel with dates older than 01/01/1900 was that the dates are not calculated and translated correctly. Hence after proper formatting the dates were loaded to Tableau and created a calculated field to handle the dates as shown below:



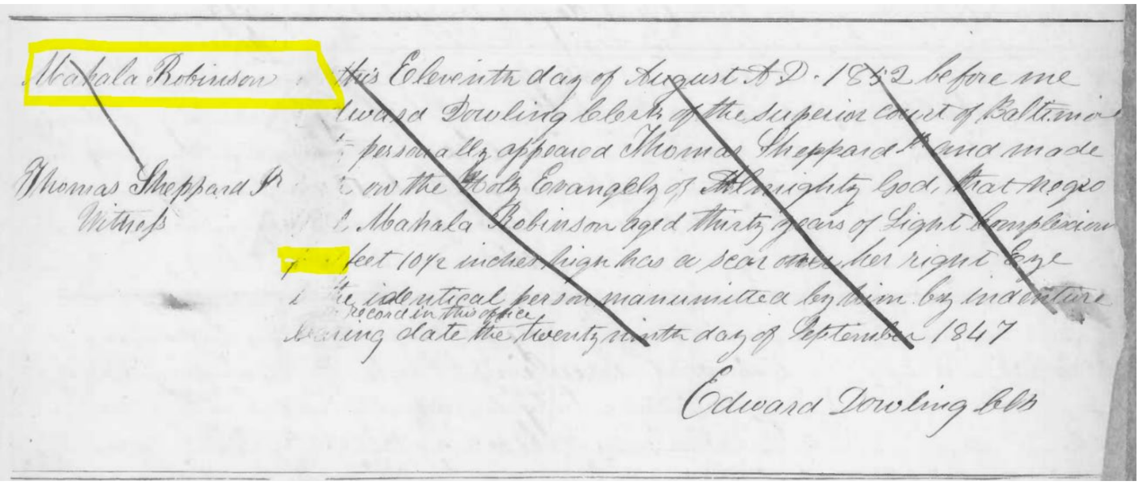
certificates of ... Date	Calculation Date_Final	certificates of freedom ... Date New
null	null	null
18,110,624	6/24/1811	null
18,110,624	6/24/1811	null
18,150,328	3/28/1815	null
18,370,710	7/10/1837	null
18,370,803	8/3/1837	null
18,261,028	10/28/1826	null
18,441,108	11/8/1844	null
18,190,127	1/27/1819	null
18,121,230	12/30/1812	null
null	null	null
18,090,219	2/19/1809	null
18,080,219	2/19/1808	null

Date_Final

```
IIF(ISNULL([Date New]),MAKEDATE(INT(left(Str([Date]),4)),int(mid(Str([Date]),5,2)),int(right(Str([Date]),2))),[Date New])
```

Height Feature -- This field is to indicate the height of the individual freed in feet and inches. There were some issues with this field mainly with some invalid values and data entry format errors.

- We used a delimiter option to parse the field into feet and inches columns separately.
- Data Entry issues -- Upon analyzing we found that there was a record with a height of 9 feet, with first name Abraham witnessed by James Wetheral, on checking the original CoF under series c931, we found that there was no CoF found for this person although there were other names found for Abraham. It mentions that under note that this person was manumitted but could not find the documents under Manumitted records as well
- Another issue with the illegible information on the original CoF as shown below for Mahala Robinson c165 page 35, it is also unclear as to why this whole certificate was striked out



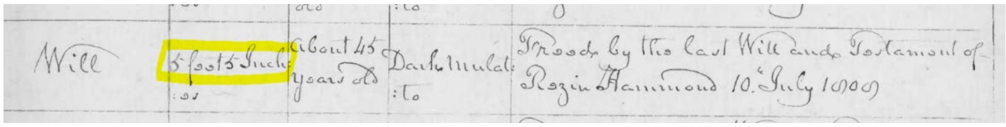
~~Mahala Robinson~~

~~this Eleventh day of August A.D. 1852 before me~~
~~[illegible]ward Dowling Clerk of the Superior Court of Baltimore~~
~~personally appeared Thomas Sheppard Sr and made~~
~~on the Holy Evangely of Almighty God, that Negro~~
~~Mahala Robinson aged thirty years of Light Complexion~~
~~7 feet 10½ inches high has a scar over her right eye~~
~~the identical person manumitted by him by indenture~~
~~recorded in this office~~
~~bearing date the twenty ninth day of September 1847~~

~~Thomas Sheppard Sr~~
Witness

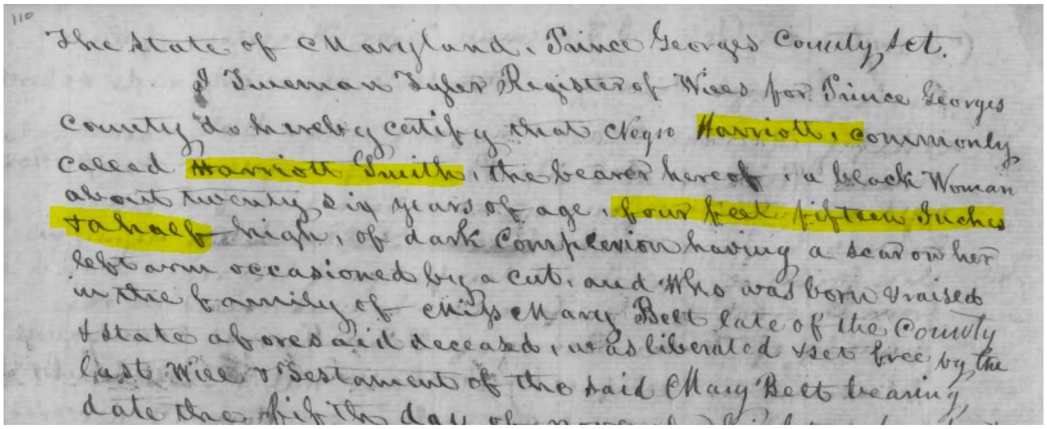
Edward Dowling Clk

- Other data capture issues were corrected by looking at the original scanned CoF as shown below: the height was noted as 5 5" which was in fact 5" 5' - 5 feet 5 inches



Will	5 feet 5 Inches	about 45 years old	Dark Mulatto	Freed by the last Will and Testament of Rezin Hammond 10. July 1809

- Cleaned up other issues related to invalid values for feet and inches like ",:,;, for inches, invalid values for inches more than 12 were also corrected, inches reported as fractions were also converted to its appropriate numeric value
- Found this issue about inches being reported as feet below:

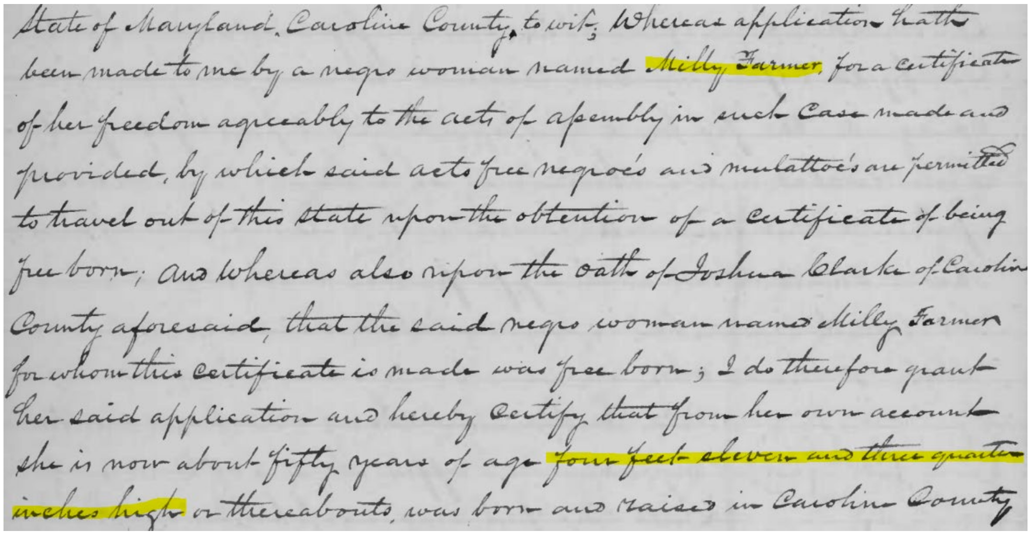


110

The State of Maryland, Prince Georges County Sct.

I Trueman Tyler Register of Wills for Prince Georges County do hereby certify that Negro Harriott, commonly called Harriott Smith, the bearer hereof, a black Woman about twenty six years of age, four feet fifteen Inches & a half high, of dark Complexion having a scar on her left arm occasioned by a cut, and who was born & raised in the family of Miss Mary Bell late of the County & state aforesaid deceased, was liberated & set free by the last Will & Testament of the said Mary Bell bearing date the fifth day of

- There is another entry where the height was mentioned as 4 feet 44.75 inches for Milly Farmer c477-2, page 200, whereas it was really 4 feet 11.75 inches as found below from the document:

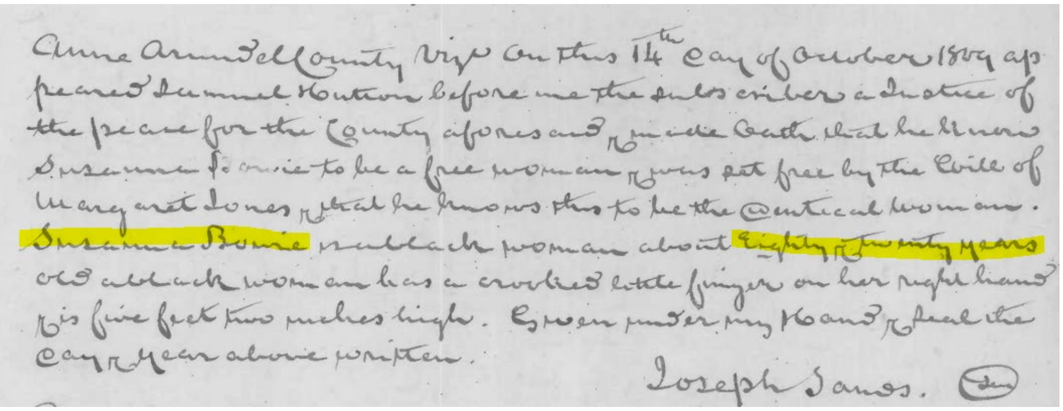


State of Maryland, Caroline County, to wit; Whereas application hath been made to me by a negro woman named Milly Farmer for a certificate of her freedom agreeably to the acts of assembly in such Case made and provided, by which said acts free negroe's and mulattoe's are permitted to travel out of this state upon the obtention of a Certificate of being free born; and whereas also upon the oath of Joshua Clarke of Caroline County aforesaid, that the said negro woman named Milly Farmer for whom this certificate is made was free born; I do therefore grant her said application and hereby Certify that from her own account she is now about fifty years of age four feet eleven and three quarter inches high or thereabouts, was born and raised in Caroline County

Age feature

- Age field was originally in the text type format, converted to number, and converted all the decimals which was entered as it is from the original document listed as months into a 12 month per year relative decimal value, for example, the original CoF noted the enslaved person as 18 months old, the dataset had this value as 0.18 under the age column which actually should be 1.5 years old

- For one case which was listed to be as 100 years old, upon checking the CoF original document, it's unclear as the document shows something like eighty & twenty years as highlighted below: This is also noted in the notes section as “Age given as eighty and twenty years. Could potentially be 28 years, not 100.”



Anne Arundel County Vizt On this 14th day of October 1809 appeared Samuel Houtton before me the subscriber a Justice of
the peace for the County aforesaid & made Oath that he knows
Susanna Bowie to be a free woman & was set free by the Will of
Margaret Jones & that he knows this to be the identical woman.
Susanna Bowie is a black woman about Eighty & twenty years
old a black woman has a crooked little finger on her right hand
& is five feet two inches high. Given under my Hand & seal the
day & year above written.

Joseph Janes.

- **Technical or tool issues**

There were issues related to limitations of tools used for example, MS Spreadsheet was unable to process dates prior to 1/1/1900 in a proper way which led us to use Tableau for date formatting. Open Refine had issues with the date formatting as well as it could not parse the date into date, month and year format from a character field.

4. Working in interdisciplinary teams

a. What disciplines are represented in your working group?

Archivists, Historians, Information technology professionals, computer scientists, researchers, students, professors.

b. What benefits/opportunities arising from interdisciplinarity have you noticed in your group?

We noticed a number of aha moments during the project as we were able to uncover certain insights unique to the collections that we would not have otherwise by virtue of working with multidisciplinary teams. These led to bonding between the team members from diverse groups who would have had no chance to meet and discuss these topics if not for this opportunity. This project also benefited the team members in understanding different perspectives of data and historical analysis.

c. What challenges arising from interdisciplinarity have you noticed in your group?

We had a good number of back and forth discussions between team members especially with the contextual background of the data being analysed as these collections were historical.

d. Are there any terminological or other confusions arising from working in interdisciplinary teams?

For non-historians and archivists, terms like ‘provenance’ were new and a good knowledge sharing experience. For non-IT team members, understanding how data could be sliced and diced to create visualizations that unearthed new insights was a good learning experience.

5. Moving forward

a. Do you see specific possibilities arising from your investigations?

- **Ideas for next steps**

As next steps, we have plans to understand more about linking the data collections so we could create networks of connected data elements that could create insights not seen or understood before.

- **Things you might want to try but haven't**

We would like to do more natural language processing analysis on certain features like notes and comments as the transcribers had entered valuable information into this feature. Also, more in-depth research on the reasons and rationale behind using different words to determine "Prior Status" and "Complexion".

6. Out-of-the-box

a. Please capture any ideas or discussions that arose from your group that do not fit into any of the previous sections.

Questions from the discussion related to the collection:

- Were the scars used for identification purposes in terms of determining which slave belongs to which owner?
- There was a spike in the number of Certificates of Freedom (CoF) from 1831 to 1832, then COF issues ceased around the year 1860. Is this because slavery was coming to an end?
- What is the significance of the differences in the prior status column? There are many records, including "Born Free", "Free Born", "Slave", "Enslaved", and "Descendant of a white female woman". Are there differences between these statuses?
- Skin complexion is very subjective, so how should we divide and classify the multiple different skin tones recorded?

Resources to read:

- A Guide to the History of Slavery in Maryland (MSA) (Read sections below)
  - III. Africans to African Americans
  - VI. Slavery and Freedom in t*he New Nation*