A Case Study in Creating Transparency in Using Cultural Big Data: The Legacy of Slavery Project

Ryan Cox
Legacy of Slavery
Maryland State Archives
Annapolis, USA
ryan.cox@maryland.gov

Sohan Shah
DCIC Center
University of Maryland
College Park, USA
sosha@terpmail.umd.edu

William Frederick
College of Information Studies
University of Maryland
College Park, USA
whfred@terpmail.umd.edu

Tammie Nelson
College of Information Studies
University of Maryland
College Park, USA
tnelson7@terpmail.umd.edu

Will Thomas
College of Information Studies
University of Maryland
College Park, USA
wthomas4@umd.edu

Greg Jansen
College of Information Studies
University of Maryland
College Park, USA
jansen@umd.edu

Noah Dibert
College of Information Studies
University of Maryland
College Park, USA
ndibert@umd.edu

Michael Kurtz
DCIC Center
University of Maryland
College Park, USA
mkclandcats@verizon.net

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

***Abstract*—The Maryland State Archives (MSA) and the Digital Curation Innovation Center (DCIC) of the University of Maryland's iSchool are collaborating on a digital project that utilizes digital strategies and technologies to create an in-depth understanding of the African-American experience in Maryland during the era of slavery. Utilizing crowdsourcing for transcription, data cleaning and transformation techniques, and data visualization strategies, the joint project team is creating new avenues for understanding the complex web of relationships that undergirded the institution of slavery. iSchool students, full participants on the project team, are learning digital curation and other technical skills while gaining insights into the multiple uses of how cultural Big Data can penetrate the past and illuminate the present.**

Keywords—Computational Thinking, Digital Curation, Computational Archival Science (CAS), Cultural Big Data, Legacy of Slavery

I. Background

For the past 18 years the Maryland State Archives has undertaken an extensive project documenting the stories of enslaved people and free Blacks in Maryland from colonial times to the latter part of the 19th century. Initially, case studies formed the core of the Legacy of Slavery program. Staff members used, among other research sources, runaway slave advertisements, census data, court records, and published materials from pre-Civil War Maryland with a connection to slavery.

As the project progressed the MSA project team began a decades-long initiative to digitize the Archives' holdings records relevant to the African-American experience. This has resulted in repurposing records originally created for other reasons. Examples include chattel records, property inventories contained in wills, and accounts of slave sales. Census records, slave statistics, manumissions and certificates of freedom contained relevant data that was captured in the digitization process. The move from analog records to relational databases resulted in a number of challenges for the staff.

The following sections of this paper explore in some depth the evolution of the Legacy of Slavery project from dispersed segments of analog records to a digital records collection. The application of digital curation and visualization tools and technologies has begun to bring together in a comprehensible manner the relationships and contexts that formed the reality of slavery in Maryland.

II. Analog Records to Relational Databases: Laying the Foundation

The mission of the Study of the Legacy of Slavery in Maryland research program is to utilize the records found in the holdings of the Maryland State Archives to help tell the story of enslavement and African American achievement in Maryland. To look at the records at the time of their creation, and by whom, has allowed staff to research how the institution of slavery was conducted, promoted, perpetuated and supported across the State. A number of record series provide a unique vantage point for researchers, archivists, educators, students and historians to look at the communities that were involved with enslavement, and the tools they created and used to maintain that institution. As archivists, our task is to provide access to the public records in our stacks for use by the patrons we serve. As techniques and technologies improve our finding aids and catalogs over time, it becomes the duty of the archivist to make the most of these advanced tools for simplified access to the documents. We do this so the user can

easily obtain the materials they need to tell their stories. This is not always a simple or conflict-free endeavor, but the keepers of the records must embrace these changes, and improve the experience of all who wish to read the public record.

A. The Process of Transforming Documents into Databases: An Overview

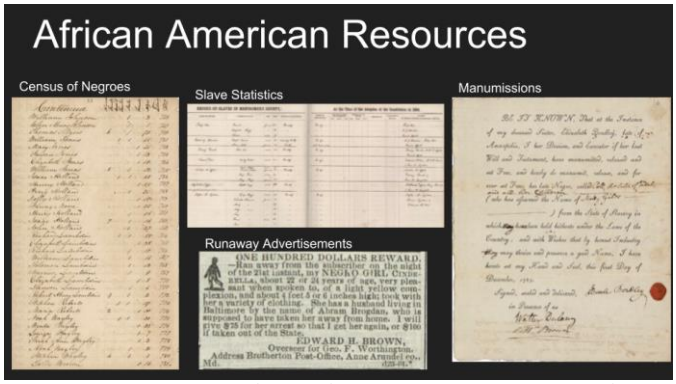

Figure 1. Sample images of some of the Maryland State Archive's African American Resources. Images from the left going clockwise: Talbot County Court Census of Negroes, 1832 Montgomery County Commissioner of Slave Statistics, 1867 Anne Arundel County Court Manumission, 1789 *Baltimore Sun* runaway advertisement, Dec. 1848.

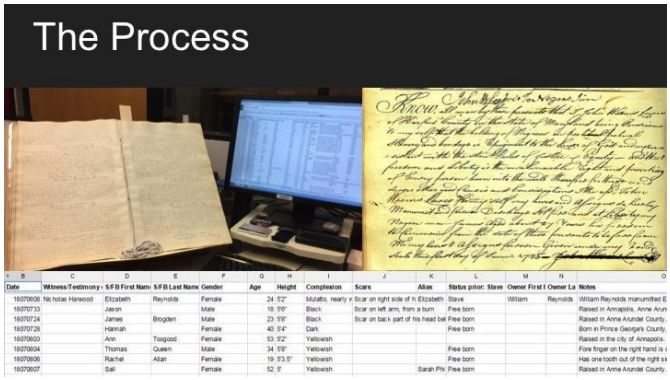

Figure 2. Transcription process of extracting information of primary document into database program, 2016. Image: Harford County Court Manumission, 1785.

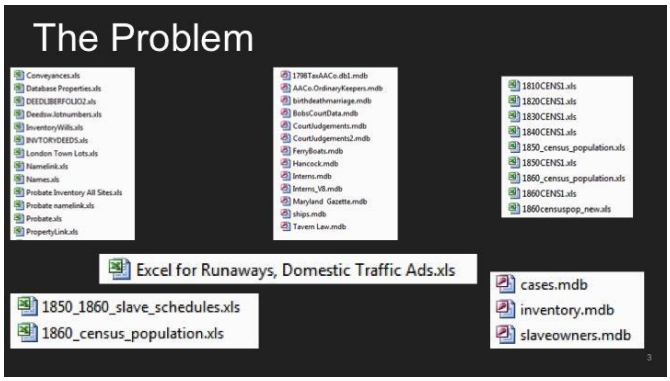

Figure 3. Selection of isolated spreadsheets based on the record series found in the Maryland State Archives mined by the Study of the Legacy of Slavery in Maryland research program

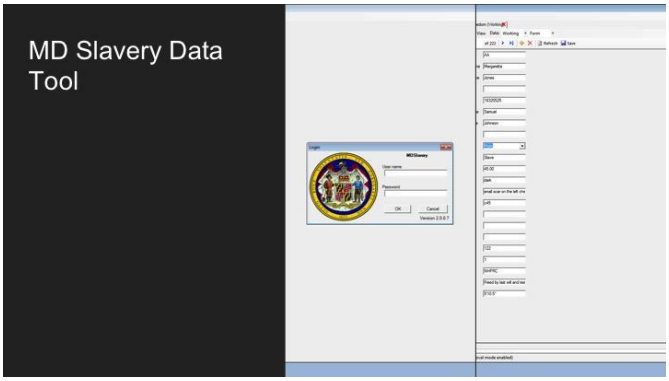

Figure 4. Sample entry from the mdslavery.exe data tool.

Figure 5. Interactive database found on http://slavery.msa.maryland.gov

B. Community Effort Spanning18 Years and 30 Record Series

Since 2001, our grant-funded program has been successful in mining the state's records for information about this population and highlighting the contributions that African Americans had on the development of the state of Maryland. During that time, our staff has looked at our holdings from 16 of Maryland's 23 counties and Baltimore City, and extracted information from over 30 record series with the assistance of dozens of archivists, staffers, interns and volunteers. This crowdsourcing initiative has allowed our team to look at the nature of enslavement across the complex and diverse areas of the state, in an attempt to uncover some of these marginalized communities.

The Maryland State Archives is legislatively mandated to obtain, house, and provide access to all government records of permanent value to the state. Across all branches of government, from state, county and municipal, we are charged as the official depository of government records.

The archival record of Maryland as a former slave state is distinct. Recovering the legacy of slavery depends upon the state archives more heavily than other such states for two reasons. Firstly, Maryland entered the Union as custodian of its own records. Maryland, at the beginning of the United States, was never a territory or district so there was no Federal body that kept Maryland's records as part of a treaty or military obligation. Secondly, as a former slave state that did not secede from the Union, it did not operate under military occupation following the war and there was no interruption of its existence as a state.

The understanding of who access to the archive was for – who the community was the Archives actually served – has changed at least twice during its existence. Before the Civil War and the adoption of the Fourteenth Amendment, the Archives envisioned a community which excluded Black people with few rights the institution need respect; the author of that view was Marylander 5RJHU Taney, Chief Justice of the Supreme Court and former slaveholder of Frederick County. Following the war, the state and institution willfully failed to embrace this expansion of its community and instead adopted de jure segregation. This meant that who the community was, changed again as Blacks and others marginalized by Jim Crow were forced to demand the right to recover their own narratives.

Some of the titles and descriptions of the record series are obvious about what they contain and who is represented in the pages. In regards to the creation of government record series that pertain to enslavement, Maryland is in a unique position to have some of these record series in our holdings. For example, the Census of Negroes from Somerset County on Maryland's lower Eastern Shore, compiled in 1832 as a result of the Nat Turner slave uprising in Northampton County, Virginia, is a census tabulation that indicated how many free blacks were living in the state at the time. This was an attempt by the state to have insight into a population believed to have contributed to the unrest, and was an unwanted, but motivating, demographic to the idea of resistance. To be able to view a Certificate of Freedom from Anne Arundel County in Central Maryland and gain insight into the process of manumission and the documentation needed by African Americans to verify their status as a free citizen is a remarkable lens into a community engaged in enslavement, and how those individuals creating the record chose to codify freedom. These are just a few of the aforementioned record series that we consult when researching the slave owning past of Maryland.

Other record series are less obvious and may have a title and description that may not completely encompass the information found in the pages. One would not necessarily think to look at a land record and find entries that note a sale of an enslaved individual, or the manumission of an enslaved person by their owner. The idea of repurposing record series to utilize their information in a new respect is an avenue we've taken to formulate a more robust analysis of enslavement in Maryland. To look at a probated inventory from Washington County in the 1820s in Western Maryland and view the information in a unique manner that differs from the reason the record was originally created is a necessarily computational challenge. While an inventory of an estate is a tax record about the deceased, our researchers use the record series as a tool to locate and identify those enslaved persons that had their names and values associated with an enslaver within a document. During this research, we are not necessarily interested in the total of a person's estate, just those that were enslaved and listed as a possession and parcel of the slaveholder's property. We would not be able to do this without the racial monikers imposed onto the name of the person, by the creator of the record. The perspective or idea of "who's history is represented in these pages, and who's history is this record series supposed to capture?" is an element that we keep in the back of our minds throughout our research.

C. On the Complexity of Creating an Online Search Tool

Over the years, we have built this institutional knowledge to recognize all of the avenues at our disposal that can lead to information about a community. Also over the years, new practices of mining the information, what to capture, record, and how to store the information has changed as our tools become more advanced and complex. The Maryland State Archives holdings stretch back over 380 years, however the finding aids, technological tools and instruments used to access this information are constantly changing. It was not until 2005 that our research archivists, with the assistance and expertise of our Information Technology department, realized the complex, tangled web we've created along the way. Over 200,000 bits of information were stored in various locations, file formats, completion stages, and imaging programs, making it impossible to cross-compare our findings between the tables within a unified search. We were unable to locate individuals that were represented on different record series. We knew that we wanted to manipulate, relate, evaluate, extract, filter and sort our findings, but we first needed to find a way to clean the data and combine the tables so that this analysis could be done effectively.

It was at this point that our Database Analysts began to look at our data, and clean it for integration. Thousands of hours over the course of months, dozens of eyes began to look at our information, come to understand the nature of the record series they derived from, and create an interface that we could combine, store, and add data. This was an important communication exercise with our IT, reference, appraisal and research departments coming together to see what information we had available, what we hoped the data would be able to do, and access the capabilities and limitations of the technological tools available to us. Concerns about the loss of the meaning of the data led to engaged conversations about what we hoped the tool could do for our patrons. During the transcription and mining process, we had to deal with the realities of how a "high-level transcription" would affect the ability of the data tool to generate accurate and complete results based on the user's search queries. Questions like "Do we define an abbreviated word? Do we keep the misspellings? How can we combine "City", "County", and "State" columns to display a "Location"? were some of the issues addressed. This is still an ongoing process. Digitization and data-mining are not always a direct avenue to making the record series accessible to our patrons in an intuitive manner.

We were able to successfully integrate all of the separate data tables and combine them into our data tool MDSlavery.exe. It is an executable file that uses Visual Basic to speak to our raw data on a Microsoft SQL Server, and display the results from a user's search inquiries on our online database found at slavery.msa.maryland.gov. To date, over 420,000 pieces of information can be accessed, sorted, filtered, and extracted with full citations, and where possible, digitized

images are also linked. We are very proud of our tool, but realize its limitations in displaying the information effectively.

Figure 6. shows the processing steps required to go from historical documents to datasets, represented as database tables stored in the Data Store. While the example starts with Runaway Slave Ads from historical newspapers, there are 16 distinct sources. These include:

- Accommodations
- Assessments
- Census 1830 & 1840
- Census 1850 & 1860
- Certificates of Freedom
- Chattels
- Deserters
- Domestic Traffic Ads
- Inventories
- Manumissions
- Pardon Dockets
- MD Penitentiary Records
- Runaway Ads
- Slave Jails
- Slave Schedules
- Slave Statistics

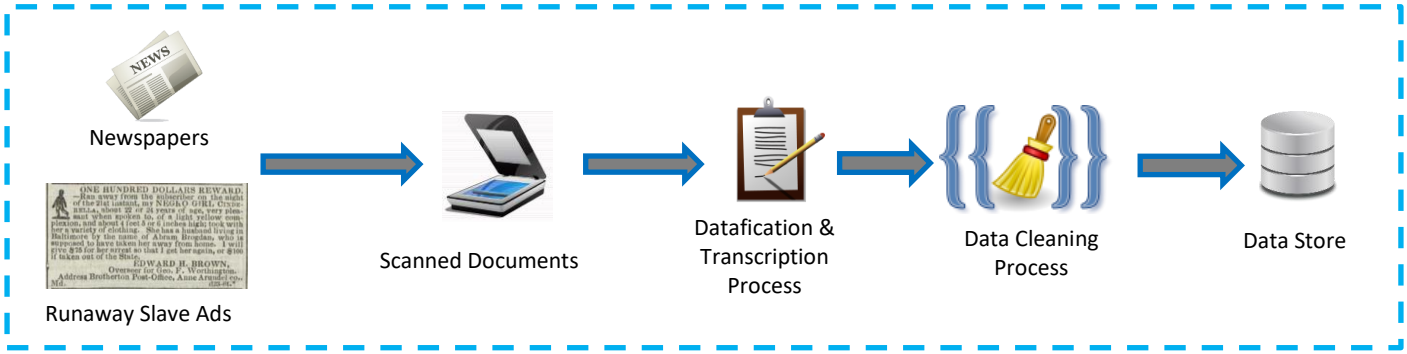

Figure 6. Processing pipeline from documents (Runaway Slave Ads in this example) to database tables stored in the Data Store.

D. Forming a Partnership with the University of Maryland's DCIC Center

This is where our colleagues from the University of Maryland's iSchool and the DCIC helped fill the void. We are grateful for their expertise in data analytics software, curation and information management. The school introduces the students not just to archival historical research and information technologies, but also utilization of the display tools and programs that find new mediums to illustrate our data in easy-to-read and useful ways. We approached this collaboration from two angles. The first was to introduce the student volunteers to our project, relevant record series, manuscript literacy, and transcription techniques. The second was to have the students and the faculty, develop ways to manipulate and display their findings from our records. This was a welcome challenge for the iSchool and appreciated effort on the part of the archivists given that there were limitations in how the MSA was able to display our information in-house. The results that we have been able to generate were incredibly useful. The students were able to illustrate the complex communal relationships between individuals, and relate this back to the record series where their names were found. Our work would not be as successful were it not for collaborations and continued relationships we have with our state's namesake institution, and we look forward to many projects in the future.

III. Data Engineering and Analytics: Building an Automated Repeatable Process

To provide additional context to this data section, this work reflects the emergence of "Collections as Data" in cultural heritage institutions over the last couple of years, where computational methods and tools are increasingly leveraged to enhance library and archives collections:

> *"Combined with an increasing flow of born digital items, digital library collections have come to represent a rich community resource for users... Yet a focus on replicating traditional ways of interacting with collections in a digital space does not meet the needs of the researcher, the student, the journalist, and others who would like to leverage computational methods and tools to treat digital library collections as data."*[1]

In the Digital Curation Innovation Center (DCIC) at the University of Maryland iSchool, we are developing a larger agenda to "infuse computational treatments" into archival science, as demonstrated in our work on Computational Archival Science (CAS) [1], defined as:

> *An transdisciplinary 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*

[1] Padilla, T, et al. (2016): "Always Already Computational: Library Collections as Data". See: https://www.imls.gov/grants/awarded/lg-73-16-0096-16

efficiency, productivity and precision in support of appraisal, arrangement and description, preservation and access decisions, and engaging and undertaking research with archival materials.

In this paper, we attempt to demonstrate the relevance of computational thinking concepts [2] through the application of data engineering and analytics processes applied to the Legacy of Slavery collection.

Transcribing digital records is a process requiring exceptional skill and attention to detail, but it is only half the battle won in making cultural big data collections accessible and thus more valuable. When working on a project to understand the information present in hundreds of thousands of digital records, we need a robust data pipeline to collect data, process it and store it in a database to make it analytics ready. We built a system that allows for creating such a data pipeline and automating the entire process. Our pipeline recommends using Apache Spark to process the data, which can increase processing speeds by up to 100x. While this case study is based on slavery data in Maryland, we built our pipeline to have a wider application, allowing slavery data from any state to be processed and analyzed using the same model.

A. *Extract Transform and Load (ETL)*

The challenge in dealing with historical collections from the 1800s is that the documents were written by different people, resulting in variations in handwriting and structure. Once data is transcribed, it needs to be consolidated, cleaned and stored in a database making it ready for analytics. We suggest using Apache Spark to execute these tasks.

- **Apache Spark.** An open source framework that allows for ingesting, processing and analyzing large volumes of data, both in real-time and in batch mode. Using Apache Spark, we can run Spark jobs against clusters in the cloud (or even on-premise) and leverage its in-memory processing engine which can result in increased processing speeds up to 100x.
- **Run ETL Jobs.** The Spark job consolidates data from the different data sources into one data frame that is now ready to be cleaned. Data is cleaned so that each column in the dataset is consistent with the schema and can be analyzed using visualization tools.
- **Data Repository (MongoDB).** Once the data is cleaned, it is stored in a MongoDB knowledge base. Since we are receiving data in batches, every new set of data will be appended to the last row in the database. Note that MongoDB is not the only option, and many other databases like PostgreSQL, MySQL, Aurora, DynamoDB and other data warehousing solutions like Redshift and Snowflake will satisfy the requirement.

B. *Data Visualization*

We imported the cleaned dataset into Tableau, a data visualization tool, to explore the data and extract insights from it. Tableau can be connected to a wide range of data sources, making it a great choice for visualizing the data. We created an interactive dashboard in Tableau which can be used to present findings from the data and can be uploaded online and made available to a wider audience, which is precisely what we did as we made our dashboard public through Tableau's website. Figures 8 and 9 show sample interactive dashboards that were built to explore the data and present statistics on the Certificates of Freedom documents.

What makes a data visualization tool like Tableau really valuable is its ability to refresh the dashboard when new data is added to the data source, without the need for us to re-create the visualizations.

C. *Data Pipeline*

Our data pipeline makes the process of data collection, processing and analytics both automated and repeatable. The flow of data in our pipeline is outlined in Figure 7. below.

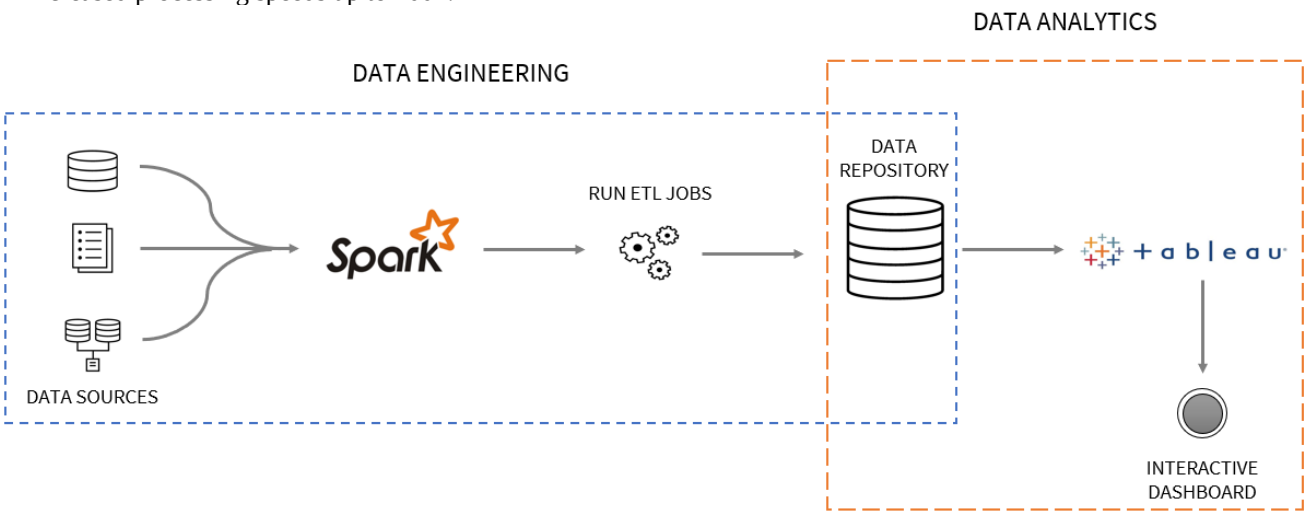

Figure 7. Data pipeline depicting the flow of data through the data engineering and analytics phases.

With hundreds of thousands of documents to transcribe, the transcription must be done in batches. Without automation, we would have spent months executing the same task for every batch of transcribed data. Instead, we created an integrated system comprised of software programs configured as spark jobs, a database that allows rows to be appended, and a visualization tool that can refresh the visualizations to reflect newly transcribed data. Our system completes each iteration of this process in minutes. There is a one-time effort of writing code to create the spark job, connect Tableau to the data source and create the visualizations for the interactive dashboard. For every future iteration, the Spark job will process the data and store it in the database and Tableau will read in the new set of data and automatically update the visualizations.

D. Data Analytics

The following two figures illustrate the power of creating an interactive dashboard to visualize and interact with the historical data at scale.

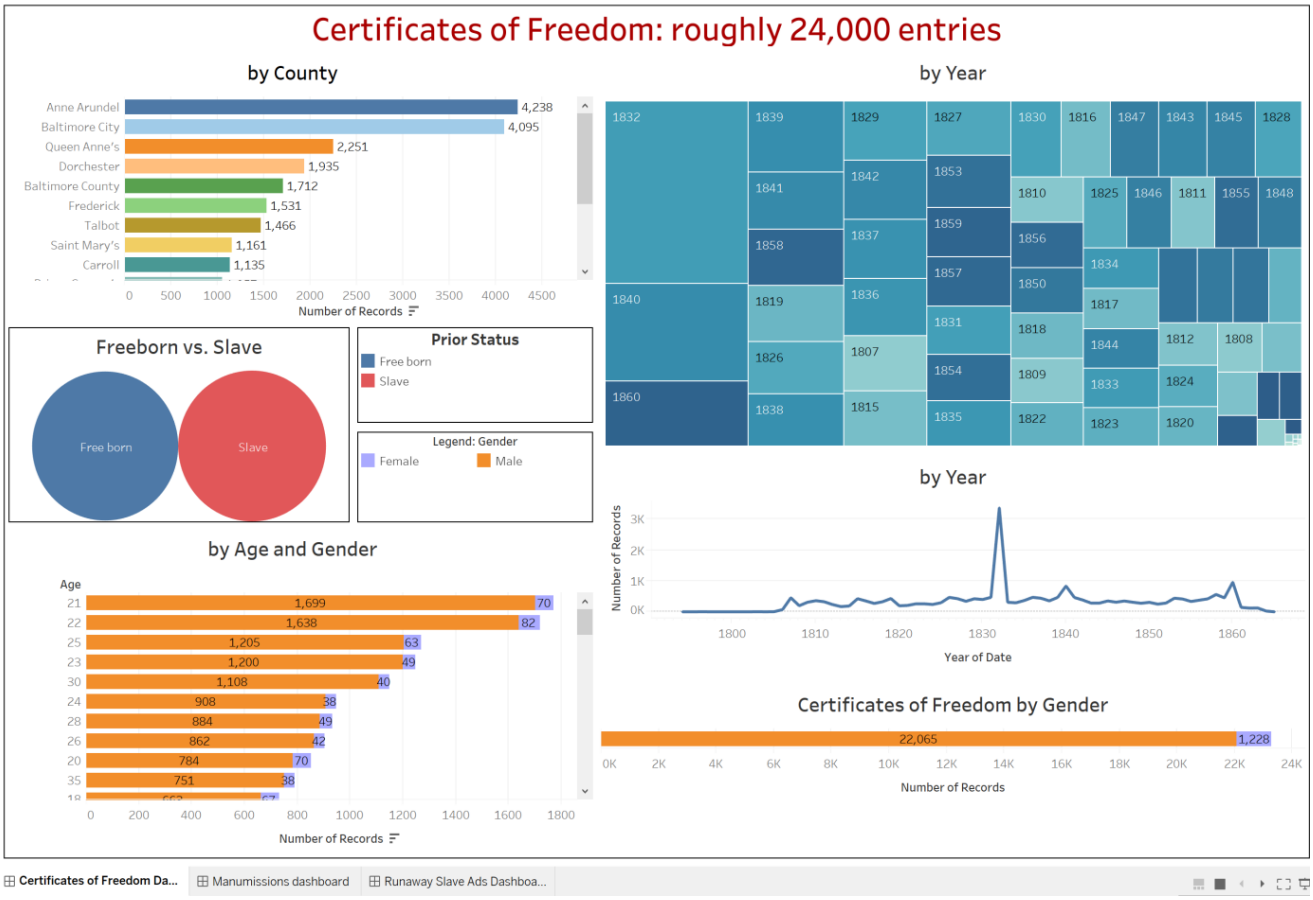

Figure 8. Certificates of Freedom Dashboard when no filter was applied. This dashboard contains data from approximately 24,000 Certificates of Freedom documents issued to then enslaved persons by their slave owners. [3]

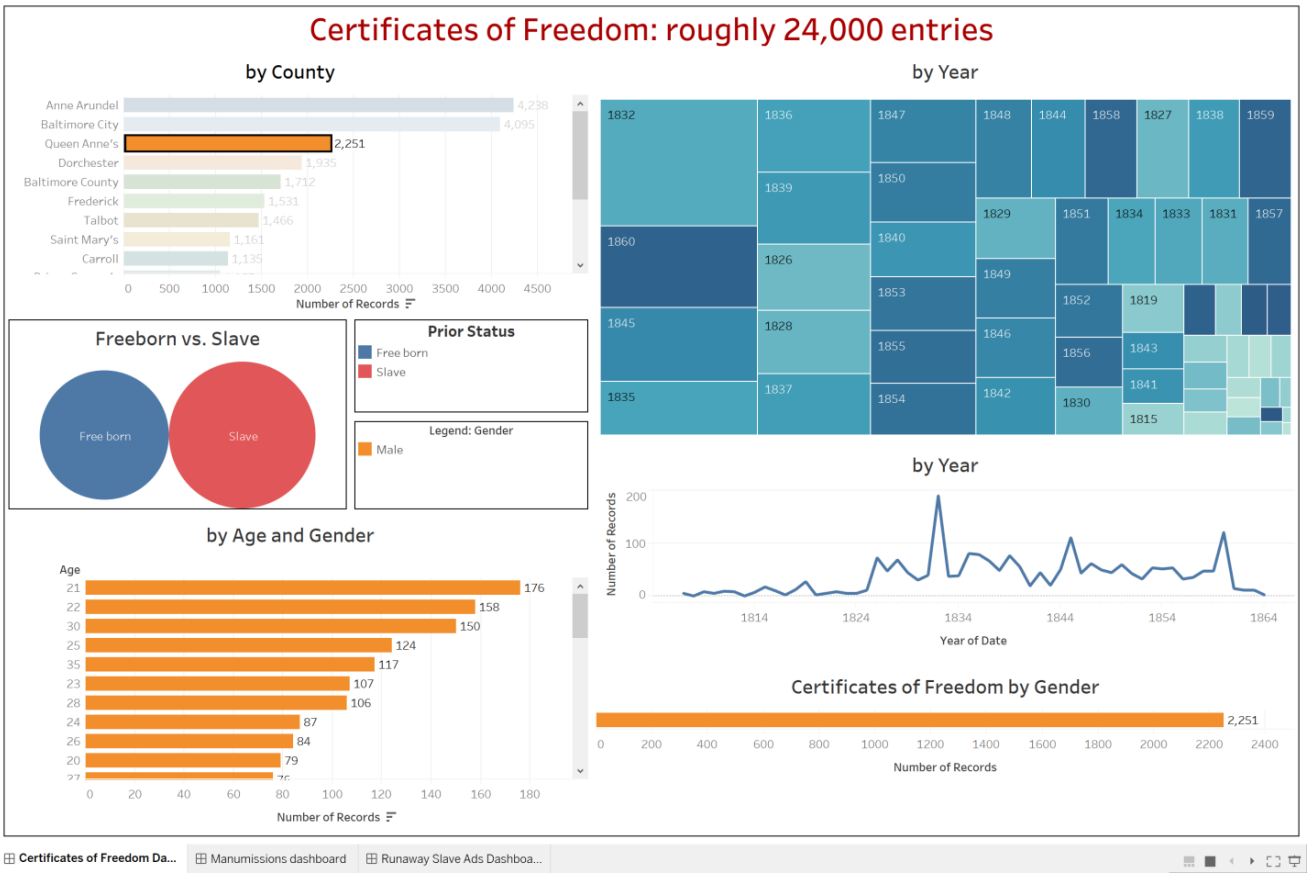

Figure 9. This demonstrates the interactive capability of the dashboard. The underlying data is the same as that in Figure 8. However, this dashboard is only filtered for Queen Anne's County. [3]

IV. Conclusions and Future Work

The Legacy of Slavery of project has demonstrated a proof of concept in creating a data pipeline that can collect data, process and store it, and make the data analytics ready. There are several elements necessary to make this a fully developed, mature system. The project team intends to take the prototype developed and apply it to all the relevant data collected and digitized by the Maryland State Archives. This will require further testing to ensure that the data pipeline can effectively manage a steady flow of processed data ready for analysis of the relationships that formed the network of the enslaved peoples, free Blacks, and society at large during the era of slavery in Maryland. A major goal is making the system robust enough to deal with the data from the Maryland political jurisdictions whose data has not been transcribed or processed through the data pipeline. This is effort will require a significant level of resources from the MSA.

In addition, as the system envisioned is fully deployed the next step is to share the results with other similar projects underway in the United States. The DCIC and the MSA will host a conference to share the results of this project with potential partners.

Acknowledgment

We wish to acknowledge major funding from the NSF "Brown Dog" project (NSF Cooperative Agreement ACI-1261582), as well as IMLS funding [LG-71-17-0159-17], which is helping us port Fedora on top of DRAS-TIC, an open-source platform (Digital Repository At Scale - That Invites Computation). More details at:

http://dcic.umd.edu/about-us/infrastructure/.

References

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

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

[3] W. Frederick (2018). MD State Archives Legacy of Slavery. Available: https://public.tableau.com/profile/william.frederick#!/vizhome/MDStateArchivesLegacyofSlavery/CertificatesofFreedomDashboard