

# *the* Taming of the Unstructured Data

Rayvn Manuel



Smithsonian  
*National Museum of African American  
History and Culture*

# Agenda

---

- About....
- Brief History of the Bureau
- Congressional Mandate
- Datasets
- Searchability
- Next Steps & //TODOs
- Qs?



# About...



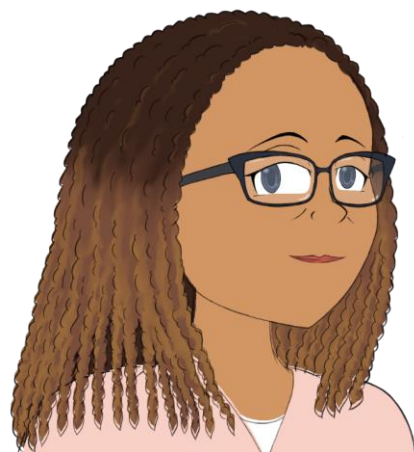
Smithsonian  
*National Museum of African American  
History and Culture*

# Me

## *& why am I standing up here speaking to you*

---

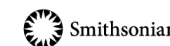
Sworn in on 9 January 2017, I serve as the **Senior Application Developer and DevOps engineer** for NMAAHC.



I am the principle developer **responsible for the planning, design and implementation** of some of the hands-on interactives scattered throughout the museum.

**Taming of the Unstructured Data** is a tale of transforming two (2) unruly and somewhat feral datasets into a searchable, web-based application and kiosk interactive.

The interactive was **designed** for the enjoyment of the museum's visitors as well as a research tool for genealogists and historians to explore the digitized archive of the Freedmen's Bureau Records.

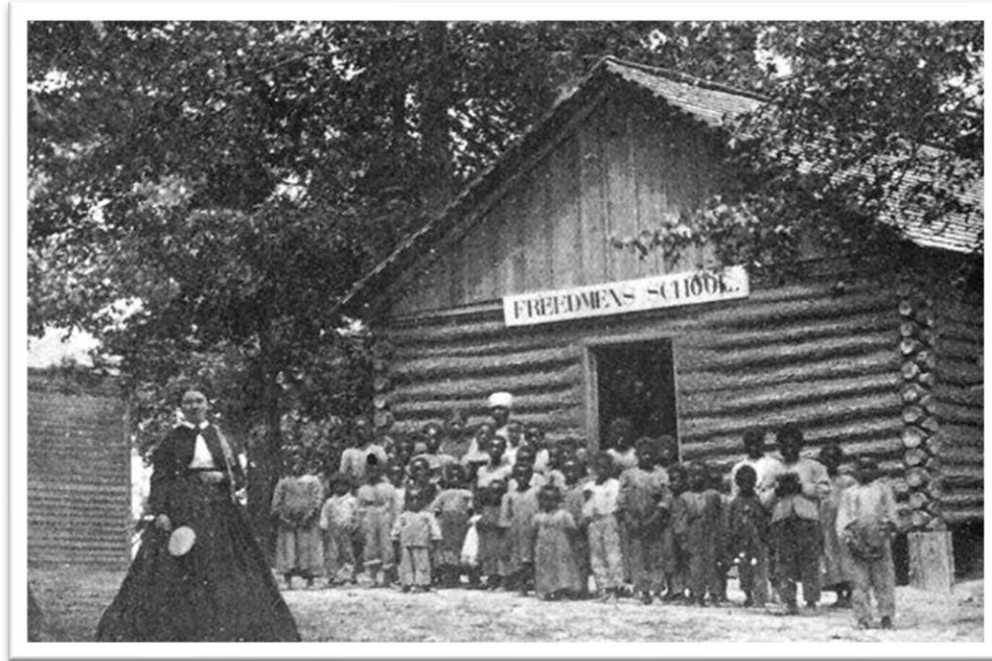


# Brief History of the Bureau



Smithsonian  
*National Museum of African American  
History and Culture*

# the “Bureau”



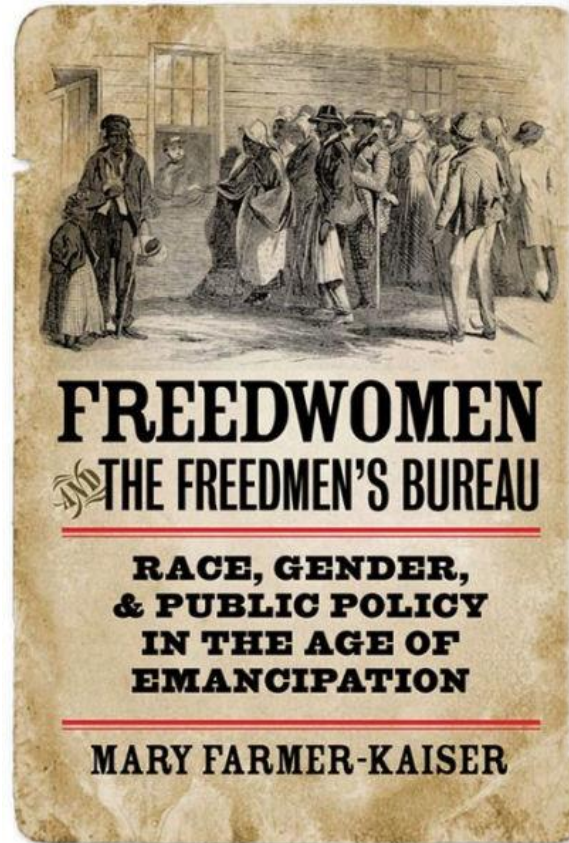
<https://images.app.goo.gl/t56fy6m1Y5diRuKY9>

The Bureau of Refugees, Freedmen and Abandoned Lands (Freedmen's Bureau) was established in 1865 by Congress to *help millions of emancipated slaves and poor whites in the South in the aftermath of the Civil War.*

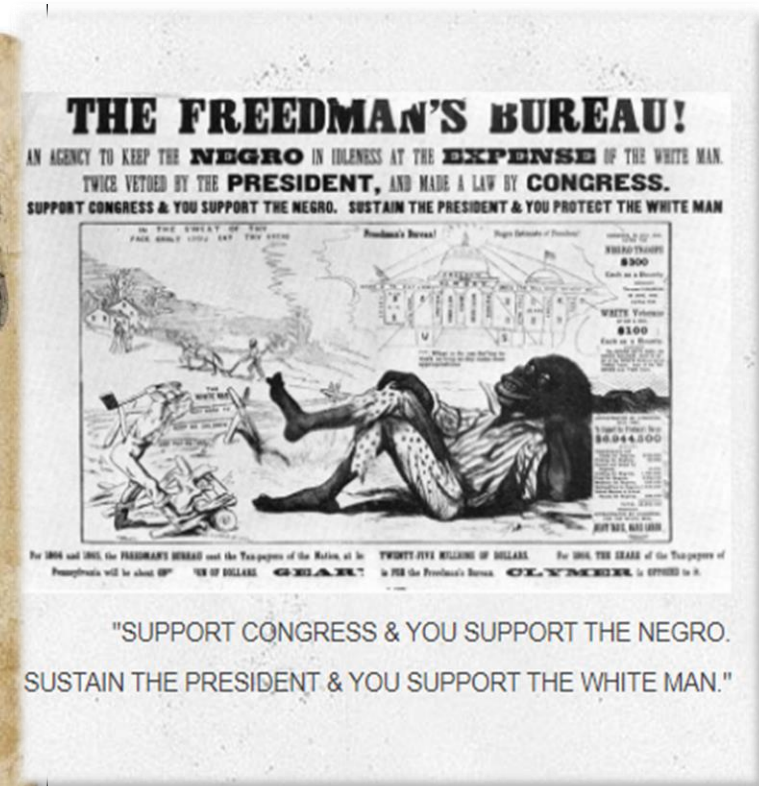
During its years of operation, the Freedmen's Bureau fed millions of people, built hospitals and provided medical aid, negotiated labor contracts for ex-slaves and settled labor disputes. It also helped former slaves legalize marriages and locate lost relatives, and assisted black veterans.

## the “Bureau” cont’d

Intended to be a temporary organization to last the duration of the war and one year afterward (1865 - 1872), the bureau failed to provide long-term protection or ensure any real measure of racial equality. In the summer of 1872, Congress dismantled the bureau *due to a shortage of funds and personnel, along with the politics of race and Reconstruction.*

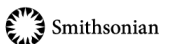


<https://images.app.goo.gl/GLBSehAebkwRiwrK8>



<https://images.app.goo.gl/FkvQEVENAtKvTGJF6>

NATIONAL  
MUSEUM of  
AFRICAN  
AMERICAN  
HISTORY &  
CULTURE





# Congressional Mandate



Smithsonian  
*National Museum of African American  
History and Culture*

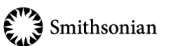


# *the* Freedmen's Bureau Preservation Act

---

In 2000, the U.S. Congress passed the Freedmen's Bureau Preservation Act, which *directed the National Archivist to preserve the extensive records of the Bureau on microfilm, and work with educational institutions to index the records.*

Announced on June 19th 2015, the Freedmen's Bureau Project was created as a partnerships between **FamilySearch International** and the **National Archives and Records Administration (NARA)**, the **National Museum of African American History and Culture**, the **Afro-American Historical and Genealogical Society (AAHGS)**, and the **California African American Museum**.



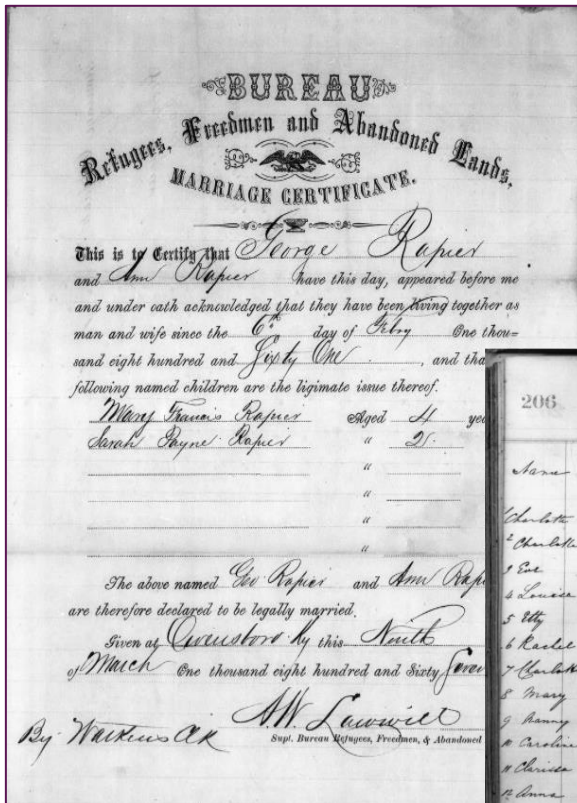
# Datasets



Smithsonian  
*National Museum of African American  
History and Culture*

# the Freedmen's Bureau Records

<https://images.app.goo.gl/hob3RjmpcCd5bhSWA>



The records constitute a major source of documentation on the operations of the Bureau, political and social conditions in the Reconstruction Era, and one of the few recorded genealogies of freed people.

NARA provided access to the microfilms to the Smithsonian and the Church of Latter Day Saint's FamilySearch to facilitate and fulfill the record preservation and indexing mandate.

Name	Age	Sex	Color	Religion	Education	Occupation	Remarks
Sarah	Master	Black	Free	Methodist	None	Washing	Capable of doing work
Charlotte	10	Female	Black	Methodist	None	Washing	Capable of doing work
1. Ene	10	Female	Black	Methodist	None	Washing	Capable of doing work
2. Louise	10	Female	Black	Methodist	None	Washing	Capable of doing work
3. Elly	10	Female	Black	Methodist	None	Washing	Capable of doing work
4. Rachel	10	Female	Black	Methodist	None	Washing	Capable of doing work
5. Mary	10	Female	Black	Methodist	None	Washing	Capable of doing work
6. Mary	10	Female	Black	Methodist	None	Washing	Capable of doing work
7. Mary	10	Female	Black	Methodist	None	Washing	Capable of doing work
8. Mary	10	Female	Black	Methodist	None	Washing	Capable of doing work
9. Mary	10	Female	Black	Methodist	None	Washing	Capable of doing work
10. Mary	10	Female	Black	Methodist	None	Washing	Capable of doing work

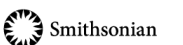
<https://images.app.goo.gl/BMF2H7ZifdzLFBUE6>

State of North Carolina  
Robeson County.

This Indenture made the fourteenth  
day of September A.D. 1865 between  
James Sinclair Agent of the Bureau of  
Refugees, Freedmen, and Abandoned Lands  
and therefore legal Guardian of Colored  
Orphans) of the one part and Archibald  
McMillan of the other part  
County and State on the other part  
Witnesseth; that the said James  
Sinclair, Agent, doth put, place, and  
bind unto the said Archibald McMillan

<https://nmaahc.si.edu>

NATIONAL  
MUSEUM of  
AFRICAN  
AMERICAN  
HISTORY &  
CULTURE



the

[illegible]

**DataSet #1:** The **Smithsonian Transcription center** is a crowdsourcing project that allows volunteers [\[shameless CALL TO ACTION plug\]](#) to access microfilmed records which need to be transcribed into digital content.

To include the Bureau records, the Transcription center is tasked with digitizing all of the Institution's historic documents and collection records.

- Data captured & transcribed: *The complete record.*
- **Digital Storage:** Internal Enterprise Digital Asset Network [EDAN]
- **Digital Image Storage:** Internal Digital Asset Management System [DAMS]



# the Bureau Records | FamilySearch Intl

**DataSet #2:** A nonprofit family history organization run by the Church of Jesus Christ of the Latter-day Saints, FamilySearch International relies on its members to transcribe genealogy records as a free service to their members and service subscribers.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
FB_DATE	FB_DAY	FB_MONTH	FB_LOCATION	EVENT_TYPE	FB_YEAR	FS_BATCH_LOCALITY	FS_DIGITIZATION_FILM_ID	FS_IMAGE_ID	FS_IMAGE_TYPE	FS_IMAGE_ID	FS_IMAGE_ID	FS_LANGUAGE	FS_RECORDS_SORT	FS_RECORDS_SORT	FS_RECORDS_SORT	FS_RECORDS_SORT
15 May 1867	15	May	Calhoun, Georgia, U Residence		1867	United States	4139606	1574230	168		004139606_00168	English	15-0061; 2 004139606 00000000C	Unit		
29 Mar 1867	29	Mar	Albany, Dougherty, Residence		1867	United States	4139606	1574230	157		004139606_00157	English	15-0061; 2 004139606 00000000C	Unit		
			Georgia, United Stal Unspecified			United States	4139606	1574230	162	Blocked	004139606_00162	English	15-0061; 2 004139606 00000000C	Unit		
			Georgia, United Stal Unspecified			United States	4139606	1574230	147	No Extractable	004139606_00147	English	15-0061; 2 004139606 00000000C	Unit		
			Georgia, United Stal Unspecified			United States	4139606	1574230	224	Blocked	004139606_00224	English	15-0061; 2 004139606 00000000C	Unit		
11 Mar 1867	11	Mar	Savannah, Chatham Residence		1867	United States	4139606	1574230	225		004139606_00225	English	15-0061; 2 004139606 00000000C	Unit		
13 Sep 1866	13	Sep	Atlanta, Fulton, Gec Residence		1866	United States	4139606	1574230	225		004139606_00225	English	15-0061; 2 004139606 00000000C	Unit		
31 May 1867	31	May	Marietta, Cobb, Gec Residence		1867	United States	4139606	1574230	216		004139606_00216	English	15-0061; 2 004139606 00000000C	Unit		
02 May 1867	2	May	McDonough, Henry, Residence		1867	United States	4139606	1574230	218		004139606_00218	English	15-0061; 2 004139606 00000000C	Unit		
20 May 1867	20	May	Marion, Georgia, Un Residence		1867	United States	4139606	1574230	220		004139606_00220	English	15-0061; 2 004139606 00000000C	Unit		
23 May 1867	23	May	Atlanta, Fulton, Gec Residence		1867	United States	4139606	1574230	231		004139606_00231	English	15-0061; 2 004139606 00000000C	Unit		
23 May 1867	23	May	Atlanta, Fulton, Gec Residence		1867	United States	4139606	1574230	231		004139606_00231	English	15-0061; 2 004139606 00000000C	Unit		
23 May 1867	23	May	Atlanta, Fulton, Gec Residence		1867	United States	4139606	1574230	231		004139606_00231	English	15-0061; 2 004139606 00000000C	Unit		
01 Apr 1867	1	Apr	Atlanta, Fulton, Gec Residence		1867	United States	4139606	1574230	231		004139606_00231	English	15-0061; 2 004139606 00000000C	Unit		
31 May 1867	31	May	Savannah, Chatham Residence		1867	United States	4139606	1574230	200		004139606_00200	English	15-0061; 2 004139606 00000000C	Unit		
28 May 1867	28	May	La Grange, Troup, G Residence		1867	United States	4139606	1574230	206		004139606_00206	English	15-0061; 2 004139606 00000000C	Unit		
			Georgia, United Stal Unspecified			United States	4139606	1574230	171	Blocked	004139606_00171	English	15-0061; 2 004139606 00000000C	Unit		
01 May 1867	1	May	Columbus, Muscogee Residence		1867	United States	4139606	1574230	177		004139606_00177	English	15-0061; 2 004139606 00000000C	Unit		
Nov 1866	31	Nov	Columbus, Muscogee Residence		1866	United States	4139606	1574230	179		004139606_00179	English	15-0061; 2 004139606 00000000C	Unit		
01 May 1864	1	May	Columbus, Muscogee Residence		1864	United States	4139606	1574230	179		004139606_00179	English	15-0061; 2 004139606 00000000C	Unit		
01 May 1867	1	May	Georgia, United Stal Residence		1867	United States	4139606	1574230	183		004139606_00183	English	15-0061; 2 004139606 00000000C	Unit		
			Georgia, United Stal Unspecified			United States	4139606	1574230	186	Blocked	004139606_00186	English	15-0061; 2 004139606 00000000C	Unit		
31 May 1867	31	May	Macon, Monroe, Ge Residence		1867	United States	4139606	1574230	212		004139606_00212	English	15-0061; 2 004139606 00000000C	Unit		
31 May 1867	31	May	Macon, Monroe, Ge Residence		1867	United States	4139606	1574230	212		004139606_00212	English	15-0061; 2 004139606 00000000C	Unit		
			Georgia, United Stal Unspecified			United States	4139606	1574230	43	Blocked	004139606_00043	English	15-0061; 2 004139606 00000000C	Unit		
			Georgia, United Stal Unspecified			United States	4139606	1574230	49	Blocked	004139606_00049	English	15-0061; 2 004139606 00000000C	Unit		
			Georgia, United Stal Unspecified			United States	4139606	1574230	54	Blocked	004139606_00054	English	15-0061; 2 004139606 00000000C	Unit		
13 May 1867	13	May	Albany, Dougherty, Residence		1867	United States	4139606	1574230	111		004139606_00111	English	15-0061; 2 004139606 00000000C	Unit		
31 May 1867	31	May	Albany, Dougherty, Residence		1867	United States	4139606	1574230	113		004139606_00113	English	15-0061; 2 004139606 00000000C	Unit		
			Georgia, United Stal Unspecified			United States	4139606	1574230	114	Blocked	004139606_00114	English	15-0061; 2 004139606 00000000C	Unit		
01 Oct 1866	1	Oct	Atlanta, Fulton, Gec Residence		1866	United States	4139606	1574230	118		004139606_00118	English	15-0061; 2 004139606 00000000C	Unit		
01 Feb 1867	1	Feb	Atlanta, Fulton, Gec Residence		1867	United States	4139606	1574230	123		004139606_00123	English	15-0061; 2 004139606 00000000C	Unit		
16 Mar 1867	16	Mar	Atlanta, Fulton, Gec Residence		1867	United States	4139606	1574230	123		004139606_00123	English	15-0061; 2 004139606 00000000C	Unit		
23 Apr 1867	23	Apr	Georgia, United Stal Residence		1867	United States	4139606	1574230	123		004139606_00123	English	15-0061; 2 004139606 00000000C	Unit		
28 May 1867	28	May	Augusta, Columbia, Residence		1867	United States	4139606	1574230	102		004139606_00102	English	15-0061; 2 004139606 00000000C	Unit		
01 May 1867	1	May	Georgia, United Stal Residence		1867	United States	4139606	1574230	105		004139606_00105	English	15-0061; 2 004139606 00000000C	Unit		
			Georgia, United Stal Unspecified			United States	4139606	1574230	143	Blocked	004139606_00143	English	15-0061; 2 004139606 00000000C	Unit		
01 Sep 1866	1	Sep	Georgia, United Stal Residence		1866	United States	4139606	1574230	149		004139606_00149	English	15-0061; 2 004139606 00000000C	Unit		
01 Mar 1867	1	Mar	Georgia, United Stal Residence		1867	United States	4139606	1574230	149		004139606_00149	English	15-0061; 2 004139606 00000000C	Unit		

- Data captured & transcribed:  
*Demographic Info*
- Digital Content Storage: Internal Database
- Digital Image Storage: Web Accessible URL



# Data Consolidation Challenges

---

**FamilySearch Data:** Data received from FamilySearch was captured & stored as thirteen (13) comma-delimited files with a combined record count of six (6) million entries.

Unedited and viewed in spreadsheet format uncovered that across all the data not one collection set shared a single piece of common data (::groan::: no primary key).

Also revealed - a number of "what not to do" with tabular & normalized data

- Blank cells & null values versus filler zeros
- Transcription formatting wonkiness
- Multiple pieces of information captured in a single cell
- Problematic field names
- Special characters mixed with the data
- Inclusion of metadata information
- Varied date formats

**SI Transcription Center Data:** Data received from SI is served from EDAN in JSON format but not accessible via client URL.



# Searchability



Smithsonian  
*National Museum of African American  
History and Culture*



# FB Digital Record Search



[Learn more about this project at the SI Transcription Center](#)

Egerton. Jos. J. | C.149  
Eaton. W.A. | W.356-H.249  
Eddy Jas | D.118  
Elyard Danl. | W.361  
Evans Amy | W.386.-387  
Epps James | W.248  
Evans Edward | N.85  
Evans Miss | C.169

Records of the Assistant  
Commissioner for the State of  
North Carolina, Bureau of  
Refugees, Freedmen and  
Abandoned Lands, 1865-1870

Registers of letters received, name index (4) to register 2

Danl Elyard

**RECORDS OF:**

Assistant Commissioner  
| North Carolina

**PUBLICATION NUMBER:**

M843

**ASSET NAME:**

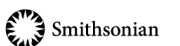
NMAAHC-004567389\_00223

The Project received financial support from a donor to create a place within the museum – Family History Center – where visitors could sit and search the records for family members who may have been displaced during this era.

This required that the two datasets be combined and the records be matched to cross-reference (as much as possible) each other.

The purpose of the combined datasets was to provide researchers & visitors with a richer and more complete representation of the records and family history.

Additional Benefit: Mapping the demographic information with the complete transcript and stored images could possibly uncover trends during the Reconstruction Era e.g. diseases among the population at that time.



# Feature set: Adding Findability



## Supporting Technology:

- SOLR 7.4 with custom schema
- Application Language: Angular 7.x

Extensive understanding of the content in order to make judgement calls during cleanup: Consultation with Genealogist & SME Curator

## Dream-worthy amount of Data Clean-up

*Sifting through each record to verify data consistency; did I mention there are 6 million records?*

- Splitting multiple pieces of data
- Standardizing column headers\*\* & making these the indexable fields in SOLR
- Creating a composite key: autoincremented ID column + contrived CollectionID
- Removing beginning spaces



# Made it so...

[illegible]

```
<!-- NMAAHC - Freedmens Bureau Records Fields -->

<field name="FB_FULL_NAME" type="string" multiValued="true" indexed="true" stored="true"/>
<field name="FB_FIRST_NAME" type="string" multiValued="true" indexed="true" stored="true"/>
<field name="FB_LAST_NAME" type="string" multiValued="true" indexed="true" stored="true"/>
<field name="FB_OCCUPATION" type="text_nmaahc" multiValued="true" indexed="true" stored="true"/>
<field name="FB_RACE" type="text_nmaahc" multiValued="true" indexed="true" stored="true"/>
<field name="FB_GENDER" type="string" multiValued="true" indexed="true" stored="true"/>
<field name="FB_YEAR" type="string" multiValued="true" indexed="true" stored="true"/>
<field name="FB_DATE" type="tdate" multiValued="true" indexed="true" stored="true"/>
<field name="FB_DAY" type="string" multiValued="true" indexed="true" stored="true"/>
<field name="FB_MONTH" type="tdate" multiValued="true" indexed="true" stored="true"/>
<field name="FB_COUNTY" type="string" multiValued="true" indexed="true" stored="true"/>
<field name="FB_LOCATION" type="string" multiValued="true" indexed="true" stored="true"/>
<field name="FB_IMAGE_APID" type="string" multiValued="false" required="true" indexed="true"
stored="true"/>
<field name="FB_IMAGE_ID" type="string" multiValued="true" indexed="true" stored="true"/>
<field name="FS_IMAGE_PAL" type="string" multiValued="true" indexed="true" stored="true"/>
<field name="text_g" type="text_general" indexed="true" stored="true" />
<field name="text_s" type="string" indexed="true" stored="true" />

<field name="BR_FULL_NAME" type="string" multiValued="true" indexed="true" stored="true"/>
<field name="BR_FIRST_NAME" type="string" multiValued="true" indexed="true" stored="true"/>
<field name="BR_LAST_NAME" type="string" multiValued="true" indexed="true" stored="true"/>
<field name="GR_FULL_NAME" type="string" multiValued="true" indexed="true" stored="true"/>
<field name="GR_FIRST_NAME" type="string" multiValued="true" indexed="true" stored="true"/>
<field name="GR_LAST_NAME" type="string" multiValued="true" indexed="true" stored="true"/>
<field name="FB_MARRIAGE_PLACE" type="string" multiValued="true" indexed="true" stored="true"/>
```

# Next Steps & //TODOs



Smithsonian  
*National Museum of African American  
History and Culture*

# BugFixes, Refactors & Enhancements *oh my*

---

## *NEXT Steps*

Creating an open-source API to make the csv records & EDAN data available to the public without exposing internal bits.

- **Use case:** Regional Chamber of Congress or Museum who wants limit records for a specific state.
- **Use case:** Family historian who wants to limit records for a specific surname
- And **other use cases** that I cannot fathom

## *//TODOs 2 GET /there*

### BugFixes

- Address performance issues
- Search is not working
- Wonky mobile view
- Adding SSL certs
- Additional security to SOLR

### Refactor

- Change up back-end [Nodejs + GraphQL]
- Adding a back-end data store [redis/mongoDB]

### Enhancements

- Pretty-fying UI while maintaining branding
- Adding more whiz-banginess
- Add keyword search



# Questions??



Smithsonian  
*National Museum of African American  
History and Culture*

# *the* Taming of the Unstructured Data

Rayvn Manuel



Smithsonian  
*National Museum of African American  
History and Culture*