



Looking for ‘Trouble’

Building a Description Audit Toolkit for Archival Metadata

Noah Huffman and Meghan Lyon
ALA Core Metadata Interest Group
July 28, 2021

Ongoing projects at Rubenstein Library

Guiding Principles for Description



-  *We will use inclusive and accessible language when describing the people represented by or documented in our materials. We commit to continually educate ourselves on evolving language and practices of inclusivity and accessibility.*
-  *We will prioritize facts and accuracy, and resist editorializing, valorizing, or euphemistic narratives or phrases in our description. This includes a commitment to revisit and revise our past description.*

How We Describe (Style guide for inclusive description)

Establish a Reparative Cataloging and Processing Queue

-  Re-process / re-describe legacy archival collections

Audit, analyze, and remediate description

-  Identify harmful language in catalog records / finding aids
-  Identify collections where BIPOC voices and experience have been ignored
-  Develop remediation work plan

Problem(s)

1. Catalog records and finding aids contain language that is:
 - Racist
 - Sexist
 - Ableist
 - Colonialist
 - Aggrandizing (upholds / reinforces white-supremacy)
2. Many catalog records & finding aids ignore, undervalue, or otherwise deny agency to BIPOC individuals / communities
 - no “find and replace” solution

Building a description audit toolset

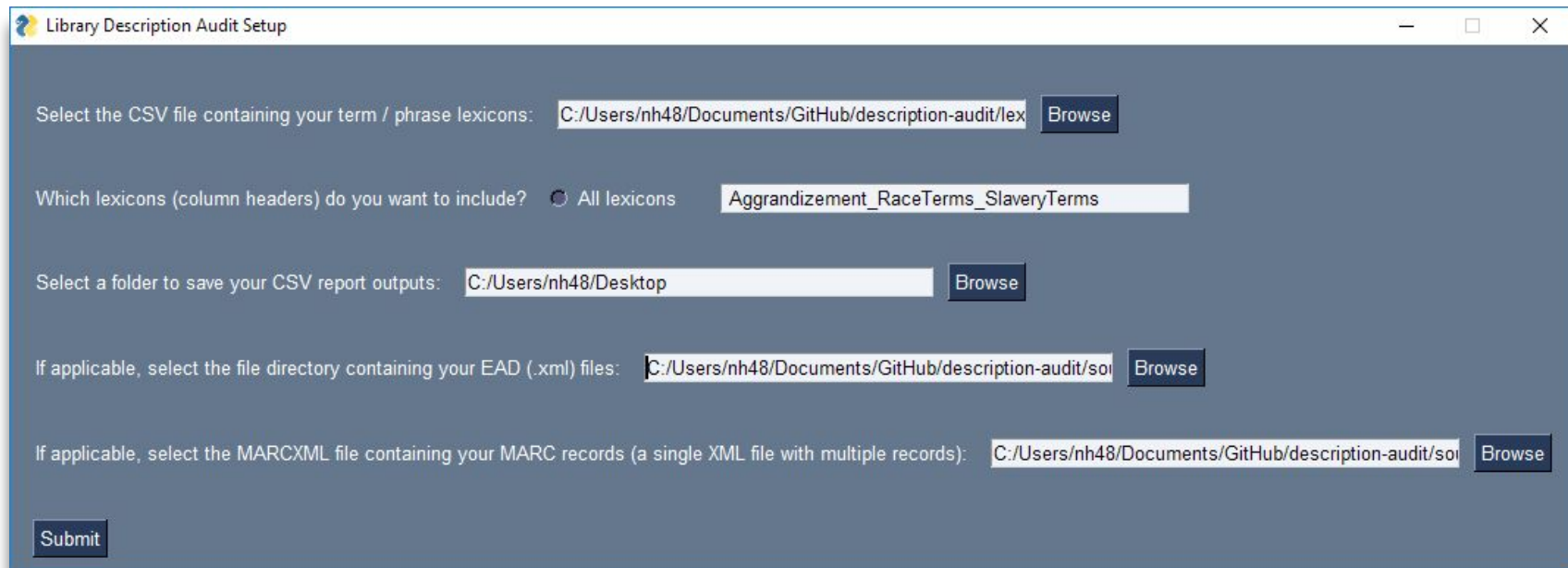
Guiding principles:

- Develop tools / techniques to automate as much as possible
 - Acknowledge need for manual review & analysis; make that work as easy as possible
- Support iterative audits (easy to modify / re-run)
- Build tools that are flexible / shareable
- Build on work from other libraries / archives
 - [Princeton University Libraries](#) (Kelly Bolding, et al.)
 - [University of Pittsburgh Libraries](#) (Kayla Heslin)
 - Many others...

Description audit tool: Building blocks

- **Input: Source Data**
 - Finding aids: **4,033 EAD (XML) files**
 - Catalog records: **11,920 MARC records** (MARC XML exported from Aleph ILS)
- **Input: Lexicons**
 - Lists of harmful terms / phrases
 - Lists of terms / phrases that might indicate presence of hidden/ignored BIPOC voices
 - [HateBase](#) (used by companies to identify online hate speech)
- **Processing: Python / [spaCy](#)** (Natural Language Processing library)
- **Output: CSV audit reports**
- **Developer: 1 Duke student (Miriam) + Noah**

description_audit.exe



The screenshot shows a window titled "Library Description Audit Setup". It contains five input fields, each with a "Browse" button, and a "Submit" button at the bottom left.

Select the CSV file containing your term / phrase lexicons:

Which lexicons (column headers) do you want to include? ☒ All lexicons

Select a folder to save your CSV report outputs:

If applicable, select the file directory containing your EAD (.xml) files:

If applicable, select the MARCXML file containing your MARC records (a single XML file with multiple records):

<https://gitlab.oit.duke.edu/dul-rubenstein/description-audit>

Lexicons managed in single CSV file

C	D	E	F	G
Aggrandizement	RaceEuphemisms	RaceTerms	SlaveryTerms	GenderTerms
acclaimed	color blind	aboriginal	abolition	miss
ambitious	colored	aboriginals	abolitionist	mistress
celebrated	coloured	aborigines	antislavery	mrs.
distinguished	negro	aliens	anti-slavery	muse
eminent	race relations	arab	bill of sale	spouse
esteemed	race situation	arabs	bills of sale	wife
expert	race-based	asians	enslaved	
father of	racial	asiatic	freed slave	
foremost	racism	blacks	freed slaves	
founding father	troubles	bushman	freedman	

Sample audit report - MARC (6800 rows)

Collection info

Match info
(where, match term/phrase, snippet, match lexicon)

	B	C	D	E	F	G	H	I	J
1	bibnumber	last_update	creator	title	extent	match_type	terms	context_snippets	rule_ids
2	000740477	20190109115046.0	Ascension (Ship), creator.	Ascension insurance account, 1793.	1 item.	summary	slaves	value of the ship Ascension and its cargo including 52 slaves. Apparently, the ship was	HateBaseFull
3	000740477	20190109115046.0	Ascension (Ship), creator.	Ascension insurance account, 1793.	1 item.	summary	slaves	value of the ship Ascension and its cargo including 52 slaves. Apparently, the ship was	SlaveryTerms
4	000740477	20190109115046.0	Ascension (Ship), creator.	Ascension insurance account, 1793.	1 item.	summary	slaves	slaves. Apparently, the ship was lost in the slaves' insurrection in 1793. Information	HateBaseFull
5	000740477	20190109115046.0	Ascension (Ship), creator.	Ascension insurance account, 1793.	1 item.	summary	slaves	slaves. Apparently, the ship was lost in the slaves' insurrection in 1793. Information	SlaveryTerms
6	000740480	20120501091907.0	Parks, Mary C., active 1827.	Mary C. Parks journals, 1827-1878, bulk	4 items.	summary	Indians	an account of a meeting with a group of Osage Indians that were visiting there. Other	RaceTerms
7	000740861	20190208100719.0	Matthews, J. B. (Joseph Brown)	J. B. Matthews papers, 1862-1986, bulk	479 linear feet	bionote	prominent	missionary, college professor, author, lecturer, and prominent conservative spokesm	Aggrandizement
8	000740862	19870326000000.0	Crane and Conley (Montgomery)	Daybook, 1834-1841 bulk 1834-1838.	1 item (204 p.)	summary	Prominent	M. Klimby and E. B. Crane of Newwark, N.J. Prominent customers included Albert Pic	Aggrandizement
9	000741861	19870402000000.0	Tuley, Joseph, 1796-1860.	Ledger, 1810-1829.	32 items.	summary	slaves	at Millwood, Clarke Co., Va. Includes shoes for slaves and occasional entries for labo	HateBaseFull
10	000741861	19870402000000.0	Tuley, Joseph, 1796-1860.	Ledger, 1810-1829.	32 items.	summary	slaves	at Millwood, Clarke Co., Va. Includes shoes for slaves and occasional entries for labo	SlaveryTerms
11	000741861	19870402000000.0	Tuley, Joseph, 1796-1860.	Ledger, 1810-1829.	32 items.	summary	slave	Includes shoes for slaves and occasional entries for labor and slave hiring. Prominen	HateBaseFull
12	000741861	19870402000000.0	Tuley, Joseph, 1796-1860.	Ledger, 1810-1829.	32 items.	summary	slave	Includes shoes for slaves and occasional entries for labor and slave hiring. Prominen	SlaveryTerms
13	000741861	19870402000000.0	Tuley, Joseph, 1796-1860.	Ledger, 1810-1829.	32 items.	summary	Prominent	slaves and occasional entries for labor and slave hiring. Prominent customers include	Aggrandizement
14	000741867	19870402000000.0	Alderman, Jacob O.	Papers, 1884-1906.	29 items.	summary	wife	correspondence, 1889-1906, between Alderman and his wife, Mary M. Alderman. Let	GenderTerms
15	000741867	19870402000000.0	Alderman, Jacob O.	Papers, 1884-1906.	29 items.	summary	wife	journal for July 1904, which Alderman sent to his wife.	GenderTerms
16	000741956	20190315161823.0	Carlton, John L., http://id.loc.gov	John L. Carlton papers, 1847-1848.	5 items.	summary	anti-slavery	the "Hunkers." The letters reflect a strong anti-slavery sentiment and mention public	SlaveryTerms
17	000741956	20190315161823.0	Carlton, John L., http://id.loc.gov	John L. Carlton papers, 1847-1848.	5 items.	summary	slavery	Hunkers." The letters reflect a strong anti-slavery sentiment and mention publication	SlaveryTerms
18	000741956	20190315161823.0	Carlton, John L., http://id.loc.gov	John L. Carlton papers, 1847-1848.	5 items.	summary	slavery	publications they believe would be supportive of their beliefs on slavery as well as o	SlaveryTerms
19	000741957	20091002113547.0	Hood, George Henry.	George Henry Hood papers, 1857-1895,	88 items.	summary	wife	(1861-1863) from Hood, to his wife "Etta" (Henrietta Janvrin Hood), while	GenderTerms
20	000741978	20191022145126.0	Coates, Sue, http://id.loc.gov	S. Coates papers, 1871.	0.1 linear ft. (1 ite	summary	wife	of Narragansett Pier, Rhode Island, the apparently affluent wife of Harry Coates. The	GenderTerms

Initial findings

- Many “false” positives
 - Harmful terms in formal titles, organization names
- Terms of aggrandizement are pervasive
 - “Prominent”, “eminent”, “pioneer”, “notable”, “successful”.....
- HateBase lexicon is not very useful in library context

Scale and scope of the reports

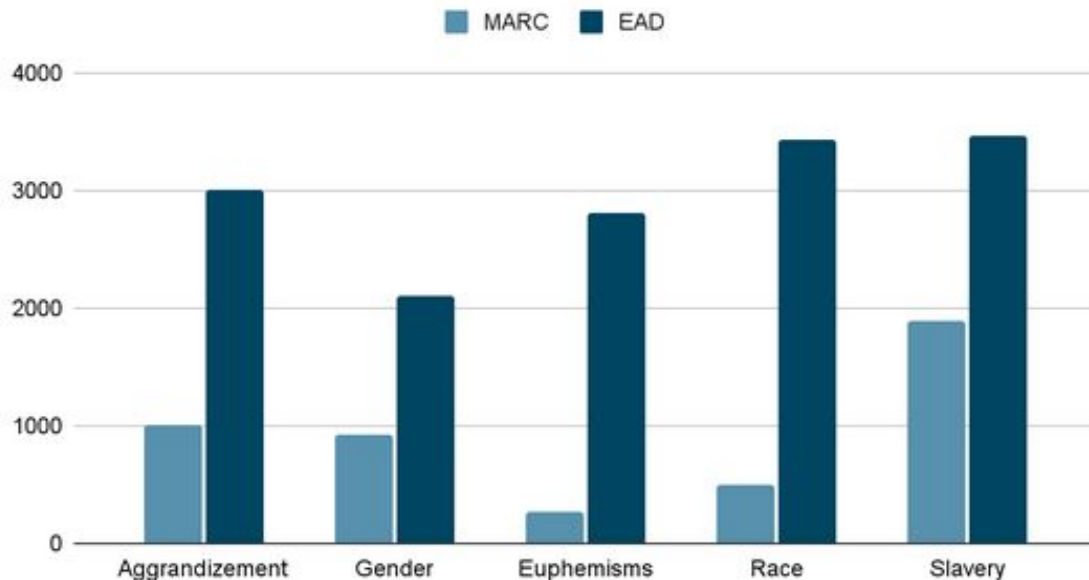
Archival catalog records:

Out of 11,920 MARC records, 2600 collections (**21%**) have at least one hit from one of the lexicons

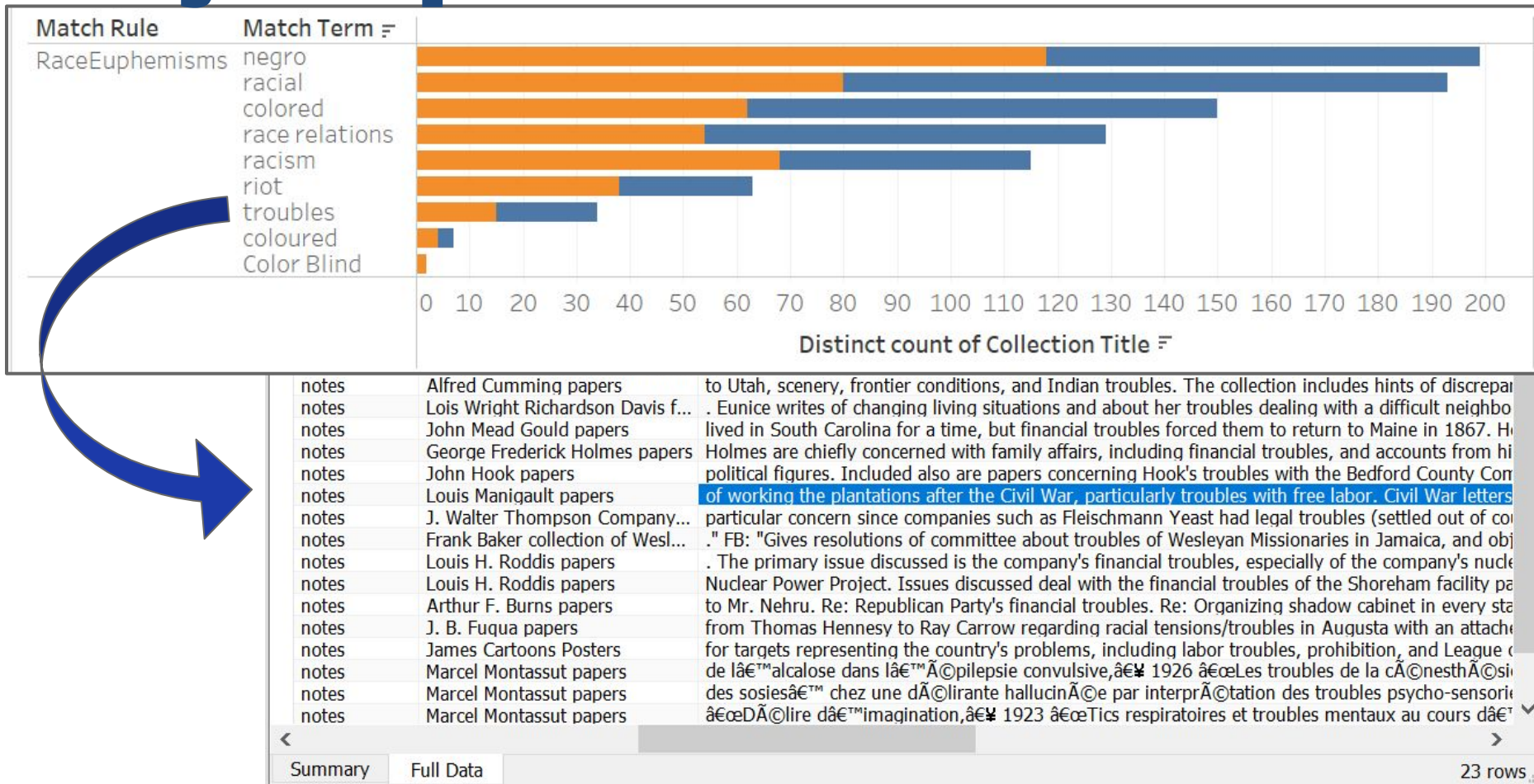
Finding aids:

Out of 4,033 EAD files (~1 million described components), 1676 collections (**41%**) have at least one hit from one of the lexicons

Catalog Records vs. Finding Aids



Using the reports



Next steps...

- Develop visualizations to assist with report analysis
- Refine lexicons based on analysis
- Identify categories of remediation work needed
 - Remove / replace harmful terms in library-produced metadata
 - Contextualize harmful terms in transcribed description (e.g. folder titles)
 - Rewrite aggrandizing / colonialist biographical notes to de-center whiteness
 - Seek to recover agency for BIPOC voices
- Develop remediation workflows

Sorting into remediation workflows

collection_title	match_type	match_term	context_snippet	proposed_remediation_category	Notes
Mary Dowdell Ashley film collection	notes	colored	pier, Boots Ashley (Sharman) in multi	False positive	
Josiah William Bailey papers	notes	Negro	to hold its circus at the same time as	Dated Rubenstein Term	
Josiah William Bailey papers	notes	Racial	for lists of 1928 Democratic state con	Dated original term	
Josiah William Bailey papers	notes	racial	one subject were placed in this sectio	Downplaying bias	
Frank Baker papers	notes	colored	, and illustrations of individuals and s	False positive	
William Watts Ball papers	notes	race relations	local, state, and national issues. Mate	Downplaying bias	
William Watts Ball papers	notes	Negro	of states' rights versus federal contro	Dated original term	
William Watts Ball papers	notes	racial	, enfranchisement and control of the	Downplaying bias	
Graham Arthur Barden papers	notes	colored	, and the establishment of a white ca	Dated Rubenstein Term	
Graham Arthur Barden papers	notes	Colored	, 1951 - several letters concern drinki	Dated original term	
Graham Arthur Barden papers	notes	racial	includes much printed material--spee	Dated Rubenstein Term	Urgent remediation
Aaron Michael Barker papers	notes	troubles	memories (circa 1942-1943) detail his	False positive	
Aaron Michael Barker papers	notes	troubles	99-147, in which he detailed his finan	False positive	
Bernard Irwin Barnes papers	notes	race relations	and traditional devotional subjects to	Downplaying bias	
Lionel Barrow papers	notes	racial	newspaper clippings, dating largely fr	Mention of racial issues	

Thanks for listening!

Feedback / Questions?

noah.huffman@duke.edu

meghan.lyon@duke.edu

Code Repository: <https://gitlab.oit.duke.edu/dul-rubenstein/description-audit>

Slides: <https://tinyurl.com/v3hk7p4w>