The Report Committee for Maria Victoria Fernandez
Certifies that this is the approved version of the following Report:

The Coloniality of Metadata: A Critical Data Analysis of the Archive of Early American Images at the John Carter Brown Library

APPROVED BY
SUPERVISING COMMITTEE:

Kelly McDonough, Supervisor

Tanya E. Clement

The Coloniality of Metadata: A Critical Data Analysis of the Archive of Early American Images at the John Carter Brown Library

by

Maria Victoria Fernandez

Report

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degrees of

Master of Arts

And

Master of Science in Information Studies

The University of Texas at Austin

May 2018

Abstract

The Coloniality of Metadata: A Critical Data Analysis of the Archive of Early American Images at the John Carter Brown Library

Maria Victoria Fernandez, M.A., M.S.I.S.

The University of Texas at Austin, 2018

Supervisor: Kelly McDonough

How do contemporary metadata practices replicate the coloniality of power embedded in European colonial documents describing the Spanish Americas? This report draws from critical theory and postcolonial thought within the field of Latin American studies to explore the extent to which standardized description and categorization practices can perpetuate a Eurocentric colonial gaze on the Spanish Americas. In order to ground this theoretical engagement within the fields of information science and critical data studies, a dataset is compiled and computationally analyzed, which contains metadata records corresponding to images derived from books, manuscripts, and broadsides printed between 1492 and 1825 about the Spanish Americas found in the Archive of Early American Images at the John Carter Brown Library. This report then applies unsupervised machine learning techniques and counts word frequencies to identify broad metadata trends across the collection. Ultimately, this report combines methods from critical data studies and Latin American cultural studies to understand how controlled vocabularies and descriptive practices can perpetuate colonial structures of power.

Table of Contents

List of Tables vi

List of Figures vii

Introduction 1

Discourse Communities: Latin American Cultural Studies and Critical Cataloging 4

Methods: Compiling, Cleaning, and Computationally Analyzing the Dataset 15

1. Comparing word frequencies of text derived from the "subject headings," "description," and "notes" fields in both datasets 18

2. Running two unsupervised machine learning algorithms, LDA topic modeling and K-means clustering, on text corpora extracted from the description and notes fields from both datasets 21

Conclusion and Avenues for Further Research 25

Appendices 28

Appendix A: Datasets and Python Scripts Used for Computational Analysis 28

Appendix B: Data Analysis Workflow 29

Appendix C: Word Frequency Visualizations 30

Appendix D: Unsupervised Machine Learning: Latent Dirichlet Allocation (LDA) Topic Modeling 34

Appendix E: Unsupervised Machine Learning: K-Means Clustering 38

Bibliography 39

List of Tables

Table 1: Latent Dirichlet Allocation (LDA) topic modeling of a text corpus derived from the description and notes fields of the Spanish America Dataset....................34

Table 2: K-Means Clustering of the Spanish America Dataset 38

List of Figures

Figure 1: Data Analysis Workflow..29

Figure 2: Frequency of Top 20 Subject Headings in the Spanish America Dataset 30

Figure 3: Frequency of Top 20 Subject Headings in the Indigenous Terms Dataset...31

Figure 4: Word Frequencies of the Top 50 Words in the Description and Notes Text Corpus Derived from the Spanish America Dataset.............................32

Figure 5: Word Frequencies of the Top 50 Words in the Description and Notes Text Corpus Derived from the Indigenous Terms Dataset 33

Introduction

This critical data analysis focuses on a dataset derived from the Archive of Early American Images, a database of digitized primary sources from the John Carter Brown Library.[1] The John Carter Brown Library (JCB) is a center for advanced research in history and the humanities at Brown University that contains extensive collections of rare books, manuscripts, and maps relating to the European discovery, exploration, settlement, and development of the Western Hemisphere until the 1820s.[2] The Archive of Early American Images contains graphic representations of colonial North and South America derived from books, manuscripts, and broadsides in the library's collections. These images are accompanied by extensive bibliographical and descriptive information.

While the database contains over 8000 unique image records from six distinct geographic regions in the Western Hemisphere (North America, Arctic, Spanish America, Brazil, Caribbean Islands, and Guianas), this project analyzes a filtered dataset of approximately 2600 image records specifically relating to Spanish America.[3] This dataset contains images derived from thousands of European books printed between the sixteenth and nineteenth centuries that contain some reference to Spanish America. These images represent European understandings and imaginaries of the New

[1] "Archive of Early American Images: The John Carter Brown Library," *The John Carter Brown Library,* accessed February 15, 2018, https://jcb.lunaimaging.com/luna/servlet/JCB~1~1

[2] Description of the John Carter Brown Library: The JCB has been a leading institution within the rare book and special collections community in its strategic plan and implementation strategies to digitize its entire collection. At present, the JCB has digitized roughly twenty percent of its collection and has made it fully available through the Internet Archive (https://archive.org/details/JohnCarterBrownLibrary) and the digital content platform, LUNA Imaging (https://jcb.lunaimaging.com/luna/servlet) In 2016, the Andrew W. Mellon Foundation awarded a grant to the JCB to move beyond static web images of its collection materials and experiment with new platforms and forms of digital engagement and discovery. For more information about the JCB's institutional history, see https://www.brown.edu/academics/libraries/john-carter-brown/about/history-library

[3] The term Spanish America refers to any area of the Western Hemisphere that was under Spanish colonial rule between the sixteenth through nineteenth centuries.

World. Topics represented in the collection include botany and the natural sciences; Catholic missions, the Church, and evangelization; transatlantic trade and commerce; navigation and exploration; natural resource extraction; warfare and conquest; law and government; contact between Europeans and Indigenous peoples; ethnology; and slavery.

This project engages with current discourses in critical cataloging and Latin American cultural studies to explore how standardized description practices can perpetuate colonial structures of power. This theoretical discourse is paired with a computational analysis of the Spanish America dataset derived from the Archive of Early American Images. Guiding research questions for this project are: How do descriptive practices used with this dataset walk the fine line between implementing standardized terminology and perpetuating a Eurocentric colonial gaze on the Spanish Americas? Can metadata unintentionally extend or replicate colonial power structures embedded in the primary sources the metadata describes?

In order to approach answering these questions, this paper begins by identifying discourse communities it seeks to engage with and contribute to. This first section establishes a theoretical framework to contextualize the data analysis in current discussions about colonialism and Eurocentrism that are taking place within the field of Latin American studies as well as recent debates about critical cataloging practices within the field of library and information studies. After establishing this theoretical foundation, this paper provides a thorough overview of the computational methods used to compile, clean, and analyze the dataset derived from the JCB's Archive of Early American Images. Even though the data analysis methods carried out are unable to fully address the guiding research questions this paper poses, these methods provide a crucial foundation for continued research into using collections as data to study how colonial structures of power manifest themselves in contemporary knowledge organization practices.

Discourse Communities: Latin American Cultural Studies and Critical Cataloging

The intended audience of this project is a combination of Latin American cultural studies scholars, Indigenous studies scholars, and information professionals working to expand critical cataloging efforts within the field of library and information science. The colonial documents that make up this project's dataset represent European understandings of the New World and would be of great interest to Latin American cultural theorists engaging in discourses about colonialism, coloniality, postcolonialism, and the power dynamics involved in Spain's colonial rule in the Americas. It is also relevant to historians and cultural studies scholars analyzing infrastructures of economic power, globalization, modernity, and the transatlantic exchange of ideas within colonial Spanish America.

The field of Latin American studies has been profoundly impacted by the critique of colonialism and coloniality in their diverse temporal and spatial manifestations. The theorization of coloniality and recognition of the far-reaching influence of Spanish imperialism played central roles in the reconsideration of postcolonial thought during the 1980s. At that time, the field of postcolonial studies was chiefly concerned with the age of high imperialism and decolonization in Asia, Africa, the Middle East, and the non-Spanish Caribbean. According to Latin American critical theorist Santa Arias, "[i]t had one major blind spot," however: the material and ideological impact of the Spanish conquest of the Americas on all subsequent forms of colonialism around the world."[4]

[4] Santa Arias, "Coloniality and Its Preoccupations," *Latin American Research Review* 48.3 (2013): 214, doi: 10.1353/lar.2013.0042.

During the postcolonial turn, prominent Latin American theorist Aníbal Quijano coined the term "coloniality of power," which contemporary literary and philosophy scholars Mabel Moraña, Enrique Dussel, and Carlos Jáuregui consider "pivotal to the understanding and critique of early and late stages of colonialism in Latin America, as well as of its long-lasting social and cultural effects."[5] Quijano's coloniality of power is a term that refers to the structures of hegemony that emerged during the conquest of the Americas, which continue to exist to the present.[6] He contends that social, cultural, political, and economic power dynamics that exist in today's globalized world and its conceptualization of modernity are rooted in the power dynamics at play between colonizer and colonized in Latin America throughout the sixteenth and nineteenth centuries. Quijano traces a continuous line from the Spanish conquest to today and discusses how the social construct of race was central to centuries of exploitation, inequality, and unequal power relations in Latin America. These power dynamics played out between individuals commanding authority and economic control over populations subjected to subordinate positions in society. Quijano frames the coloniality of power as an ever-present living legacy of the colonial encounter in the Americas. He asserts that after the end of colonialism, early instruments of social domination survived and continued to shape Eurocentric forms of rationality and modernity. Through his far-reaching theoretical contributions, Quijano presented the foundational idea within Latin American

[5] Mabel Moraña, Enrique Dussel, and Carlos Jáuregui, "Colonialism and Its Replicants," in *Coloniality at Large: Latin America and the Postcolonial Debate*, eds. Mabel Moraña, Enrique Dussel, and Carlos Jáuregui. (Durham: Duke University Press, 2008), 2.

[6] Aníbal Quijano, Coloniality of Power, Eurocentrism, and Latin America," trans. Michael Ennis, *Nepantla: Views from South* 1, no. 3 (2000): 533-580.

critical thought that "modernity,[7] the Americas, and capitalism were borne in the context of the "discovery" of the Americas."[8]

This critical data analysis draws from Quijano's theoretical framework to assert that the JCB's Archive of Early American Images is an archival embodiment of the "coloniality of power." Further, this project maintains that this database of images derived from European books depicting Western European perceptions of the Spanish Americas throughout the sixteenth through nineteenth centuries clearly manifests the profound influence of Eurocentrism and European rationalizations of modernity. Quijano defines Eurocentrism as a "trait common to all colonial dominators" that was rooted in "the success of Western Europe in becoming the center of the modern world-system."[9] Non-western cultures were integrated into what Moraña et al call a "solidly Eurocentric frame of consciousness" that "rel[ied] on European vocabulary."[10] Using this Eurocentric frame and vocabulary, Europeans carved out what Latin American literary scholar

[7] According to Quijano, "Modernity refers to a specific historical experience that began with America, when new material and subjective and intersubjective social relations have been produced, alongside the emergence of the new Euro-centered, capitalist, colonial world power structure. Above all, there was a new place for the idea of future in the world imaginary, especially among the peoples that configured Europe. So developed a new perspective on space/time and on the place of humankind in such a new world. But it was Western Europe that, since the seventeenth century formally and systematically elaborated the new intersubjective universe in a new knowledge perspective. And it was Western Europe that termed that knowledge perspective 'modernity' and 'rationality'." Aníbal Quijano, "Coloniality of Power and Eurocentrism in Latin America," *International Sociology* 15, no. 2 (2000): 220-221.

[8] Nelson Maldonado-Torres, "Colonialism, Neocolonial, Internal Colonialism, the Postcolonial, Coloniality, and Decoloniality." in *Critical Terms in Caribbean and Latin American Thought: Historical and Institutional Trajectories,* edited by Yolanda Martínez-San Miguel, Ben Sifuentes-Jáuregui, and Marisa Belausteguigoitia, (New York: Palgrave Macmillan, 2016), 75. Also see Aníbal Quijano, "La modernidad, el capitalismo, y América nacen el mismo día," *Boletín Ilia* 10 (January 1991): 42-57.

[9] Aníbal Quijano, "Coloniality of Power, Eurocentrism, and Latin America," trans. by Michael Ennis, *Nepantla: Views from South* 1, no. 3 (2000): 541.

[10] Moraña et al., 18.

Santa Arias identifies as “the historical memory of colonialism and its epistemological, social, economic, and political legacy.”[11]

In order to understand the extent of Eurocentrism in the Archive of Early American Images, it is useful to consider the concept of locus of enunciation. According to Walter Mignolo, a leading cultural theorist within the field of Latin American studies, the locus of enunciation is “the disciplinary, geocultural and ideological space from which discourses of power and resistance are elaborated.”[12] In the particular case of the Archive of Early American images, the locus of enunciation is deeply entrenched within a Eurocentric power structure that occupies a privileged epistemological space. Critical theorist Eduardo Mendieta extends Mignolo’s insights by framing the importance of identifying loci of enunciation in order to determine: “Who is the subject who thinks what object? And, more acutely still, where is this subject, and how does it project and localize its object of knowledge? Who speaks for whom and who speaks over or about whom?”[13] In the case of the Archive of Early American Images, it is the European that speaks for all non-European subjects in the Spanish Americas. Mendieta extends this line of inquiry by stating that “we always speak about something, or someone, from a given perspective, and when we do so, we are enacting, performing, [and] deploying certain forms of knowledge-power.”[14]

By considering Eurocentrism as a locus of enunciation from which certain forms of knowledge-power are enacted, we can tie contemporary manifestations of knowledge-power to

[11] Arias, 219.

[12] Walter D. Mignolo, “Epistemic Disobedience, Independent Thought and De-Colonial Freedom,” *Theory, Culture & Society* 26, no. 7-8 (2009): 1-23, quoted in Moraña et al., 3.

[13] Eduardo Mendieta, “Remapping Latin American Studies: Postcolonialism, Subaltern Studies, Post-Occidentalism, and Globalization Theory” in *Coloniality at Large: Latin America and the Postcolonial Debate*, ed. Mabel Moraña, Enrique Dussel, and Carlos Jáuregui (Durham: Duke University Press, 2008), 294.

[14] Ibid., 295.

their historical, colonial legacies. In the case of the Archive of Early American Images, the metadata and standardized classification practices used to describe collection material are part of universalized knowledge organization systems that perpetuate bias and preserve colonial power structures. This leads us to the second discourse community this data analysis project engages with: critical cataloging in the field of library and information science.

The practice of classifying knowledge in libraries is not a neutral act. Library knowledge organization in the form of classification and subject headings terminology are powerful—and problematic—forms of representation. According to activist librarian Emily Drabinski, "When an item is placed in a particular category or given a particular name, those decisions always reflect a particular ideology or approach to understanding the material itself."[15] In the case of the Archive of Early American Images, cataloging and descriptive practices walk the fine line between implementing standardized terminology and perpetuating colonial power structures embedded in universalized knowledge organization systems.

Library and information studies professionals have actively critiqued universalized classification structures and controlled vocabularies since the 1970s, when a critical cataloging movement began to flourish in the United States. This movement addressed the problem of bias in knowledge organization structures and called attention to the fundamentally political project of sorting materials into categories and then giving those categories names. In his foundational book, *Prejudices and Antipathies: A Tract on the LC Subject Heads Concerning People,* critical cataloger Sanford Berman presents the case that Library of Congress Classification (LCC) and Library of Congress Subject Headings (LCSH) fail to accurately and respectfully organize library

[15] Emily Drabinski, "Queering the Catalog: Queer Theory and the Politics of Correction," *The Library Quarterly: Information, Community, Policy* 83, no. 2 (2013): 110.

materials about social groups and identities that lack social and political power, such as non-White, non-Christian, and non-heterosexual peoples.[16] Berman's powerful breakdown and analysis of problematic subject headings continue to be extended by critical scholars today. For instance, in a 2017 article in the *International Indigenous Policy Journal,* activist librarian Michael Dudley convincingly presents a case study of how LCSH schemes are overwhelmingly Euro- and Christian-centric in nature, display many sexist and racist tendencies, and ultimately use "pejorative language to describe, exclude, or misrepresent marginalized knowledge domains."[17] In an often cited passage within the critical cataloging literature, Berman states that LCSH

> can only 'satisfy' parochial, jingoistic Europeans and North Americans, white-hued, at least nominally Christian (and preferably Protestant) in faith, comfortably situated in the middle- and higher-income brackets, largely domiciled in suburbia, fundamentally loyal to the Established Order, and heavily imbued with the transcendent, incomparable glory of Western civilization.[18]

Since the publication of Berman's influential tract, many librarians, catalogers, and information science scholars have contributed to a growing body of literature[19] in the field of library and information science that illuminates the ways in which knowledge organization schemes "are rife with Euro-, Christian-, hetero-, and ethnocentric biases and sexism."[20]

In the case of the Spanish America dataset at the center of my critical data analysis project, one area of bias that became evident through both a close and distant reading of individual records

[16] Sanford Berman, *Prejudices and Antipathies: A Tract on the LC Subject Heads Concerning People* (Jefferson, N.C.: McFarland & Co., 1993).

[17] Michael Q. Dudley, "A Library Matter of Genocide: The Library of Congress and the Historiography of the Native American Holocaust," *The International Indigenous Policy Journal* 8, no. 2 (2017): 8.

[18] Berman, 15.

[19] Consult the bibliography for full citations of the following important contributions to the critical cataloging literature. Beall 2006; Biswas 2018; Bone & Lougheed 2018; Drabinski 2013; Duarte & Belarde-Lewis 2015; Dudley 2017; Greenblatt 1990; Lee 2011; Nuckolls 2016; Olson 2001; Olson 2002; Roberto & Berman 2008; Yen-Rah Yeh 1971.

[20] Dudley, 7.

was the use of problematic subject headings such as “Indians of South America”; “Indians, treatment of”; “Indians—first contact with Europeans”; “first contact of aboriginal peoples with Westerners”; “Indian—captivity narratives; and “conquest”. Many of these subject headings have already been extensively analyzed within the critical cataloging literature.[21] For instance, academic librarian Jeffrey Beall studies the representation of ethnic groups in LCSH and breaks down the structure of the main subject heading “Indian,” presenting how more specific terms that Indigenous communities use to self-identify are nested as narrower sub-headings while “Indian” remains the heading at the topmost level.[22] From a Canadian context, librarians Christine Bone and Brett Lougheed also critique the use of “Indian” as the top-most subject heading and analyze the problematic way geographic distinctions are used within the LCSH term hierarchy.[23] Finally, law librarian Karen Nuckolls discusses how no subject heading for “Indigenous peoples” exists; instead, that much-needed term remains only a Used For (UF) term within the LCSH framework while “Indians of North, South, and Central America” continue to be used despite the much documented preference for using “indigenous peoples” over “Indians” in a contemporary, international context.[24]

Not only do scholars address the biases of subject headings within the critical cataloging literature, but they also consider terminology within other descriptive metadata contexts as well.

[21] Consult Beall 2006; Berman 1971; Biswas 2018; Bone & Lougheed 2018; Duarte & Belarde-Lewis 2015; Dudley 2017; Lee 2011; Moorcraft 1992.

[22] Jeffrey Beall, “Ethnic Groups and Library of Congress Subject Headings,” *Colorado Libraries* 32, no. 4 (2006): 39.

[23] Christine Bone and Brett Lougheed, “Library of Congress Subject Headings Related to Indigenous Peoples: Changing LCSH for Use in a Canadian Archival Context,” *Cataloging & Classification Quarterly* 56, no. 1 (2018): 89.

[24] Karen A. Nuckolls, “LC Subject Headings, FAST Headings, and Apps: Diversity Can Be Problematic In the 21st Century” in *Rethinking Technical Services,* ed. Bradford Lee Eden (Lanham, M.D.: Rowman & Littlefield, 2016), 88.

For instance, there are many politically-charged terms to refer to Indigenous peoples that are deeply rooted in colonial histories and power dynamics. A small sample of these terms include: indian, natives, native people, native American, indigenous people, tribe, and aboriginal. While many of these terms are used synonymously in various contexts, from the disciplinary perspective of Indigenous Studies, this practice is problematic because "a term can be a loaded word, used as a powerful method to divide peoples, misrepresent them, and control their identity."[25] In a useful guide for navigating these terms, the First Nations & Indigenous Studies program at the University of British Columbia addresses how "terminology can be critical for Indigenous populations, as the term for a group may not have been selected by the population themselves but instead imposed on them by colonizers."[26] Perceptions of Indigenous identity can be complicated and terms such as "Indian," "American Indian," and "Native American" have complex historical origins as identifiers used by colonizers to name and exert power over Indigenous populations. When metadata describes primary source materials that are a direct product of colonial society, as is the case with the Archive of Early American Images, each term holds a singular identity as a vestige of colonialism. The past, present, and future interpretation of these loaded words is also variable and in flux depending on different socio-historical and geo-political contexts.

A relevant study that considers the power to name and assign terminology to Indigenous populations is a 2015 article by Indigenous information studies scholars Marisa Elena Duarte and Miranda Belarde-Lewis titled "Imagining: Creating Spaces for Indigenous Ontologies." Writing from the intersection of Indigenous studies and Information Science, Duarte and Belarde-Lewis

[25] "Terminology," First Nations & Indigenous Studies Program, The University of British Columbia, 2009. http://indigenousfoundations.web.arts.ubc.ca/terminology

[26] Ibid.

discuss how naming practices within controlled vocabulary systems reflect colonial power dynamics. For instance, when it comes to terminology within standardized classification systems to identify Indigenous groups, they explain how

> the term American Indian emerged out of common use by Spanish colonial authorities and settlers who, since the late 1500s, were erroneously describing Indigenous inhabitants as *indios*, or 'Indians.' The terms 'Native American,' 'American Indian,' and 'Indian' are all terms of conflation designed for governmental racial and class management.[27]

For Duarte and Belarde-Lewis, continued use of these terms reflects how the cataloging and classification of information become enduring vestiges of colonialism. They insist that to move away from using these terms is a "step toward the redress of colonial power."[28]

To explain how the naming of Indigenous peoples continues to be a tool for racial and class management, the authors analyze the contemporary use of the term "Native American" in the United States. In the 1990s, the US federal government adopted the term "Native American" to replace "American Indian" as a response to the social policy of multiculturalism.[29] Duarte and Belarde-Lewis frame this action as an attempt to reduce Indigenous people to an ethnic minority and state that it is "important to understand the term 'Native American' as a colonial tool for describing an Indigenous U.S. population in aggregate, regardless of the social, political, and philosophical distinctions of the many tribal peoples of the United States."[30] Having to depend on this "imprecise term," leads to "categorical misunderstanding" that "occludes and erases a wide range of distinctive epistemologies, philosophies, languages, and experiences." This

[27] Duarte, 680.
[28] Ibid., 682.
[29] Ibid., 680.
[30] Ibid.

misunderstanding, in turn, allows for “governments and elite classes of citizens [to] continually benefit”[31] from the power to name and perpetuate a key mechanism of colonialism.

The concerns Duarte and Belarde-Lewis present about the structures of colonialism influencing the representation of Indigenous peoples in knowledge organization systems and the biases found in LCSH by critical catalogers are all directly relevant to issues of representation and misrepresentation in the Spanish America dataset derived from the Archive of Early American Images. One problematic colonial legacy that is present in the dataset is that metadata description practices have a tendency to reinforce language used to describe the historical treatment of Indigenous peoples in the Spanish Americas by Europeans “as objects of knowledge or of domination and exploitation.”[32]

I take a post-colonial stance grounded in the field of Latin American cultural studies in order to understand how using standardized terminology to describe inherently Eurocentric colonial primary sources can extend that Eurocentrism as a vestige of colonialism into contemporary knowledge organization systems. Questions guiding this critical data analysis are: How do the description practices used within the dataset walk the fine line between implementing standardized terminology and reinforcing a Eurocentric colonial gaze on the Americas? Can metadata unintentionally extend or replicate colonial power structures embedded in the primary sources the metadata contextualize? How can we describe historical colonial documents without reinforcing the Eurocentric ideology within them? In order to begin to answer these questions, I

[31] Ibid., 681.

[32] Quijano, “Coloniality of Power,” 221.

conducted a computational analysis of various metadata fields in the dataset. This process is outlined and detailed in the following section.

Methods: Compiling, Cleaning, and Computationally Analyzing the Dataset

In order to compile a dataset derived from the Archive of Early American Images, I contacted Ross Mulcare, the JCB's Assistant Director for Digital Engagement and Discovery. With his help, I was able to obtain a CSV containing all the image records held within the Archive of Early American Images database. While the database contained 8324 unique item records from six distinct geographic regions in the Western Hemisphere (North America, Arctic, Spanish America, Brazil, Caribbean Islands, and Guianas), I compiled a reduced dataset of 2654 records containing Spanish American content. To create this subset of records, I used OpenRefine, an open source desktop application for data cleanup and transformation, to filter and extract all image records assigned "Spanish America" as the primary geographic area in the "geographic area" metadata field. To access a copy of the Spanish America dataset compiled for this project, please refer to appendix A.

With my Spanish America dataset, I then set out to batch edit individual item records using the "facets," "clustering," and "custom text transform expressions" features of OpenRefine. In OpenRefine, clustering refers to the operation of "finding groups of different values that might be alternative representations of the same thing."[33] I employed a range of methods and algorithms available through OpenRefine's facet and clustering features to: (1) discover possible redundancies and to standardize place names appearing in the "place of publication" field; (2) clean up "image date" and "source date" fields that ranged widely in format including approximate date ranges, circa dates, and roman numerals; and (3) correct typos in creator names and subject areas. First, I

[33] Delpeuch, Antonin. "Clustering in Depth." *Open Refine.* Last modified September 27, 2017. https://github.com/OpenRefine/OpenRefine/wiki/Clustering-In-Depth

used a series of key collision methods, including fingerprinting and phonetic fingerprinting. Key collision methods are easy to use in OpenRefine, however they constantly shift between being too strict or too lax in their assessment of how much difference to tolerate between strings analyzed. Therefore, I also applied a series of nearest neighbor methods, including both Levenshtein Distance and Prediction by Partial Matching (PPM) available in OpenRefine. These latter methods allow for fine tuning of distance thresholds between strings. In every case of a possible redundancy, I examined the records individually to verify if the duplication was an error and edited the records if necessary.

Originally, my data set contained 71 metadata fields, but through the data clean-up process in OpenRefine, I reduced it to 40 fields. During this extensive data clean up, I removed blank, redundant, and repetitive fields. For instance, after selecting all records containing "Spanish America" in the "geographic area 1" field, I noticed that all the following "geographic area #2-6" fields were blank, so I removed them from the dataset. I also removed the "Creator Role", "Creator Dates" and "Creator" fields because they represented a second set of identical fields that had already been accounted for elsewhere in the dataset.

Once the OpenRefine data clean-up process was complete, I transferred the Spanish America dataset containing 2654 records into a list of lists in Python and began the critical data analysis. In order to conduct a distant reading of descriptive terminology used to contextualize the collection material, I set up a comparative analysis of two parallel datasets: 1) the Spanish America dataset containing 2654 records and 2) a sub-dataset derived from the Spanish America dataset labeled the "Indigenous terms dataset." I wanted to compare the most frequently used descriptive terms across the entire collection to those used to describe images specifically referencing some aspect of Indigenous culture within the Spanish Americas. The goal was to have a basis for comparison and see how descriptive terminology trends for part of the dataset compared to those trends for the whole dataset. I compiled the Indigenous terms dataset by using a “for loop” to search iteratively across every field in the Spanish America dataset for common terminology that are used as variant forms or synonyms of “indigenous peoples”. If an item record in the dataset contained a text string of any of the following terms—’indian', 'native american', 'natives', 'native people', 'indigenous', 'tribe', 'aboriginal’—across any of the metadata fields, that record was added to the new Indigenous term dataset. This dataset ended up containing 1482 records derived from the 2854 records in the Spanish America dataset. To access a copy of both datasets, please refer to appendix A.

Once I had these two parallel datasets, my data analysis consisted of the following two stages: (1) compare word frequencies of text derived from the ‘subject headings’, ‘description,’ and ‘notes’ fields in both datasets; and (2) run two unsupervised machine learning algorithms, LDA topic modeling and K-means clustering, on text corpora extracted from the description and notes fields from both datasets. For a helpful visualization of this workflow, please consult the data analysis flowchart in figure 1, appendix B.

1. Comparing word frequencies of text derived from the “subject headings,” “description,” and “notes” fields in both datasets

In order to study word frequencies, I extracted the subject headings from both the Spanish America and the Indigenous terms datasets. I created two list of lists of the subject headings[34] and counted the frequency of each word. I then wrote these lists of lists to CSVs and created bar graphs to visualize the top 20 subject heading frequencies in each dataset (see figures 1 and 2 in appendix B). In a similar way, I created a text corpus derived from the description and notes fields within the full Spanish America dataset. I iterated across every item record and combined the text within the description and notes fields into a list of lists variable called “combined_description_notes_list.” I followed the same procedure with the Indigenous terms dataset, combining the description and notes fields, and created a text corpus within the “combined_indig_description_notes_list” list of lists variable. With these two parallel text corpora, I then used the textblob and nltk python libraries to remove stop words, punctuation, and count word frequencies.[35] I then wrote these list of lists to CSVs and created bar graphs to visualize the word frequencies of the top 50 words in the description and notes text corpus. After manually cleaning up some remaining stop words from the mix such as “was” and “de” (meaning “of” in Spanish), the top 50 words used in the description and notes metadata fields can be seen in figures 4 and 5 of appendix C.

[34] Consult the variables “master_subjectheading_list_no_subheadings” and “master_indig_subjectheading_list_no_subheadings” in the following Jupyter Notebook that contains the Python scripts used to conduct this data analysis: https://github.com/fernandezmv/JCBDataAnalysis/blob/master/JCBDatasetAnalysis_Spring2018.ipynb

[35] Consult the variables “sorted_freq_fulldataset_sans_stopwords” and “sorted_freq_indigdataset_sans_stopwords” in the following Jupyter Notebook ” in the following Jupyter Notebook that contains the Python scripts used to conduct this data analysis: https://github.com/fernandezmv/JCBDataAnalysis/blob/master/JCBDatasetAnalysis_Spring2018.ipynb

When comparing the most frequent subject headings across both the Spanish America and the Indigenous terms datasets, the subject headings “natural history,” “botany,” “Indians of South America,” and “Indians of Mexico” all rank within the top five for both datasets. These subject headings indicate the most common topics represented across both datasets and it is important to point out the dominating prevalence of subject headings with the term “Indians.” This word, while seemingly benign on the surface, is controversial in the spheres of critical cataloging and knowledge organization because it is a politically-charged term with colonial historical origins. For over four hundred years it was used by colonizers to name and exert power over Indigenous populations in the Americas. When metadata describes primary source materials that are a direct product of colonial society, as is the case with the Archive of Early American Images, each term holds a singular identity as a vestige of colonialism. For a more extensive engagement with these concepts, consult the discourse communities section of this paper.

Within the colonial context of this dataset, the top subject headings “natural history” and “botany” also maintain deep ties to critical theorist Aníbal Quijano’s concept of “coloniality of power” and perspectives on Eurocentrism since most of the image records tagged with these subject headings are derived from books written by Europeans interested in extracting the flora and fauna of the Americas in order to make contributions to the new and developing fields of science and medicine in early modern Europe. In the case of the Archive of Early American Images, I maintain that the metadata and standardized classification practices used to describe collection material are part of universalized knowledge organization systems that perpetuate Eurocentrism and colonial power structures.

In addition to counting subject heading frequency, I also calculated the frequency of terms used in the description and notes text corpora derived from both the Spanish America and Indigenous terms datasets. To see the top 50 words used within the description and notes metadata fields across the entire dataset, see figure 3 in appendix B. In a similar way to the subject heading frequency count, the top terms included "plant," "Native," and "American," which serve as references to botany and Indigenous peoples. It is interesting to note that while the subject headings did not include the terms "Indigenous" or "Native American" and depended instead on the term "Indians," the content within the description and notes metadata fields never included the term "Indian." Instead, these fields using the terms "native," "Native American," and "Indigenous." Further investigation is needed to understand the factors that may be contributing to these varying descriptive practices across different metadata fields in this collection.

Analyzing word frequencies across the subject headings and description and notes metadata fields not only indicated various descriptive practices referring to Indigenous communities, but it also brought to light broader themes related to the collection material. For instance, when viewing all the word frequency counts as an aggregated whole, major themes begin to stand out such as: Indigenous peoples; navigation, ships, voyages; Aztec, Maya, Inca; flowers, trees, medicinal plants, botany, natural history; Theodor de Bry; soldiers, Spain, conquerors; clothing and dress; Saints, Catholic church, Catholic religious orders, Virgin Mary. Many of these themes are reinforced through both close and distant readings of the datasets. The following data analysis section describes the use of unsupervised machine learning algorithms to identify many of the themes that surfaced through counting word frequencies.

2. Running two unsupervised machine learning algorithms, LDA topic modeling and K-means clustering, on text corpora extracted from the description and notes fields from both datasets

The next step in my analysis was to run unsupervised machine learning algorithms to extract topics and thematic patterns from my dataset. I ran LDA topic modeling and k-means clustering on the two text corpora I derived from the description and notes fields in the Spanish America and Indigenous terms datasets (see appendix D for LDA topic models and appendix E for K-Means clustering results). In order to prepare my text corpora for these unsupervised learning techniques, I removed stop words, tokenized the words, and then vectorized the tokens.

I chose to use unsupervised machine learning algorithms for this analysis because supervised learning algorithms require training data, which I was not able to compile given the nature of my dataset and the size of the text corpora I was working with. Unsupervised learning methods are usually referred to as clustering algorithms since they "try to solve the problem of grouping the predictive contexts x into coherent groups."[36] In the case of the text corpora derived from my datasets, I was hoping these algorithms would detect patterns and similarities between words and across records in order to cluster them into groups. I wanted to not only compare these clusters to themes I had identified while manually reviewing item records in the datasets but also to themes established by scholars and curators who had compiled an exhibition catalog featuring material from the original rare book collection this dataset was derived from.

This exhibition catalog, *The Literature of the Encounter: A Selection of Books from European Americana*, grouped books at the JCB into six thematic sections: geography and history;

[36] Chris Brew, "Language Processing: Statistical Methods," in *Encyclopedia of Language & Linguistics* ed. Keith Brown (Amsterdam: Elsevier Science, 2006), 603.

missions and religious history; ethnology; science; commerce and government; and literature.[37] I wanted to see if the algorithms picked up on similar or different themes that the curators with subject expertise had identified to organize the bibliographic content within the exhibition catalog as well as the themes I had determined from my close reading of the records in my dataset. Would the algorithms just reaffirm the patterns already determined by humans? Or would they pick up on nuances and thematic patterns that neither I had recognized in my close reading of the dataset nor the scholars and curators had identified in their close readings of the actual primary source documents my dataset was derived from?

Even though I attempted to run the LDA topic modeling and the K-means clustering algorithms on both the full Spanish America dataset and the Indigenous terms dataset, the resulting outputs were only meaningful for the Spanish America dataset. The outputs for the Indigenous terms dataset were quite random, which made them difficult to interpret. I think this may have been due to the fact that the text corpus derived from the Indigenous terms dataset was significantly smaller than that of the Spanish America dataset and the algorithms may have required larger text corpora than what ended up being input. For comparison of scope, the full Spanish America dataset led to a vectorized vocabulary list of 13041 tokens while the Indigenous terms dataset resulted in a smaller vectorized vocabulary list of 9426 tokens.

Focusing on the Spanish America dataset and running an LDA topic modeling algorithm for 10 topics, I was able to identify several key topics within the clustering results. These included sea voyages and traveling expeditions; European navigation and cartography; Aztec history and

[37] Dennis Channing Landis, *The Literature of the Encounter: A Selection of Books from European Americana: Catalogue of an Exhibition* (Providence, RI: John Carter Brown Library, 1991), 18.

religion; saints and the Catholic Church in Mexico; natural sciences and botany; and conflict/warfare between colonizers and Indigenous populations (see table 1, appendix D for a detailed table of my observations for each topic). When comparing the results of both my own close reading of the dataset and the themes identified by scholars in the exhibition catalog, it was pretty clear that the LDA topic models very closely resembled the major themes humans had previously observed in this dataset. While the themes in the exhibition catalog were much broader—for instance, "missions and religious history" and "geography and history"—they closely paralleled some of the more specific themes the algorithms picked up on, such as Aztec religion, the Catholic Church, navigation, and cartography. The results of the LDA topic modeling served to enhance and reinforce the thematic interpretations scholars have already taken when engaging with the primary source material represented in this dataset.

Following my application of LDA topic modeling to the Spanish America Dataset, I shifted to using k-means cluster analysis. Once again, I tried running this algorithm on both the Spanish America and Indigenous terms datasets, but the clusters that resulted from the Indigenous terms dataset were very random and did not show clear patterns that I could discern as meaningful. The clusters for the Spanish America dataset, however, did indicate some clear patterns that could be critically analyzed (see table 2, appendix E for a detailed table of the themes interpreted from each record cluster). Implementing the k-means clustering algorithm resulted in eight clusters. Surprisingly, five of the eight clusters consisted of image records that were derived from the same books. For instance, in Cluster 5, all 159 records that were clumped together were images from a single text, *Nova plantarum, animalium et mineralium mexicanorum historia,* describing the flora and fauna of sixteenth-century Mexico to a European audience. Two of the remaining clusters

(clusters 3 and 6) were more open ended and contained records that included the terms “Indigenous peoples” or “Indians” as a subject area or subject heading field. The last remaining cluster, which was also the largest, contained 1127 image records. It proved difficult for me to distinguish a clear pattern or unifying topic within this cluster but there may still be some underlying connection the algorithm detected that may require further investigation in order to decipher.

When comparing the outcomes of applying LDA topic modeling and K-Means clustering algorithms to my dataset, I think the LDA topic modeling was more successful in identifying relevant topics that could speak directly to scholarship that has already engaged with this collection material. Specifically, having the exhibition catalog as a basis for comparison helped me gauge the relative success of the algorithms in identifying topics that would be seen as significant to scholars from disciplines such as colonial Latin American history that work with these primary sources.

Conclusion and Avenues for Further Research

There are many discoveries to be made from computationally analyzing the Archive of Early American Images. One way I would like to extend this current research project is to conduct a comparative close and distant reading of the dataset to see how Eurocentrism and colonial power dynamics are at play on both the individual record level and the collection level. Due to scope and time constraints, this project had to focus very specifically on analyzing subject heading, description, and notes fields within the dataset. There are over seventy metadata fields in this dataset, however, and future avenues of research include identifying description trends across centuries and visualizing the ebb and flow of particular themes across space and time. I am deeply interested in integrating timeline and map elements into my analysis in order to make the most of the content-rich metadata records to 1) geo-reference places of publication, 2) trace the stylistic influence of specific engravers and illustrators such as Theodor de Bry, and 3) conduct more nuanced word frequency and unsupervised machine learning studies, experimenting with different topic and cluster sizes, for instance, and comparing results.

One vital contextualizing element that I would need for further investigation would be administrative metadata about the collection. For instance, the current dataset I am using does not contain crucial metadata documenting record creation and modification practices. In order to extend my current research on subject headings and descriptive metadata within universalized knowledge organization systems, I will need more contextualizing information about the people involved in record creation for this database (catalogers, curators, student assistants, etc.) and the workflows and record standardization practices they followed for inputting records into the database.

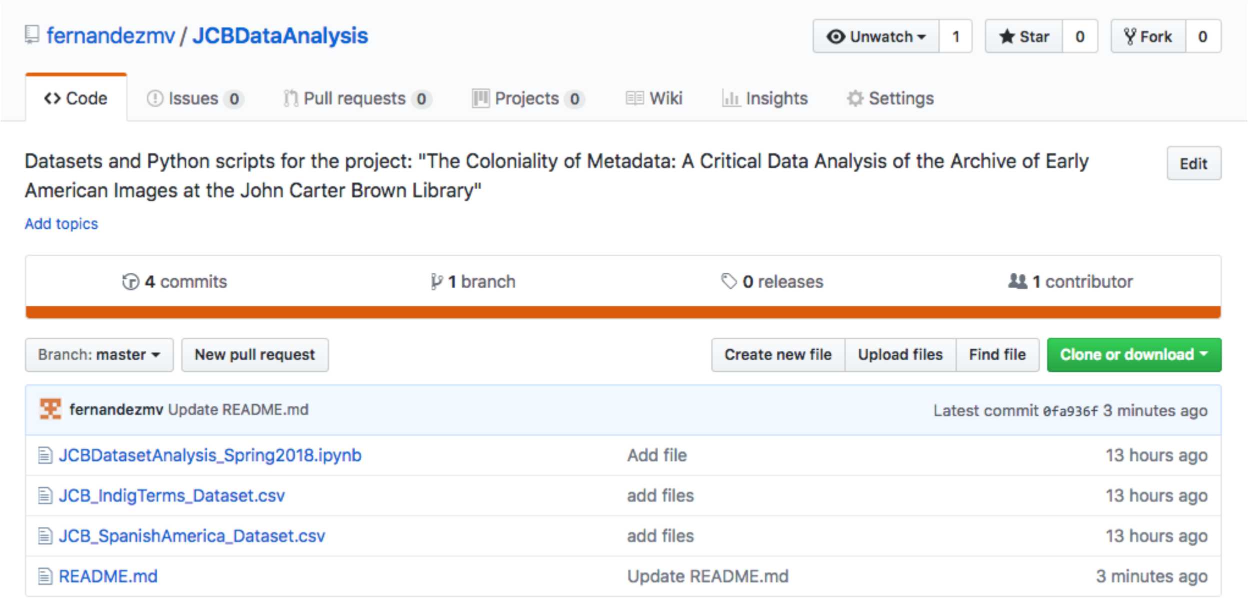
In order to carry out my data analysis, I depend on counting word frequencies and applying unsupervised machine learning algorithms to my dataset. While I was successful in identifying broad metadata trends used across this collection using these methods, I realized that these results would also have to be paired with close readings of individual records in order to develop the necessary evidence to support the broader trends found through distant reading. For instance, while it is significant that "Indians—treatment of" appeared as one of the top 20 subject headings used in the Spanish America dataset I analyzed, this evidence has to be paired with close readings of image records tagged with that subject heading to show how this subject heading remains a vestige of colonialism in the way that it functions as a euphemism for massacre and even genocide of Indigenous communities in colonial Spanish America. In order to make my case framing the JCB's Archive of Early American Images as an archival embodiment of the "coloniality of power," I will need to ground this argument within the close-reading methods that are foundational to Latin American and Indigenous cultural studies. The evidence from my distant reading cannot stand alone to show how these theoretical phenomena, which are seen clearly through close-reading of individual records, permeate broader infrastructures of knowledge organization and classification.

A central lesson I learned while developing this data analysis project has been the increasing importance of engaging with collections as data in order to pursue traditional lines of humanistic inquiry in new ways. As cultural heritage institutions like the JCB move beyond static web images of their digitized collection materials and experiment with new forms of digital engagement and discovery, they are responding to the demands of new research methods that depend on the power of computational analysis to help answer profoundly humanistic questions. In the case of the Archive of Early American Images, computational analysis can begin to shed light on how Eurocentrism and the coloniality of power embedded in colonial primary sources are reflected in the contemporary metadata infrastructures used to describe them.

Appendices

Appendix A: Datasets and Python Scripts Used for Computational Analysis

The Spanish America Dataset, the Indigenous Terms Dataset, and a Jupyter Notebook containing all the Python scripts used in this computational analysis are fully available at the JCB Data Analysis GitHub Repository: https://github.com/fernandezmv/JCBDataAnalysis



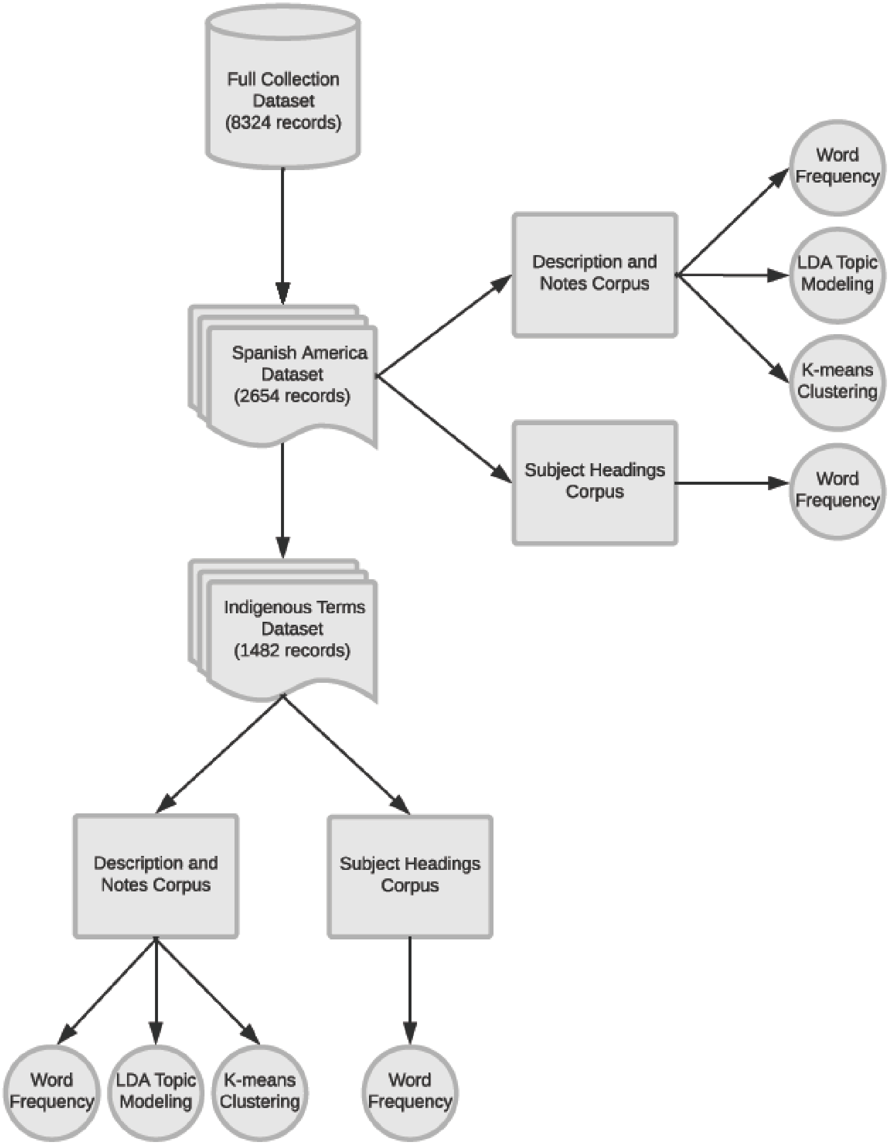
Datasets and Python scripts for the project: "The Coloniality of Metadata: A Critical Data Analysis of the Archive of Early American Images at the John Carter Brown Library"

fernandezmv Update README.md — Latest commit 0fa936f 3 minutes ago

JCBDatasetAnalysis_Spring2018.ipynb	Add file	13 hours ago
JCB_IndigTerms_Dataset.csv	add files	13 hours ago
JCB_SpanishAmerica_Dataset.csv	add files	13 hours ago
README.md	Update README.md	3 minutes ago

Appendix B: Data Analysis Workflow

Figure 1: Data Analysis Workflow



Appendix C: Word Frequency Visualizations

Figure 2: Frequency of Top 20 Subject Headings in the Spanish America Dataset

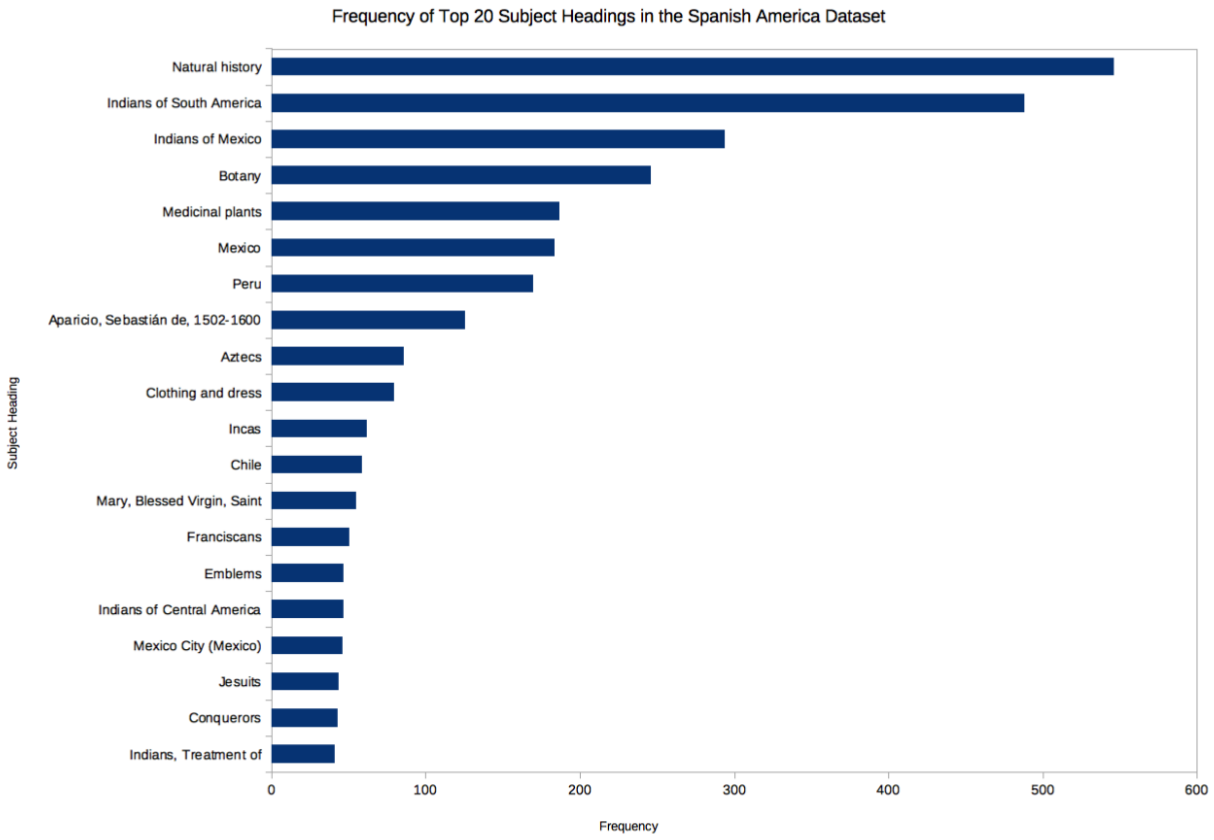


Figure 3: Frequency of Top 20 Subject Headings in the Indigenous Terms Dataset

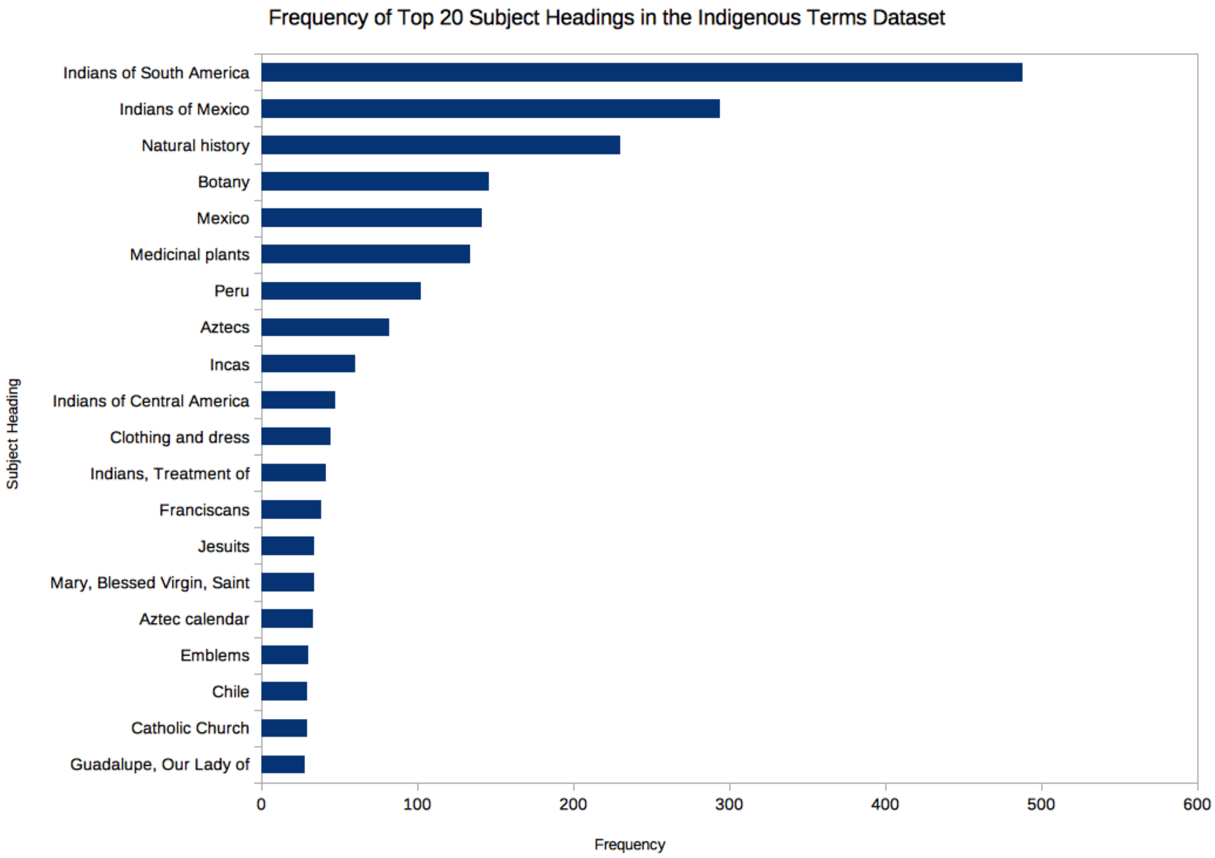


Figure 4: Word Frequencies of the Top 50 Words in the Description and Notes Text Corpus Derived from the Spanish America Dataset

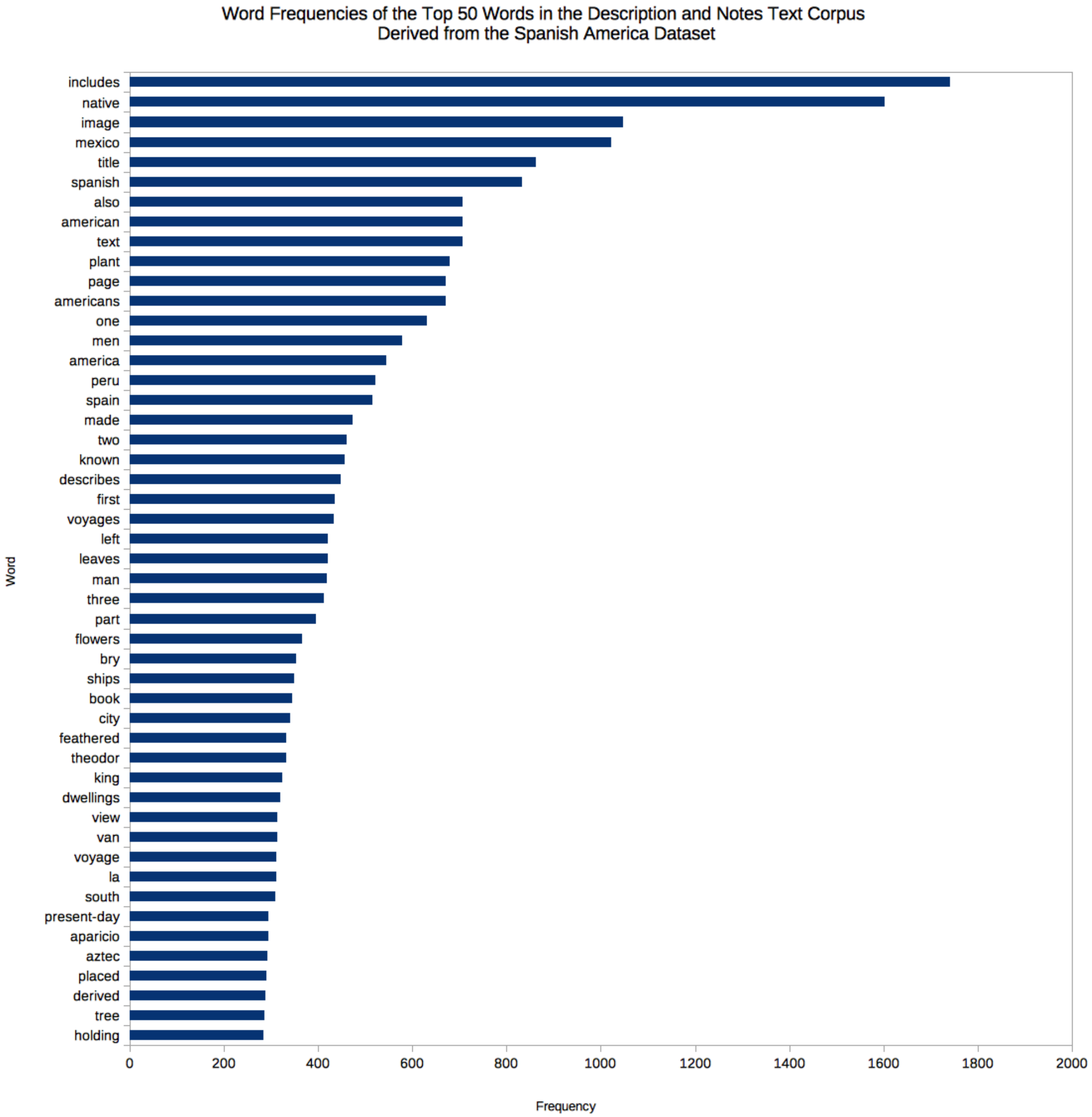
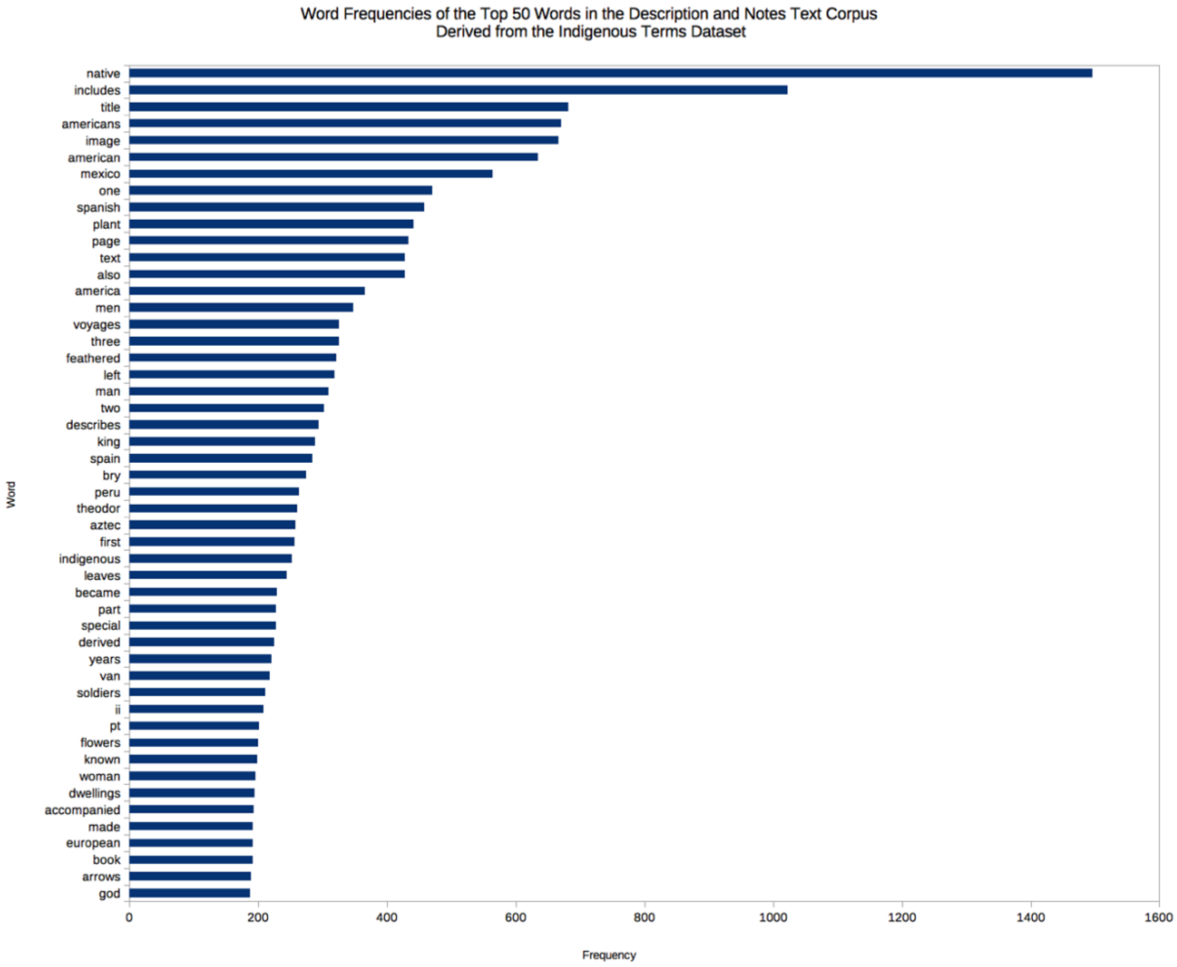


Figure 5: Word Frequencies of the Top 50 Words in the Description and Notes Text Corpus Derived from the Indigenous Terms Dataset



Appendix D: Unsupervised Machine Learning: Latent Dirichlet Allocation (LDA) Topic Modeling

Table 1: Latent Dirichlet Allocation (LDA) topic modeling of a text corpus derived from the description and notes fields of the Spanish America Dataset

Topic	Human interpretation/observations about topic
0	Sea voyages, traveling expeditions, European exploration of the Americas
1	Aztec history, Aztec religion and spirituality, Tovar Codex
2	Symbolism of Aztec Mexico; Meeting of Hernán Cortés and Moctezuma
3	Indigenous clothing and dress
4	Saint Sebastián de Aparicio, Catholic Church, Mexico
5	Natural sciences, botany, Louis Feuillée (famous 17-18th century French explorer, astronomer, geographer, and botanist who documented plant varieties on research expeditions to the Spanish colonies)
6	Warfare and conflict between Spanish colonizers and Indigenous populations
7	Sea voyages, ships, traveling expeditions, European exploration of the Americas
8	Botany, medicinal plants, collection of specimens by naturalists, Francisco Hernández (naturalist and court physician to the King of Spain)
9	Maps, cartography, topography, navigation, voyages, ships

- Size of vectorized vocabulary list: 13041
- Total item records in dataset: 2654

```
Topic 0:
title de page voyages special van herrera this tordesillas separate parts
author collection image antonio pagination register follows arranged period
consists covering leiden 127 1246 1696 chronologically native 1706 en et des
paris part observations peru journal 1725 botaniques physiques mathematiques
chile door togten los made taken gedaan na scheeps

Topic 1:
aztec god manuscript temple left two image calendar mexican de three la sun
tovar one mogrovejo lima huitzilopochtli also mexico year month section
divided may peru second sections toribio placed alfonso found book gods page
days main war september 1580 traveled includes english months day letters
natives paris horizontally engraver

Topic 2:
native includes describes text americans men feathered mexico cortés american
man moctezuma headdresses america also de image mexicans musical spanish
```

```
people eagle part hernán european instruments aztec snake top ii cactus
emperor arrows bottom two probably bry would garments made bird raft aztecs
background river city pt dance ore work

Topic 3:
image native includes american page placed horizontally woman man book
wearing made text two river bound volumes view originally issued also bolivia
century 11 peru headdress fascicles 90 author 1847 typically 1834 holds one
women spanish dress argentina feathered holding men paris people americans
lima include bow peoples title explored

Topic 4:
de mexico aparicio includes roads spain order franciscan known born sebastián
angel made joined traveled building carts goods carry age carriages arms 70
beatified fortune substantial 1787 also book saint coat lima brother
sebastian church holding portrait lay tertiary virgin background cross angels
crown san man francisco monk bishop mary

Topic 5:
text plant french tree scientific botanist flowers name fruit leaves used des
also peru known describes showing académie sciences native pierre member
several attributed feuillée items served royal louis expeditions 1723 father
françois engraving scientist 1647 giffart identification flower seeds roots
woodcuts america lettered image found chile branch engraver south

Topic 6:
de native spanish americans includes soldiers bry theodor america derived
pizarro men inca part one scene swords pt peru work spears governor atahualpa
european warfare background spaniards image dwellings american first guns
gold natives present del published title panama day king benzoni diego pedro
historia bows attack francisco muskets arrows

Topic 7:
includes ships de view city voyage image men mexico items identification
dwellings boats dutch magellan van america also town one strait lettered ship
drake key present native spanish fleet day first churches text voyages thomas
plan european published early account giants bry expedition showing known
written amsterdam fortifications cavendish south

Topic 8:
plant title leaves hernández mexico left flowers botany three one spain roots
image philip years king indigenous chapter became ii medicinal collected
specimens seven accompanied physician philippines personal body 1571 artists
expert linked humors includes detail leaf top bottom probably tree native
root flower identify may possible hern genus ndez

Topic 9:
spanish elements map include juan de south american island cartographic
includes coast image day voyage present british ship chile sea native anson
la first william world compass english ships islands rio details
topographical scale decorative guadalupe location french rivers rose
settlements spain expedition plata near latitude george maps diego colony
```

Latent Dirichlet Allocation (LDA) topic modeling of a text corpus derived from the description and notes fields of the Indigenous terms dataset

- Size of vectorized vocabulary list: 9426
- Total item records in dataset: 1482

```
Topic 0:
hansen eryngium ethnography adelantado issue embedded ibati acolhuan games
axixcoça pleaded pit orange facsimile adjoining authorization pour huaxacensi
peña bison inscription decoctions 07639 craftsmen 1905 1755 1748 pendulum
failure avocado penco epitome funeral bookseller impeded field coldest pours
eupatorium electricity fools choose estados kotzebue metropolitan chloephaga
malambo geometrically intially mopocho

Topic 1:
moxitania inscription forfeit estados americano electrical ethnography
chepauri hansen grey december figurar islay avaramo alms marguerite
adelantado crest jiménez pampa acaciae 1531 misguidedly alive chamaeleonidae
buckled maypo founded mar pomet beetles outina pour american albion admired
cave beltran howler dusicyon caught morantes dammed adjoining chichona
participants 1654 branta geographic forcibly

Topic 2:
goosander hansen caroni funeral gregorillo avaramo adelantado chepauri
finishing edge annointing orange kotzebue erect ingapirca fireworks
guapalaches ethnography columbian hope crest gonzalo brussels brick aiphanes
ethnographies penco canabalism participants advertisement decoctions
bullfighter mines dugout plinia lantana memorializes marquez initiation
malinche 1481 nurses administer eryngium coche averrhoa order flails castles
eighteen

Topic 3:
hansen plinia ethnography adelantado gate mount completion ethnographic mines
moxitania cuitlacopalli additamentum 37 acolhuan cobra deterioration checkers
keeps original cooked fearful libertador chepauri aranda bench fireworks
mineria hope lantana adjoining beyond meet deafness penco gifts maipo bucket
chichona cornucopias furnaces malambo games calash impetus colombini
hucipochotl michoacan face preparation im

Topic 4:
outina chepauri plinth ibati muerte origin bursera imprint eryngium mesa
icarus loading dam folia inland edited 1624 alongside cascade candlesticks
09519 09274 palquin bramins aiarmango 1634 confesses physics hansen coahuila
orange pagan ethnography friendly descend haiti gluttony maytensillo clara
adjoining moreno impetus porcupine evangelio promote cuitlacopalli crest
flails describe funerary

Topic 5:
hansen adjoining chepauri ethnography banded ore additamentum moxitania
gluttony chirripo choose moreno orange impetus landscape mayor adelantado
anonymous mucuna games iva cooked poineer happel keeps pr mowed assasinated
colombini crest dust lantana dignitaries dragon acolhuan ewer ayres malambo
```

alphonso checkers haas ii eryngium butchers passed hope poet eggs depict notorious

Topic 6:
annointing dieu oil garsilasso eryngium goosander hansen nigra battista ornament par adelantado enduring annona penco color equestrian hope foals guarani doctor gonzalo cleric mending plantain poet brick basketwork prison mahu brothers inga mendonça misbound mindanao cheese portrayals acolhuan difficulties mollusk issue narciso different atlat palisade plaque ibati guaraní electrical fermented

Topic 7:
chepauri hansen eryngium ethnography goosander adjoining movement cuitlacopalli outina hope heads andrew acolhuan avaramo prague fireworks crest penco eupatorium grates brick nurses fermented finishing prisoner adelantado atomaria anonymous decaptitated married cultivating descent branta conquests frederick alone plinth colombini gasca moxitania kotzebue floribus alphonso gifts eggplant martyr kings meats december mexicas

Topic 8:
item outina edible flute goosander cylinders foals ornament evangelize avoid michoacán instruction movement arizona erect hope figurar glauca burriel 35379 altera coronado 1527 prisoner instructs intended australia insana mulder eryngium blood manríquez forster engravers ethnography chronologically passage flowed palisade aware hansen gifts cuÿ̆u harp ecuador jolloxcochitl equator babylonian manos kotzebue

Topic 9:
chepauri physics landscape orange adopted pr powder consort dell hansen conquered hondius arizona hemisphere additamentum mecaxochitl assist colleges plinth cloud peña original conquer decorative pliny 1546 pour 1549 ehecatl albara furnaces circled gesture heads 1536 manuscripts cuitlacopalli ghost 1551 bean knee destruens address draw cans evangelio jounael pilosiuscula appendix felled

Appendix E: Unsupervised Machine Learning: K-Means Clustering

Table 2: K-Means Clustering of the Spanish America Dataset

Cluster	# of image records in cluster	Human interpretation/observations about topic/themes in clusters
0	18	All image records are from the same book (JCB call number H719 G322v) : [Giro du mondo. French] Voyage du tour du monde ... Tome sixieme (Paris, 1719)
1	126	All image records are about Saint Sebastián de Aparicio and the vast majority of image records (only 4 record outliers) are derived from the same book *Coleccion de estampas que representan los principales pasos, hechos y prodigios del Bto. Frai Sebastian de Aparizio religo. Franciscano de la Provincia del Sto. Evangelio de Mexico* (Rome, 1789).
2	1127	Could not find distinct pattern or common topic across all records
3	764	All image records contain some subject area entry containing “Indigenous peoples” or a subject heading containing the term “Indians.”
4	113	All records correspond to images derived from Louis Feuillée’s *Journal des observations physiques, mathématiques et botaniques* published in 1714 and 1725. (Feuillée was a famous 17-18th century French explorer, astronomer, geographer, and botanist who documented plant varieties on research expeditions to the Spanish colonies)
5	159	Botany, flora and fauna of 16th century Mexico. All records correspond to images derived from *Nova plantarum, animalium et mineralium mexicanorum historia* ... by Francisco Hernández (1517-1587), naturalist and court physician to the King of Spain.
6	210	All image records either contain some subject area containing “Indigenous peoples,” a subject heading containing the term “Indians,” or a reference to “native Americans” in the notes and description fields.
7	137	All image records are derived from the same book, *Naaukeurige versameling der gedenk-waardigste zee en land-reysen na Oost en West-Indiën ... zedert het jaar 1492 tot 1499* (Leiden, 1707).

Bibliography

“Archive of Early American Images: The John Carter Brown Library.” *The John Carter Brown Library.* Accessed February 15, 2018. https://jcb.lunaimaging.com/luna/servlet/JCB~1~1

Arias, Santa. “Coloniality and Its Preoccupations.” *Latin American Research Review* 48.3 (2013): 214-220. doi: 10.1353/lar.2013.0042.

Beall, Jeffrey. “Ethnic Groups and Library of Congress Subject Headings.” *Colorado Libraries* 32, no. 4 (2006): 37-44.

Berman, Sanford. *Prejudices and Antipathies: A Tract on the LC Subject Heads Concerning People*. Jefferson, N.C.: McFarland & Co., 1993.

Biswas, Paromita. “Rooted in the Past: Use of “East Indians” in Library of Congress Subject Headings.” *Cataloging & Classification Quarterly* 56, no. 1 (2018): 1-18.

Bone, Christine, and Brett Lougheed. “Library of Congress Subject Headings Related to Indigenous Peoples: Changing LCSH for Use in a Canadian Archival Context.” *Cataloging & Classification Quarterly*, 56, no. 1(2018): 83-95. doi: 10.1080/01639374.2017.1382641

Brew, Chris. “Language Processing: Statistical Methods.” In *Encyclopedia of Language & Linguistics*, Edited by Keith Brown. Amsterdam: Elsevier Science, 2006. 597–604.

Chris Brew, “Language Processing: Statistical Methods,” in *Encyclopedia of Language & Linguistics* ed. Keith Brown (Amsterdam: Elsevier Science, 2006), 603

Delpeuch, Antonin. “Clustering in Depth.” *Open Refine.* Last modified September 27, 2017. https://github.com/OpenRefine/OpenRefine/wiki/Clustering-In-Depth

Drabinski, Emily. "Queering the Catalog: Queer Theory and the Politics of Correction." *The Library Quarterly: Information, Community, Policy* 83, no. 2 (2013): 94-111. doi:10.1086/669547.

Duarte, Marisa Elena, and Miranda Belarde-Lewis. “Imagining: Creating Spaces for Indigenous Ontologies.” *Cataloging & Classification Quarterly* 53, no. 5-6 (2015): 677-702.

Dudley, Michael Q. “A Library Matter of Genocide: The Library of Congress and the Historiography of the Native American Holocaust.” *The International Indigenous Policy Journal* 8 no. 2 (2017). doi: 10.18584/iipj.2017.8.2.9

Greenblatt, Ellen. “Homosexuality: The Evolution of a Concept in Library of Congress Subject Headings.” In *Gay and Lesbian Library Service* edited by Cal Gough and Ellen Greenblatt. Jefferson, NC: McFarland (1990).

“History of the Library.” *The John Carter Brown Library*. Accessed February 15, 2018. https://www.brown.edu/academics/libraries/john-carter-brown/about/history-library

“Image Collections: The John Carter Brown Library.” *The John Carter Brown Library.* Accessed February 15, 2018. https://jcb.lunaimaging.com/luna/servlet

“John Carter Brown Library Digital Collections.” *The Internet Archive.* Accessed February 15, 2018. https://archive.org/details/JohnCarterBrownLibrary

Landis, Dennis Channing. *The Literature of the Encounter: A Selection of Books from European Americana : Catalogue of an Exhibition*. Providence, RI: John Carter Brown Library, 1991. http://books.google.com/books?id=7IY7AAAAYAAJ.

Lee, Deborah. “Indigenous Knowledge Organization: A Study of Concepts, Terminology, Structure and (Mostly) Indigenous Voices.” *Partnership: the Canadian Journal of Library and Information Practice and Research* 6, no. 1 (2011): 1-33.

“Luna API Documentation.” *LUNA Documentation.* Accessed February 15, 2018. https://doc.lunaimaging.com/display/V73D/LUNA+API+Documentation

Maldonado-Torres, Nelson. “Colonialism, Neocolonial, Internal Colonialism, the Postcolonial, Coloniality, and Decoloniality.” In *Critical Terms in Caribbean and Latin American Thought: Historical and Institutional Trajectories.* Edited by Yolanda Martínez-San Miguel, Ben Sifuentes-Jáuregui, and Marisa Belausteguigoitia. New York: Palgrave Macmillan, 2016: 67-78.

Mendieta, Eduardo. “Remapping Latin American Studies: Postcolonialism, Subaltern Studies, Post-Occidentalism, and Globalization Theory.” In *Coloniality at Large: Latin America and the Postcolonial Debate*. Edited by Mabel Moraña, Enrique Dussel, and Carlos Jáuregui. Durham: Duke University Press, 2008: 286-306.

Mignolo, Walter D. “Epistemic Disobedience, Independent Thought and De-Colonial Freedom.” *Theory, Culture & Society* 26, no. 7-8 (2009): 1-23. doi: 10.1177/0263276409349275

Moorcraft, Heather. “Ethnocentrism in Subject Headings.” *The Australian Library Journal* 41, no. 1 (1992): 40-45.

Moraña, Mabel, Enrique Dussel, and Carlos Jáuregui. “Colonialism and Its Replicants.” In *Coloniality at Large: Latin America and the Postcolonial Debate*, Edited by Mabel

Moraña, Enrique Dussel, and Carlos Jáuregui. Durham: Duke University Press, 2008: 1-22.

Nuckolls, Karen A. “LC Subject Headings, FAST Headings, and Apps: Diversity Can Be Problematic In the 21st Century” In *Rethinking Technical Services.* Edited by Bradford Lee Eden. Lanham, M.D. : Rowman & Littlefield, 2016: 87-94.

Olson, Hope A. “Patriarchal Structures of Subject Access and Subversive Techniques for Change.” *Canadian Journal of Library and Information Science.* 26, no. 2/3 (2001): 1–29.

———. *The Power to Name: Locating the Limits of Subject Representation in Libraries.* Dordrecht, Netherlands: Kluwer Academic Publishers, 2002.

Quijano, Aníbal. “Coloniality of Power and Eurocentrism in Latin America.” *International Sociology* 15, no. 2 (June 1, 2000): 215–32.

———. "Coloniality of Power, Eurocentrism, and Latin America." Translated by Michael Ennis. *Nepantla: Views from South* 1, no. 3 (2000): 533-580.

———. “La modernidad, el capitalismo, y América nacen el mismo día.” *Boletín Ilia* 10 (January 1991): 42-57.

Roberto, K.R. and Sanford Berman. *Radical Cataloging: Essays at the Front*. Jefferson, N.C.: McFarland & Co, 2008.

“Terminology.” First Nations & Indigenous Studies Program The University of British Columbia. 2009. http://indigenousfoundations.web.arts.ubc.ca/terminology

Yeh, Thomas Yen-Ran. “The Treatment of the American Indian in the Library of Congress E-F Schedule” *Library Resources & Technical Services* 15, no. 2 (1971): 122–126.