NOTES ON OPERATIONS

Text Mining Bibliographic Metadata for Inclusivity

Analyzing Most Frequent Words in Titles, Summaries, and Subjects

Janelle Bitter

Academic libraries have embraced diversity, equity, and inclusion (DEI) principles as core tenets for serving their users. Many of these libraries have undertaken a diversity audit of their collections, evaluating content as well as authorship and amending acquisition processes to increase representation of historically marginalized groups. Techniques used in an audit can include comparison to bibliographies and peer institutions, but few libraries have used text mining of bibliographic metadata to uncover the inclusivity of their collections. This article describes one such study, performed at Raritan Valley Community College, to determine whether language displayed in the title, summary, and subject fields was inclusive and welcoming to library users. Prompted by a new functionality available for WorldCat Discovery that would allow for local updates to problematic subject headings, the process involved uploading MARC metadata to Voyant Tools to learn the most frequent terms in each bibliographic field. Results demonstrated that while the metadata includes welcoming language, improvements could be made by updating subject headings, deaccessioning outdated titles, and educating users in navigating the library catalog.

Raritan Valley Community College (RVCC), an associate degree-granting institution in Branchburg, New Jersey, serves approximately 6,500 students, including dual enrollment students, older adults, first-generation college students, immigrants, Black or African American students, and Hispanic or Latin students. RVCC recently attained the status of Hispanic-Serving Institution, which is granted to institutes of higher education whose undergraduate FTE is at least 25 percent Hispanic. RVCC's Evelyn S. Field Library serves students, faculty, and staff as well as individuals living and working in the two counties affiliated with the college. With this diverse set of library patrons in mind, RVCC's librarians make efforts to ensure that programming, instruction, and collections reflect and appeal to a community with a wide range of backgrounds and interests.

Promoting diversity, equity, and inclusion (DEI) in all areas of library operations has been a guiding principle of America's professional library associations.[1] DEI work has been widely adopted at academic libraries across the nation.[2] These efforts are understood to be important and relevant in creating a welcoming campus environment, especially to groups who have been historically excluded from or neglected by libraries.[3] Some libraries have concentrated on assessing and improving the inclusivity of their collections, although there are challenges to doing so.[4] One area that has received attention is Library of Congress Subject Headings (LCSH), which can be Eurocentric, outdated, or offensive, resulting in a need for alternative vocabulary options.[5] OCLC, the global library cooperative responsible for WorldCat, has contributed to inclusivity efforts, taking actions such as their 2022 release of a local subject remapping function. Libraries that use OCLC's discovery service (WorldCat Discovery) can choose to locally alter the appearance of subject headings in the public interface. As part of the

Janelle Bitter (janelle.bitter@raritanval.edu), Assistant Professor—Systems and Technical Services Librarian, Raritan Valley Community College.

Library Resources & Technical Services | October 2024
https://doi.org/10.5860/lrts.68n4.8329

WorldCat Discovery Diversity, Equity, and Inclusion initiative, libraries can view, contribute to, and locally use a spreadsheet of recommended terms that "aims to reduce harm in item description."[6] This is especially beneficial to libraries that use OCLC's library services platform WorldShare Management Services (WMS), of which RVCC is one, as they cannot make local edits to bibliographic metadata. These libraries' holdings are set on the WorldCat record (previously called the Master record), which means they cannot replace globally used terminology without affecting the thousands of libraries that utilize WorldCat.

When this affordance was released, librarians at Raritan Valley Community College were eager to update subject headings in WorldCat Discovery to improve inclusivity. However, they were also interested in learning what words were most prominent in other metadata fields. They hoped to determine how welcoming the collection was based on the terminology included in such fields. Because the author was interested in digital humanities techniques, it was decided that text mining would be used to evaluate the language in metadata descriptions. Text mining has been a lesser used approach to analyzing collection diversity, as delineated in the following literature review.

Literature Review

To situate this study in the context of relevant literature, the author decided to review articles in two categories: text mining monographic metadata and academic librarians assessing collections for diversity and inclusion. The literature in this latter group includes comparisons against bibliographies as well as analyses of authors, subject matter, and acquisitions processes. The author noted that text mining monographic metadata was a less common undertaking than performing a diversity audit of library collections. Even less common were analyses that addressed both; three studies fell into both of these categories and will be discussed first.

Text Mining for Diversity

Although the primary goal of researchers at the Indiana University Pervasive Technology Institute was enhancing metadata, their work on the public domain portion of HathiTrust also uncovered whether male and female authors were equally represented.[7] After extracting MARC fields 100 and 700 from the nearly 3 million records, the research team used several methods of name matching to determine gender, since most records of interest included author names. Researchers made API calls to VIAF (the Virtual International Authority File) and compared names against US Census data, baby name websites, and a set of names from an earlier study to establish the gender of each author. Although not all methods were reliable, the study indicated similar numbers of male and female authors. The gender information was stored in a Solr index, allowing users to access author gender by searching author name as well as other fields.

In another study on author gender using names, librarians at the University of California, Irvine used text mining to determine how many history monographs (Library of Congress classes C-F) in their collections were written by women.[8] After exporting MARC records from Ex Libris's Alma, their library

management system, they used the program C# MARC Editor to create a .csv file, in which each row represented a book and each column a bibliographic field. They uploaded the data from several fields, including author name, to the free online digital humanities platform Voyant Tools to learn which terms appeared most often. Names like John, Robert, and David were most common, although researchers did not make a declarative statement about gender and acknowledged the need to compare names against a name registry database.

Evaluating a different element of monographic metadata, Jordan Pedersen, a metadata librarian at the University of Toronto, used bibliographic data to determine diversity of content location.[9] Pederson undertook this study to determine whether parts of the world were over- or under-represented in the library's collections. After evaluating MARC control fields, fields that use thesauri, and free text fields, Pederson selected field 651 subfield $a (geographic names), a field that uses a thesaurus, concluding it was a middle ground between control fields and free text. The researcher wrote a Python program and used an SQLite database to determine which countries and regions were best represented in the library and found expected results: the United States and United Kingdom appeared most often in the metadata, with Europe, the Americas, and Asia being far more represented than Africa and Oceania.

Text Mining Bibliographic Metadata

In addition to the above-mentioned studies, several other researchers performed text mining techniques on bibliographic metadata. This approach to research can tackle large quantities of data and reveal previously unthought-of questions and answers. The article "Toward a Metadata Generation Framework" describes the efforts of researchers at Johns Hopkins University to develop and use ANAC (Automated Name Authority Control).[10] This tool was designed to identify the authorized names for individuals mentioned in free text statements of responsibility (MARC field 245 subfield $c) from items in the Lester S. Levy Collection of Sheet Music. Their goal was to automatically generate controlled name metadata, and while the tool was successful more than 50 percent of the time, they concluded that it was not generalizable and would not benefit other collections. In a similar study focused on music, Weitz et al. analyzed free text in MARC statements of responsibility and notes fields (500, 505, 508, 511, 520) to enhance coded name fields (7XX) through the addition of Relator Terms (subfield $e).[11] This work was performed on all of the musical sound recordings and scores cataloged in WorldCat, which was 19 million bibliographic records. Researchers matched names and roles to their controlled versions as well as refining roles that were too generic.

Another article on text mining in the field of music was published in 2016. Tuppen, Rose, and Drosopoulou describe their study on bibliographic datasets from RISM (Répertoire International des Sources Musicales) that totaled more than 1 million records.[12] Using the program MarcEdit, the team selected desired metadata fields for export as tab-delimited records, which they analyzed in Excel. They used additional tools, such as Google Fusion Tables and OpenHeatMap, to further study the text data. This research had several objectives, including determining how certain composers' works had spread geographically and over time. Researchers also noted the vast number of composers who were prolific throughout history but whose works were no longer performed.

Jillian Tomm and colleagues at McGill University, like the team of librarians from UC Irvine, used Voyant Tools, the same digital humanities platform as the author of this article, to analyze textual data from bibliographic records.[13] Researchers at McGill focused on pre-nineteenth century materials in their special collections by exporting MARC records from their catalog and using MarcEdit to save the data in tab-delimited format. Their approach was more experimental, determining what questions they might ask as they viewed the data in new formats, such as word clouds. They mined the titles to learn, for example, how use of a word changed over time. They were also able to determine the distribution of works across nations and decades. This new knowledge led them to identify themes for displays or collection development and provided researchers with new ideas to pursue.

Assessing Collection Diversity

Academic librarians have taken multiple approaches to performing collection diversity audits that have been undertaken by academic libraries. Ciszek and Young provide an overview of some diversity audit techniques, including subject analysis, comparison to bibliographies, usage statistics, surveys, and focus groups.[14] Researchers at other libraries have taken up many of these methods and have used the results to amend collection development policies as well as plan outreach and instruction.

Several studies compared library holdings against award lists or bibliographies of diverse titles. Pettingill and Morgan built a list by combining "Ethnic Studies Reviews" from ACRL's *Choice Reviews* series with several other bibliographies about ethnic groups or multiculturalism.[15] Delaney-Lehman used book reviews (including *Choice*), internet lists, periodical articles, online discussion groups, and reference works to compile a bibliography for comparison at Lake Superior State University.[16] Proctor compared the holdings of Pennsylvania State University Libraries (PSUL) to lists of Lambda Literary and Stonewall Book Award winners, which are granted to LGBTQ books and authors.[17] Kristick pulled together diversity literary award winners from sources such as the American Library Association, LibGuides from other institutions, and an internet search.[18] Monroe-Gulick and Morris built on Kristick's award list to add thirty-four diversity awards that were part of GOBI's Adult Awards Program for their comparison against the University of Kansas Libraries' collections.[19] Bradley-Ridout, Mahetaji, and Mitchell undertook a study with a unique scope.[20] Calling their list-comparison process a "reverse diversity audit," the authors assessed the dermatology collection at the University of Toronto's Gerstein Science Information Centre to ascertain whether diverse skin tones were represented. Materials referenced to develop the list included academic literature, library resource guides, and other websites.

Some researchers focused on assessing the content of the books in their collections in ways other than or in addition to comparison against bibliographies. Backowski and Morton shared the efforts of two libraries in their article.[21] At the University of Wisconsin–Eau Claire, Backowski's research determined which e-books in the collection contained content on topics such as nondominant cultures or identities, then used COUNTER reports to assess usage statistics. Morton's portion of the report, which took place at the University of Virginia, detailed efforts to learn whether books in the collection about Africa or African nations were published on that continent. Proctor used several methods to analyze the collections at PSUL for LGBTQ-related materials.[22] The study compared holdings containing subject

headings such as Gay, Lesbian, Queer, and Transgender against peer libraries who were also in the Association of Research Libraries' 2016 list of top-ten ranked institutions. Salem assessed the childrens picture book collection at San Diego State University (SDSU) for Black, Indigenous, and People of Color diversity using a tool called DBF CAT, or Diverse BookFinder.[23] The study compared character representation to the ethnicity of students at SDSU and those enrolled at San Diego County public schools.

Other studies examined authorship in their collections, measuring the representation of characteristics such as race/ethnicity, Indigeneity, gender expression, and sexual orientation. A main concern in these cases was that the collections' authors reflected the diversity of the student populations. All three studies focused on only a portion of their collections. At Monash University, Manuell, McEntee, and Chester assessed items such as books, e-books, and audio-visual materials in the design collection by researching the identity and location of the author and the location of the publisher.[24] In addition to investigating whether authors were Indigenous or first peoples, women or non-binary, or persons of color, the researchers also determined whether authors and publications from the Global South were included, and if smaller publishers or nontraditional methods of publishing were represented. Stone analyzed playscripts at UC Irvine Libraries, seeking demographic information of playwrights and whether the demographics of their collection had changed between 2011 and 2019.[25] The first step in this process involved searching GOBI for titles purchased each year. Stone then reviewed playwrights' websites, publishers' websites, Wikipedia, and New Play Exchange, a digital library of playscripts by living writers, to determine each playwright's nationality, ethnicity, gender, and sexual orientation. Lastly, Emerson and Lehman focused on print books by a single author published in 2000 or more recently at Augustana College's Thomas Tredway Library.[26] The team decided to research author gender, sexuality, race, and ethnicity as they felt these parts of one's identity provide them with a unique perspective. To learn this information about the authors in their collection, they turned to primary sources like personal blogs, social media, book dedications, and author interviews, only using secondary sources like university websites when the researchers knew the author was still closely affiliated.

The studies undertaken by Kristick, Stone, and Monroe-Gulick and Morris also evaluated their libraries' acquisition processes. Kristick discovered that many diversity-award-winning titles were published by the Big Five publishers (Hachette, HarperCollins, MacMillan, Penguin Random House, and Simon & Schuster) or independent presses.[27] The Oregon State University Library's collection development relied on approval plans from the book vendor YBP (now GOBI) for titles primarily from university presses; therefore, they held a low percentage of diverse books from the bibliography developed for the study. Stone determined that all the publishers from which UC Irvine purchased playscripts were based in North America or England and published playwrights from their geographic region.[28] The research also revealed which publisher's output included the greatest percentage of works by authors from underrepresented groups, such as women, playwrights of color, and LGBTQ+ individuals. Monroe-Gulick and Morris only analyzed orders from GOBI and learned that the highest percentage of award-winning titles held by their library came from approval plans, not firm orders.[29] Further,

their bibliography of award-winning books included five publishers with at least ten winners; their library held none of those books because the publishers were not part of GOBI's profiling program and therefore had to be identified in another way.

Methodology

This study used text mining techniques to learn which terms were most prevalent in certain metadata fields with the goal of determining whether the language was welcoming to users of the library catalog. The first step in analyzing the inclusivity of bibliographic data was downloading MARC records for print books and e-books held by the Evelyn S. Field Library, OCLC Symbol SOC, using a Query Collection in OCLC's Collection Manager. The selection criteria li:SOC AND x0:Book was used to construct the appropriate Query Collection, resulting in the export of 223,884 records in three files, due to file size limits. Next, the program MarcEdit was used to create separate files for titles, summaries, and subject headings, which would be used for text mining. MARC records were prepared using the Export Tab Delimited Records tool, which involved normalizing field 245, subfields $a and b (title); field 520 (summary); and field 650, subfields $a, b, c, d, e, g, v, x, y, z, and 3 (subjects). These tab-delimited records were saved as text files and reviewed to ensure diacritics were displaying correctly and subfield indicators had been removed. This process was performed three times for each field since there were three separate files of raw MARC records. The three text files for each MARC field were combined, resulting in three final text files, one for titles, one for summaries, and one for subjects, which are the text corpora used in this study. Each was uploaded to the open-source digital humanities platform Voyant Tools to determine the most frequent terms and visualize the text corpus in a word cloud format.

The most frequent 500 words in each text corpus were downloaded from Voyant's Terms tool for further analysis and manipulation.[30] Files were opened in Excel and terms were alphabetized to combine words with the same roots (e.g., "learn" and "learning"), then returned to frequency order. Since combining words resulted in a count of fewer than 500, the next most frequent words were added, and the process repeated until the final number of unique terms for each MARC field was 500. Although Voyant Tools has a function called "Categories" that may have been able to do this, the site states "Categories are a new experimental feature, expect things to go wrong," which led the researcher to decide against using it.[31] Automated lemmatization was also considered, but manual intervention was ultimately preferred for greater control of the chosen terms. Combining word counts for terms with the same root was done at the researcher's discretion and primarily consisted of combining singular words with plural words ("economic" and "economics") or words with and without an ending ("design" and "designing," "high" and "higher"). This was done to ensure the greatest number of unique terms would be reviewed. These actions were based on knowledge of the contents of the collection, understanding of bibliographic records, and awareness of the purpose of the research, which was to uncover how users of WorldCat Discovery may react to the terminology they encounter. For example, "programming" and "programs" from the subjects list were not combined, because "programming" in LCSH refers to "individual and types of computers, microprocessors, and programmable calculators," while "programs"

is used for "works on printed lists providing the order of events and other pertinent information for public presentations."[32] However, "poetry" and "poems" from the titles list were combined, as they carry the same connotation for users of a community college library catalog. After completing the task of combining word counts, it was noted that the word "book" appeared in about half of the summaries. Because this widespread use would provide little specific insight into the contents of any given book, it was decided to remove "book" from the summaries list to leave space for another more unique term. The subject list contained a number of words in French, as well as a few in other languages, which were combined with their English language translations. For French words with ambiguous meanings (for example, "histoire" can mean "history" or "story" in English), Répertoire de vedettes-matière de l'Université Laval (RVM), the French subject heading vocabulary prevalent in WorldCat, was searched to learn how the word was used as a subject heading.

Initial Observations and Hypotheses

Titles

The most apparent hypothesis after seeing the word cloud and reviewing the most frequent terms in the titles corpus was that RVCC held many works that promoted student success, containing words such as "development," "education," "guide," "handbook," "introduction," "learning," "practice," and "research." It was also noted that "women('s)" appeared 2,432 times while "men" or "man" came up only 1,196 times. This may mean that our collection elevates the experiences of women, but likely points to the normalization of men, where they are not named because they are the presumed group. Other researchers have also addressed the issue of markedness of women in metadata, most focusing on subject headings.[33] While Voyant Tools ignores some parts of speech in English, such as articles and pronouns, it does not do the same for terms in foreign languages, unless manually configured to do so. It was observed that German words such as "das," "für," "mit," "von," and "zur" were in the top 500 terms, while similar parts of speech in other languages (primarily Spanish, spoken by many of our students) were not. This indicates a potential Northern European preference in our collection's foreign-language titles, which may be a product of the ease of acquiring these materials over others. In addition, words indicating nationality heavily favored America over other nations (see table 1). This, too, may indicate a preference for particular cultures. Finally, the researcher noticed a wide variety of disciplinary terms

Table 1. Occurrence of national or geographical terms in book titles

National or Geographical Term	No. of Occurrences in Titles Corpus
america(n)(ns)('s), u.s, usa	12,235
china, chinese	2,255
europe(an)	2,222
africa(n)	1,707
german(y)	1,428
india(n)	1,215
france, french	1,191
english	1,169
asia(n)	1,091
japan(ese)	1,063
british, uk	837
italy	759
spain	550
latin	476
canada	374

relating to varied academic subjects, such as “art,” “business,” “communication,” “computer,” “design,” “economics,” “education,” “engineering,” “environmental,” “finance,” “health,” “history,” “language,” “mathematics,” “medical,” “music,” “political,” “social,” “science,” and “technology.” It was hypothesized that these terms indicate the library has a balanced collection covering many disciplines taught at the college.

Summaries

The list of most frequent terms in the book summaries, as with the titles list, contained terms related to student support, such as “guide,” “help,” “learn,” “research,” “tools,” “students,” and “understanding.” However, there were also words that indicated a marketing or advertising approach to describing the book, including “make,” “new,” “offers,” “practical,” “provides,” and “work.” While summaries play a valuable role in connecting users to resources through keyword searching, this finding led the researcher to question whether this type of representation would benefit community college students, who may be persuaded to choose an inferior title based on subjective language. This potential barrier to successful research could be exclusionary to first-time, first-generation college students who are unfamiliar with using bibliographic metadata.

Subjects

As with the titles list, there were high word counts for terms that seemed to promote the United States, such as “united,” “states,” “etats,” and “unis.” The prevalence of the French language name for the US (États-Unis) in our subjects may also indicate the Northern European preference hypothesized previously, which is not representative of our student body. North American, British, and Western European libraries also contribute a disproportionally high number of records to WorldCat, so these subject headings in our catalog may be a product of global overrepresentation, not our collection development decisions.[34] Another similarity to the title field was the high occurrence of “women” (19,662) over “men” (2,103). This disparity, like that in the titles, could be due to several factors.

As the initial motivation for undertaking this research was determining which outdated subject headings should be replaced in WorldCat Discovery, the text in MARC field 650 was compared to the subjects listed on the Cataloging Lab’s “Problem LSCH” webpage.[35] This was challenging to do using the list of top terms from Voyant Tools, as many subject headings are multi-word phrases and Voyant Tools treats each word individually in the Terms tool.[36] However, hypotheses for further investigation could be developed by examining the most frequent terms. The word “climatic” had 2,056 appearances, which likely means the confusing subject heading “climatic changes” is frequently used in the library catalog. “Indians” appeared 5,823 times, and when combined with the French “Indiens” totaled 8,025 appearances. This implied that “Indians of North America” may be a common subject heading. “Juvenile” was in the catalog 4,776 times, which could mean that “juvenile delinquents” was utilized many times as a subject heading. The term “race” appeared 5,536 times in subject headings, so “race relations” or “race riots” may be widely used at the Evelyn S. Field Library. Problematic single-word subject headings, such as “prisoners” (1,368 counts) and “slavery” (1,804 counts) were also in the list of most frequent terms.

Findings and Discussion

Initial hypotheses were developed based on the most frequent words in each bibliographic field. These hypotheses were tested in a few ways, including close reading of sentences and text strings, and comparison of works across Library of Congress classifications. For the latter process, MarcEdit was used to export tab-delimited files from the same set of MARC records, this time containing title (field 245), summary (field 520), and subjects (field 650) together, along with field 050 subfields $a and b, which contains the Library of Congress call number. This field was chosen rather than the call number in the local holdings record because nearly 75 percent of the records in the Evelyn S. Field Library catalog represent e-books, for which we do not provide call numbers locally. This new file was saved as an Excel spreadsheet, and, along with the original spreadsheets exported from MarcEdit, was searched using Excel's Find function in ways detailed below. In addition to the spreadsheets, WorldCat Discovery and Voyant Tools were used to glean further information.

Reflecting on the manner in which words were combined in the initial process, it may have been better to either eliminate non-English terms from the study altogether or analyze them separately. Although it was relevant for the purposes of this research to note that the non-English terms were primarily French and German, the inclusion of those languages may have skewed the analysis. For one, the default setting in Voyant Tools is to ignore parts of speech like articles, pronouns, and prepositions, but only in English. The inclusion of these types of words in other languages took up space in the list of top words that could otherwise have been utilized by terms with more substance. Furthermore, unless a library user is bilingual, words with the same meanings in different languages would not evoke the same response. For example, someone who is not a francophone would not see "Angleterre" and think of England, but the frequency of those two words was combined, implying they had the same effect and resulting in a higher number.

Titles

"Women" was found more frequently than "men" in the titles at the Evelyn S. Field Library, and it was determined by searching the titles spreadsheet that both possible explanations were true. There were instances where the word "women" in a title was used to elevate the experiences of women, for example, *Women Who Shaped History*; *America's Working Women*; *Women in Congress, 1917–2006*; and *Women Warriors of the Afro-Latina Diaspora*. However, the word "women" could also be used to treat women as a novelty, or an unexpected subject, as in the titles *Fear of Women*; *Women: Their Changing Roles*; *Women and Achievement*; and *Women, Work, and Computin*g. The researcher acknowledges that these assessments are subjective and that some titles could be viewed as fitting into both categories. In learning more about these titles by searching for them in WorldCat Discovery, it was determined that the titles in the former group were published more recently than those in the latter (1966, 1976, 2006, and 2012; and 1968, 1973, 1975, and 2003, respectively). While this is a small sample of the titles held by the library, it indicates a need to assess the recency of our collections and consider weeding outdated titles with little usage.

In more closely examining the nationality observation in the titles list, where "America" was more prevalent than other countries, the author discovered that "English" most often referred to the language. This provides further strength for the argument that the library materials are highly focused on American or Western culture. In addition, "Indian" meant Indigenous North Americans far more often than it referred to the Indian Ocean or anything about the nation India. As RVCC continues to make efforts to diversify the curricula, library materials should be assessed to ensure they reflect the cultural diversity in coursework.

The hypothesis that varied disciplinary terms in the titles list suggested that our collection was well-balanced among subject areas was determined to be misleading. While the Evelyn S. Field Library collection may be balanced, the appearance of particular terms in titles had no real significance, as many of those words were used in a fashion that was more like conversational language than academic vocabulary. Using Excel's Find tool in the spreadsheet of titles, subjects, summaries, and call numbers, disciplinary terms were searched for in the titles column. The author noted that in many instances, these titles did not correspond to their anticipated subject matter. This was determined by reading the summaries and viewing the call numbers for each record. A prominent example of this is the word "history," as books in our collection are about "a history of" numerous topics that fall outside the disciplines of American and world history (which would be in the Library of Congress classes D, E, or F). Examples of this include *Cockroaches: Ecology, Behavior, and Natural History* (LC class QL505.5); *Insights from Accounting History* (HF5611); and *The Secret History of the War on Cancer* (RC268.25). "Art" and "science" were commonly used together to discuss the "art and science" of a topic that was neither art nor science, for example, *The Art and Science of Business Valuation* and *The Profitable Art and Science of Vibratrading*. Those terms were also used individually in the same manner, as seen in *The Art of the Deal* and *The Science of Reading*. "Design" was used in similar ways, such as in the titles *Ethics by Design: Strategic Thinking and Planning for Exemplary Performance, Responsible Results, and Societal Accountability* and *Guided Inquiry Design*, which are not about design as an artistic pursuit. This lack of true meaning in title terminology points to the importance of controlled vocabularies for identifying what a book is about, and the need for information literacy instruction so students can make informed decisions about which books are most relevant to their needs.

Summaries

One method used to assess the impact of marketing or commercial language observed in the summaries list was to determine what percent of records with summaries contained the words "make," "new," "offers," "practical," "provides," and "work." The Excel spreadsheet containing titles, summaries, subjects, and call numbers was sorted by field 520 (summary), and all rows lacking content in that field were removed. This decision was made because records without summaries could not possibly contain any of the terms in question, and including those records would result in lower percentages and misrepresent the potential significance of this issue. Of the total MARC records used in this research, 182,400 contained 520 fields. Before performing any calculations, each term was spot-checked using the Find function in Excel to read examples of summaries containing those words and verify whether each term was consistently used in a marketing manner. The author decided that "make" and "work"

Table 2. Use of marketing terminology in book summaries

Marketing Term	No. of Occurrences in Summaries Corpus	No. of Books with Summaries Containing Term	% of Books with Summaries Containing Term
new	68,959	49,531	28%
offers	17,970	16,415	9%
practical	21,764	18,161	10%
provides	33,396	28,890	16%

were used in too many other contexts and would skew the data, so they were left out of further steps in the process. The Find All function in Excel was used to learn the number of cells in which each term appeared. Because Find All counts the number of cells and not the number of instances, it provided an accurate count of summaries, and therefore books, containing each word: if a word appeared more than once in the same summary, it would only be counted once since each summary was in a single cell. Table 2 details the occurrence of the terms "new," "offers," "practical," and "provides" in the summary fields across the entire collection.

Additionally, the appearance of these terms was compared across Library of Congress classes to learn whether a certain subject area was more affected by this type of language use. For this process, the rows in the spreadsheet discussed in the paragraph above were sorted by call number (field 050) and divided into twenty-one different sheets in the Excel workbook, one for each letter of the alphabet that represents a Library of Congress class (there are no classes under I, O, W, X, or Y). It should be noted that 177,634 of the records contained a 050 field with at least a Library of Congress class in it: 4,654 did not have any content, five had a Dewey Decimal classification, two had an ISBN, and 105 additional records said "Internet Access." Only those with a Library of Congress class or full classification were used in this process. The number of rows in each sheet was recorded, and the terms were searched for again. Rather than focusing on the total number of times a term appeared in each class, when classes contained drastically different numbers of books, a percentage of books containing each term was calculated for every letter of the alphabet. Then, a comparison was made to learn which classes had the highest percentage of summaries containing each term. Class T (technology), had the highest percentage of "practical," the second highest of "provides," and the third highest of "new." Similarly, class R (medicine) had the highest percentage of "provides" and the second highest of "practical," and class L (education) had the highest percentage of "offers" and the second highest of "practical" (tied). The highest percentage of "new" was in F (American history), but that class had the second lowest percentage of "practical." See figure 1 for this data. Because class F was the only sizeable section with both a very highly and very infrequently used term (class V, with only 123 books, was similar), additional evaluation was undertaken. To determine why the summaries in class F had the highest percentage of the word "new," they were saved as a text file and uploaded to Voyant Tools, then the Contexts tool was reviewed. While there were some instances of "new" being used in the marketing manner addressed above, most uses referred to current places in the United States (New Hampshire, New Jersey, New York), former names for places in what is now North America (New Netherland, New Spain), or common phrases in this discipline (new world, new republic).

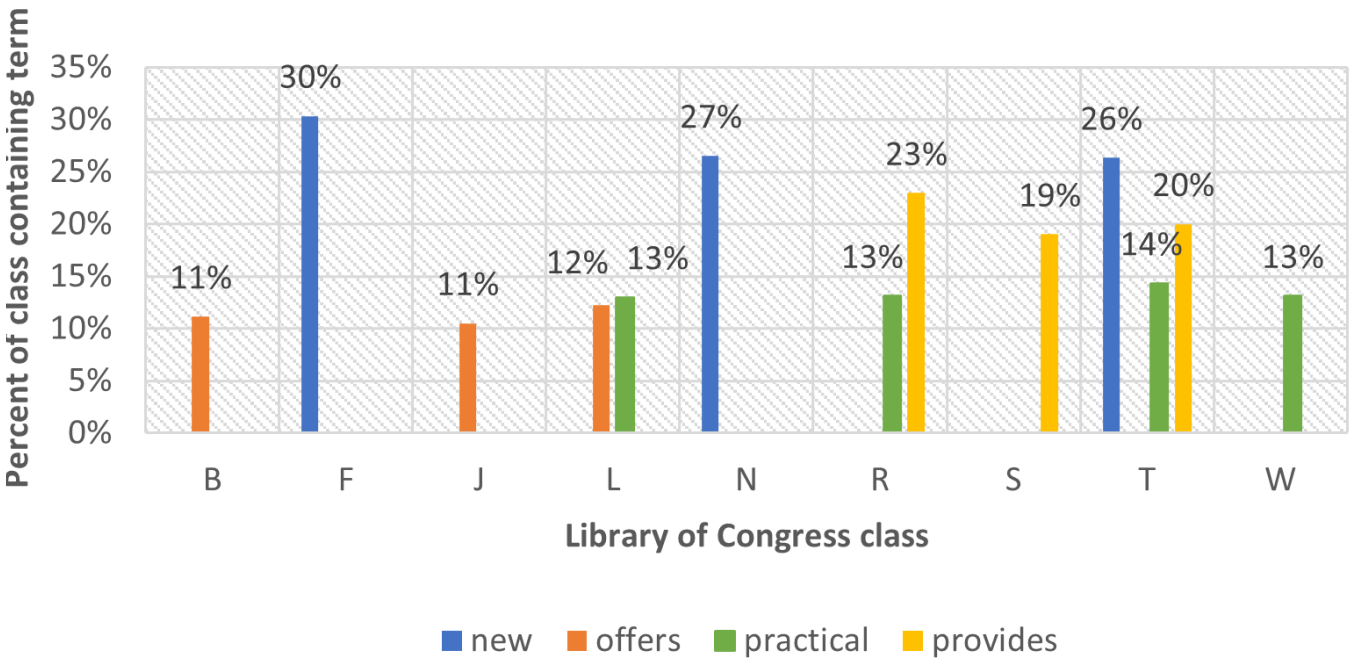

Figure 1. LC classes with highest percentage of marketing terms in summary

Other observations were made when reviewing textual data from the summaries. Terms such as "cover," "jacket," and "publisher" demonstrated that many summaries originate at a location where the goal is to sell the book, not necessarily provide an accurate description of it. These summaries are added to the metadata with quotation marks around them, indicating where they are copied from, but this may not be noticed by a new college student. In reading the context of the words "new," "offers," "practical," and "provides" in full summaries, further subjective language was uncovered. Some examples of this include "answers all the important questions of today," "indispensable guide," "most original," and "surprising insights." As with the 245 field, the ambiguous language in the 520 field demonstrates the importance of controlled vocabulary in learning a book's contents, and the need for instruction on the use of the catalog. The detection that two of the disciplines heavily affected by the marketing language were technology and medicine is highly concerning, as those fields advance rapidly, and a book will not be "new" for long. The perpetual flux and conflict in the field of education also demand that students remain informed and up-to-date.

Subjects

The significant difference between the number of times the terms "women" and "men" were used in the subject field (19,662 and 2,103, respectively) was first investigated using Find in Excel to determine in what situations each word was used. This process entailed making note of various contexts until patterns emerged, and therefore, not all instances of either term were analyzed. It was noted that "women" was commonly used to qualify a noun, implying that the given or assumed state of that noun when not qualified is "men," another example of markedness in bibliographic metadata. This was further supported by the fact that some LCSH do not have a "men" version. Examples from the Evelyn S. Field Library catalog include "women occupations" (there is no "men occupations," although there is

"male nurses," another example of gendered markedness), "mentally ill women" (there is no "mentally ill men"), "women authors" (the opposite gendered term is "male authors"), and "women fiction" (any instance of "men fiction" requires an additional adjective, such as "abusive," "Jewish," or "single"). This leads to an observation about the term "men" in our subject headings, which is that it was frequently further qualified, rather than stand-alone: "African American men," "gay men," "Jewish men," "older men," and "young men."

Table 3. Occurrence of problematic subject headings

Subject Heading	No. of Book Records Containing it
race relations	2,208
Indians of North America	796
climatic changes	743
slavery	722
sex role	645
prisoners	458

To determine which problematic subject headings were most prevalent in the library catalog, the spreadsheet containing just subjects was searched using Find All in Excel. The resultant counts were compared to results from a subject search in WorldCat Discovery to ensure quantities were accurate. However, most subject searches did not exactly equal the number derived from the spreadsheet. One issue is that Discovery has multiple ways of searching by subject (subject and subject phrase searches), which produce different results. Because all other analyses in this research used data that was exported from Collection Manager, formatted in MarcEdit, and manipulated in Excel, it was decided the same data should be used in this phase of the process. Find All was used to count the number of books containing a word or phrase. All subject fields for a book were contained within one cell, and therefore a word or phrase would only be counted once per book, even if it was repeated. The heading "race relations" was significantly more prevalent than other subjects that were searched for. This type of language may be a euphemism for mistreatment of certain groups, and also promotes the concept of race as a legitimate way of categorizing humans. "Indians of North America" was the next most common problematic heading. This phrase reflects colonialist perspectives; a more modern option is "Indigenous peoples of North America." "Climatic changes" is considered problematic because the language is confusing and not widely used, as "climate change" is. Other commonly occurring subject headings in our catalog included "slavery" (some feel "enslavement" is a more person-first term), "sex role" (this conflates sex and gender; the suggested replacement is "gender role"), and "prisoners" (some states use the person-first phrase "adults in custody").[37] See table 3 for this data. Although "juvenile" was a frequently found word in the subjects corpus, analyzing its context indicated that it was used most commonly in "juvenile fiction" and "juvenile literature"; "juvenile delinquents" was only found in eighty-three records.

As with the summary analysis, the most frequent problematic subject headings were compared across Library of Congress classifications. Using a similar procedure, a spreadsheet containing titles, summaries, subjects, and call numbers was first sorted to remove any records lacking a 650 field. Rows were then sorted by call number and divided into different sheets for each letter of the alphabet. Terms were searched for, and percentages calculated for each class to determine which class contained a particular subject heading in the greatest percentage. Classes E and F (both American history) had the

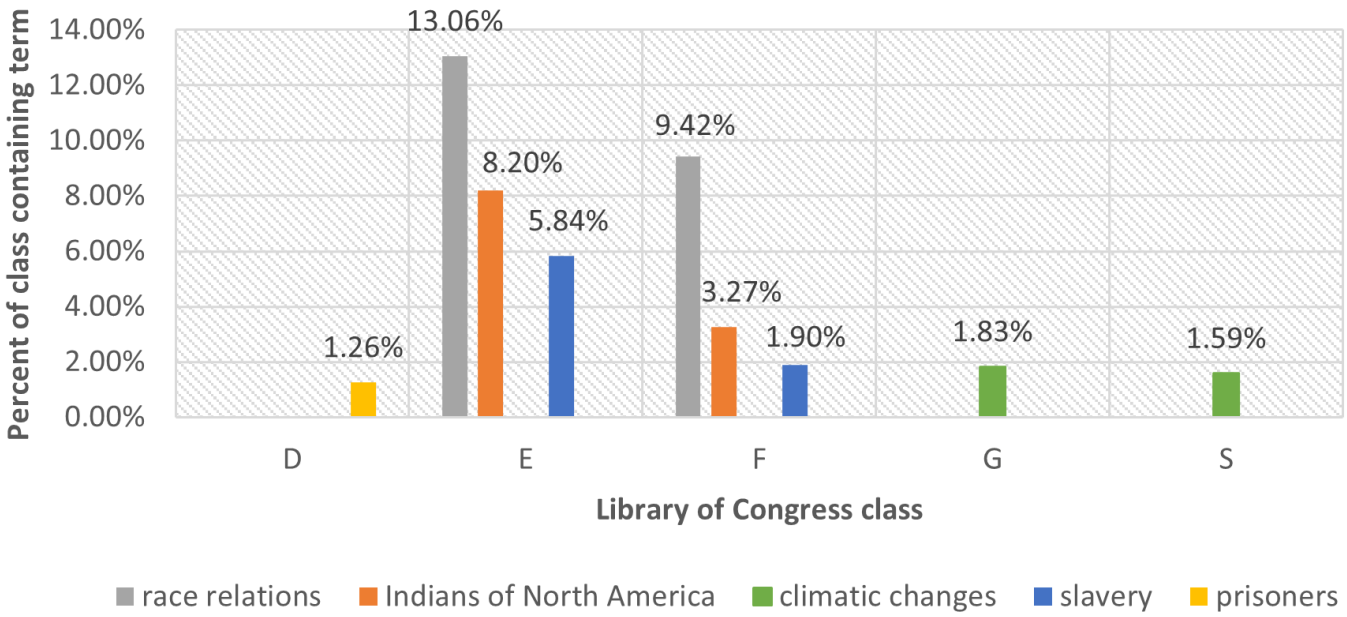

Figure 2. LC classes with highest % of problem subject headings

highest and second highest percentages, respectively, of "race relations," "Indians of North America," and "slavery." "Climatic changes" was most prevalent in class S (agriculture), then class G (geography, anthropology, and recreation). "Prisoners" was only present in more than 1 percent of records in class D (world history). Figure 2 details this data. "Sex role" was present in the highest percentage of records in class H (social sciences), then class E, although both were less than one percent.

In the case of subject headings, information literacy instruction is not enough to ensure our students are using the catalog in a way that helps them with their research and ensures they feel comfortable using the library. Updates must also be made to the subject headings themselves. As a library that uses WMS, our catalog records are the global bibliographic records; we cannot create a local version to remove outdated subject headings. But, using OCLC's recently added function of making the WorldCat Discovery metadata editable, librarians at RVCC can now substitute new terminology for those subject headings that may be offensive, outdated, or confusing.

Conclusions and Future Directions

Analyzing the most frequently used words in titles, summaries, and subjects at RVCC's Evelyn S. Field Library revealed several areas in which improvements could be made, or future research was necessary. Aside from indicating which harmful subject headings had the greatest chance of being viewed in WorldCat Discovery, the analysis also demonstrated a need for user education and for weeding outdated titles. As community college students may be new to the higher education environment and academic libraries, instruction is paramount to ensuring they can access the information they need. Librarians should build an understanding of problematic subject headings and update them based on the new WMS capabilities. They should also provide guidance in navigating the library catalog's metadata fields so students can locate the most pertinent, timely resources. As not all students will

interact with a librarian in their search for information, it is also necessary to evaluate collections and deaccession titles that are outdated. Although the recent capability to edit subject heading displays in WorldCat Discovery is useful, other fields such as summaries and titles cannot (and often should not) be changed. Rather than updating the display of information on a surface level, it may be better to simply remove the book from the collection altogether if it is offensive or outdated. While research institutions may want to retain this type of material, librarians at RVCC often feel that newer titles about historic topics are often better for our students than titles from decades ago. In addition to sharing recent research on a topic and amending outmoded perspectives, these newer titles are often written in a more approachable manner and cataloged with more robust metadata according to improved encoding standards and updated guidelines from AACR2 and RDA.

Although the most frequent word counts are factual data, much of the analysis of what the data means is speculation. To truly understand the effects of confusing or offensive metadata on community college library users, future research should include asking library patrons about their responses. This question-asking could be in the form of surveys, focus groups, or individual conversations with students at the reference desk. Whatever the approach, this quantitative study should be followed with qualitative research to better comprehend the impact of bibliographic metadata.

Notes

1. "Diverse Collections: An Interpretation of the Library Bill of Rights," American Library Association, last modified June 24, 2019, www.ala.org/advocacy/intfreedom/librarybill/interpretations/diversecollections; "Equity, Diversity, and Inclusion," Association of College and Research Libraries, accessed November 21, 2023, www.ala.org/acrl/issues/edi; "Equity, Diversity, Inclusion: An Interpretation of the Library Bill of Rights," American Library Association, last modified June 27, 2017, www.ala.org/advocacy/intfreedom/librarybill/interpretations/EDI; Joint ALA/ARL Building Cultural Proficiencies for Racial Equity Framework Task Force, *Cultural Proficiencies for Racial Equity: A Framework*, August 2022, www.ala.org/advocacy/sites/ala.org.advocacy/files/content/diversity/ALA%20ARL%20Cultural%20Proficiencies%20for%20Racial%20Equity%20Framework.pdf.
2. Freeda Brook, Dave Ellenwood, and Althea Eannace Lazzaro, "In Pursuit of Antiracist Social Justice: Denaturalizing Whiteness in the Academic Library," *Library Trends* 64, no. 2 (Fall 2015): 246–84, https://doi.org/10.1353/lib.2015.0048; Otis A. Chadley, "Addressing Cultural Diversity in Academic and Research Libraries," *College & Research Libraries* 53, no. 3 (May 1992): 206–14, https://doi.org/10.5860/crl_53_03_206; Alice M. Cruz, "Intentional Integration of Diversity Ideals in Academic Libraries: A Literature Review," *The Journal of Academic Librarianship* 45, no. 3 (May 2019): 220–27, https://doi.org/10.1016/j.acalib.2019.02.011; Jenny Lynne Semenza, Regina Koury, and Sandra Shropshire, "Diversity at Work in Academic Libraries 2010–2015: An Annotated Bibliography," *Collection Building* 36, no. 3 (July 2017): 89–95, https://doi.org/10.1108/CB-12-2016-0038; Roberto G. Trujillo and David C. Weber, "Academic Library Responses to Cultural Diversity: A Position Paper for the 1990s," *Journal of Academic Librarianship* 17, no. 3 (July 1991): 157–61, ERIC; Mark D. Winston and Haipeng Li, "Managing Diversity in Liberal Arts College Libraries," *College & Research Libraries* 61, no. 3 (May 2000): 205–15, https://doi.org/10.5860/crl.61.3.205.

3. Myrna Morales, Em Claire Knowles, and Chris Bourg, "Diversity, Social Justice, and the Future of Libraries," *portal: Libraries and the Academy* 14, no. 3 (July 2014): 439–51, https://doi.org/10.1353/pla.2014.0017; Ethelene Whitmire, "Cultural Diversity and Undergraduates' Academic Library Use," *Journal of Academic Librarianship* 29, no. 3 (May 2003): 148–61, https://doi.org/10.1016/S0099-1333(03)00019-3.

4. Matthew P. Ciszek and Courtney L. Young, "Diversity Collection Assessment in Large Academic Libraries," *Collection Building* 29, no. 4 (October 2010): 154–61, https://doi.org/10.1108/01604951011088899; Lori M. Jahnke, Kyle Tanaka, and Christopher A. Palazzolo, "Ideology, Policy, and Practice: Structural Barriers to Collections Diversity in Research and College Libraries," *College & Research Libraries* 83, no. 2 (March 2022): 166–83, https://doi.org/10.5860/crl.83.2.166; William H. Walters, "Assessing Diversity in Academic Library Book Collections: Diversity Audit Principles and Methods," *Open Information Science* 7, no. 1 (July 2023): 60–68, https://doi.org/10.1515/opis-2022-0148.

5. Anna M. Ferris, "Evolution of a Subject Heading: The Story Continues," *Library Resources & Technical Services* 66, no. 2 (April 2022): 66–76, https://doi.org/10.5860/lrts.66n2.66; Rachel K. Fischer, "Using the Homosaurus in a Public Library Consortium: A Case Study," *Library Resources & Technical Services* 67, no. 1 (January 2023): 4–15, https://doi.org/10.5860/lrts.67n1.4; Karen A. Nuckolls, "LC Subject Headings, FAST Headings, and Apps: Diversity Can Be Problematic In the 21st Century," *Law Faculty Scholarly Articles* 555 (2015), https://uknowledge.uky.edu/law_facpub/555; Karl Pettitt and Erin Elzi, "Unsettling the Library Catalog: A Case Study in Reducing the Presence of 'Indians of North America' and Similar Subject Headings," *Library Resources & Technical Services* 67, no. 2 (April 2023): 44–52, https://doi.org/10.5860/lrts.67n2.4; Crystal Vaughan, "The Language of Cataloguing: Deconstructing and Decolonizing Systems of Organization in Libraries," *Dalhousie Journal of Interdisciplinary Management* 14 (2018), https://ojs.library.dal.ca/djim/article/view/7853/7247; Hollie C. White, "Decolonizing the Way Libraries Organize" (paper presented at IFLA World Library and Information Congress, Kuala Lumpur, Malaysia, August 28, 2018), https://library.ifla.org/id/eprint/2221.

6. "Create Locally Preferred Subjects for Display and Search Expansion," OCLC Support, last modified November 29, 2023, https://help.oclc.org/Discovery_and_Reference/WorldCat_Discovery/Display_local_data/Create_locally_preferred_subjects_for_display_and_search_expansion.

7. Zong Peng et al., "Author Gender Metadata Augmentation of HathiTrust Digital Library," *Proceedings of the Association for Information Science & Technology* 51, no. 1 (April 2014): 578–81, https://doi.org/10.1002/meet.2014.14505101098.

8. Sarah Wallbank et al., "Exploring Bibliographic Records as Research Data," *Catalogue and Index*, no. 197 (December 2019): 3–9, https://cdn.ymaws.com/www.cilip.org.uk/resource/collection/F71F19C3-49CF-462D-8165-B07967EE07F0/C&I_197.pdf.

9. Jordan Pedersen, "Measuring Collection Diversity via Exploratory Analysis of Collection Metadata," *Serials Librarian* 82, no. 1–4 (March 2022): 186–93, https://doi.org/10.1080/0361526x.2022.2028499.

10. Mark Patton et al., "Toward a Metadata Generation Framework: A Case Study at Johns Hopkins University," *D-Lib Magazine* 10, no. 11 (November 2004), https://doi.org/10.1045/november2004-choudhury.

11. Jay Weitz et al., "Mining MARC's Hidden Treasures: Initial Investigations into How Notes of the Past Might Shape Our Future," *Journal of Library Metadata* 16, no. 3–4 (December 2016): 166–80, https://doi.org/10.1080/19386389.2016.1262653.

12. Sandra Tuppen, Stephen Rose, and Loukia Drosopoulou, "Library Catalogue Records as a Research Resource: Introducing 'A Big Data History of Music,'" *Fontes Artis Musicae* 63, no. 2 (2016): 67–88, https://doi.org/10.1353/fam.2016.0011.
13. Jillian Tomm, Cheryl Smeall, Christopher Lyons, and Richard Virr, "Excavations in Library Metadata: Drawing from New Tools in Support of Discovery and Valorization in a Changing Special Collections Landscape," *Argus* 42, no.3 (2014): 44–48.
14. Ciszek and Young, "Diversity Collection Assessment."
15. Ann Pettingill and Pamela Morgan, "Building a Retrospective Multicultural Collection: A Practical Approach," *Collection Building* 15, no. 3 (September 1996): 10–16, https://doi.org/10.1108/01604959610126000.
16. Maureen J. Delaney-Lehman, "Assessing the Library Collection for Diversity," *Collection Management* 20, no. 3–4 (1996): 29–37, https://doi.org/10.1300/J105v20n03_05.
17. Julia Proctor, "Representation in the Collection: Assessing Coverage of LGBTQ Content in an Academic Library Collection," *Collection Management* 45, no. 3 (2020): 223–34, https://doi.org/10.1080/01462679.2019.1708835.
18. Laurel Kristick, "Diversity Literary Awards: A Tool for Assessing an Academic Library's Collection," *Collection Management* 45, no. 2 (2020): 151–61, https://doi.org/10.1080/01462679.2019.1675209.
19. Amalia Monroe-Gulick and Sara E. Morris, "Diversity in Monographs: Selectors, Acquisitions, Publishers, and Vendors," *Collection Management* 48, no. 3 (2023): 210–33, https://doi.org/10.1080/01462679.2022.2163019.
20. Glyneva Bradley-Ridout, Kaushar Mahetaji, and Mikaela Mitchell, "Using a Reverse Diversity Audit Approach to Evaluate a Dermatology Collection in an Academic Health Sciences Library: A Case Presentation," *The Journal of Academic Librarianship* 49, no. 6 (November 2023), https://doi.org/10.1016/j.acalib.2022.102650.
21. Roxanne Marie Backowski and Timothy Ryan Morton, "Something to Talk About: The Intersection of Library Assessment and Collection Diversity," *Proceedings of the Charleston Library Conference*, 2019, https://doi.org/10.5703/1288284317148.
22. Proctor, "Representation in the Collection."
23. Linda Salem, "How Diverse Is the Academic Library Children's Picture Book Collection? Using Diverse Bookfinder's Content Analysis, Demographic Data, and Historical Bibliographies to Analyze a Picture Book Collection," *Collection Management* 47, no. 2–3 (2022): 115–35, https://doi.org/10.1080/01462679.2021.1960668.
24. Romany Manuell, Kate McEntee, and Marcus Chester, "The Equity Collection: Analysis and Transformation of the Monash University Design Collection," *Art Libraries Journal* 44, no. 3 (2019): 119–23, https://doi.org/10.1017/alj.2019.16.
25. Scott M. Stone, "Whose Play Scripts Are Being Published? A Diversity Audit of One Library's Collection in Conversation with the Broader Play Publishing World," *Collection Management* 45, no. 4 (December 2020): 304–20, https://doi.org/10.1080/01462679.2020.1715314.

26. María Evelia Emerson and Lauryn Grace Lehman, "Who Are We Missing? Conducting a Diversity Audit in a Liberal Arts College Library," *The Journal of Academic Librarianship* 48, no. 3 (May 2022), https://doi.org/10.1016/j.acalib.2022.102517.
27. Kristick, "Diversity Literary Awards."
28. Stone, "Whose Play Scripts Are Being Published?"
29. Monroe-Gulick and Morris, "Diversity in Monographs."
30. These lists of 500 terms can be accessed on RVCC's Library website at https://library.raritanval.edu/ld.php?content_id=76891621.
31. "Categories," Voyant Tools Help, accessed April 25, 2024, https://voyant-tools.org/docs/#!/guide/categories.
32. "Programming," Linked Data Service: LC Subject Headings (LCSH), Library of Congress, accessed December 11, 2023, https://id.loc.gov/authorities/subjects/sh00007512.html; "Programs," Linked Data Service: LC Subject Headings (LCSH), Library of Congress, accessed December 11, 2023, https://id.loc.gov/authorities/subjects/sh93005169.html.
33. Examples include Sanford Berman, *Prejudices and Antipathies: A Tract on the LC Subject Heads Concerning People* (Jefferson, NC: McFarland, 1993), https://www.sanfordberman.org/prejant/prejant.pdf; Judith Hudson and Victoria A. Mills, "Women in the National Online Bibliographic Database," in *Women Online: Research in Women's Studies Using Online Databases*, ed. Steven D. Atkinson and Judith Hudson (New York: Haworth, 1990), 237–57; Joan K. Marshall, *On Equal Terms: A Thesaurus for Nonsexist Indexing and Cataloging* (New York: Neal-Schuman, 1977), https://archive.org/details/onequaltermsthes0000mars; Hope A. Olson, "How We Construct Subjects: A Feminist Analysis," *Library Trends* 56, no. 2 (2007): 509–41, https://doi.org/10.1353/lib.2008.0007; Hope A. Olson, "Patriarchal Structures of Subject Access and Subversive Techniques for Change," *Canadian Journal of Information & Library Sciences* 26, no. 2/3 (June 2001): 1–29; Hope A. Olson, "The Power to Name: Representation in Library Catalogs," *Signs: Journal of Women in Culture & Society* 26, no. 3 (Spring 2001): 639–68, https://doi.org/10.1086/495624.
34. Jay H. Bernstein, "From the Ubiquitous to the Nonexistent: A Demographic Study of OCLC WorldCat," *Library Resources & Technical Services* 50, no. 2 (2006): 79–90, https://doi.org/10.5860/lrts.50n2.79.
35. "Problem LCSH," Cataloging Lab, accessed September 13, 2021, https://cataloginglab.org/problem-lcsh/.
36. At this stage in the research, other tools in Voyant, such as Phrases, were not yet considered. Further, because Excel was used for other parts of this research beyond the counting of most frequent words, it was decided for consistency to use Excel to analyze subject headings.
37. Suggestions for updated subject headings are from Cataloging Lab, "Problem LCSH."