On the State of Genre/Form Vocabulary

A Quantitative Analysis of LCGFT Data in WorldCat

Colin Bitter and Yuji Tosaka

The purpose of this paper is to report on a quantitative analysis of the LCGFT vocabulary within a large set of MARC bibliographic data retrieved from the OCLC WorldCat database. The study aimed to provide a detailed analysis of the outcomes of the LCGFT project, which was launched by the Library of Congress (LC) in 2007. Findings point to a moderate increase in LCGFT use over time; however, the vocabulary has not been applied to the fullest extent possible in WorldCat. Further, adoption has been inconsistent between the various LCGFT disciplines. These and other findings discussed here suggest that retrospective application of the vocabulary using automated means should be investigated by catalogers and other technical services librarians. Indeed, as the data used for the analysis show somewhat uneven application of LCGFT, and with nearly half a billion records in WorldCat, it remains a certainty that much of LCGFT's full potentials for genre/form access and retrieval will remain untapped until innovative solutions are introduced to further increase overall vocabulary usage in bibliographic databases.

When the Library of Congress Genre/Form Terms for Library and Archival Materials (LCGFT) project began in 2007, the principal aim was to develop a vocabulary separate from Library of Congress Subject Headings (LCSH) to describe what a resource is rather than what it is about.[1] While LCSH has been also used to describe "is-ness" for decades in certain situations, there were several problems with using LCSH terms to describe genre and form.[2] Through the efforts of the Library of Congress (LC) partnering with various parties in the greater cataloging community, the LCGFT project has been successful in establishing a separate vocabulary that is both broad and deep. As of March 2020 (when the data were compiled for this study), 2,357 terms are organized under eight disciplines (art, cartography, law, literature, moving images, music, religion, and non-musical sound recordings) plus "general library materials," with twenty-one "top terms" that have other narrower terms organized hierarchically in each category.

The purpose of this paper is to conduct a quantitative analysis of a large set of MARC bibliographic data retrieved from the OCLC WorldCat database (henceforth WorldCat). Previous publications about LCGFT have been primarily limited to providing a broad overview of the history of genre and form and establishing a clear need for a robust genre/form vocabulary, while some have also outlined the process to create the new vocabulary. What is lacking in the literature is a detailed analysis of the outcomes of the LCGFT project within

Colin Bitter (bitterc1@tcnj.edu) is the Head of Cataloging and Metadata at The College of New Jersey. **Yuji Tosaka** (tosaka@tcnj.edu) is Cataloging/Metadata Librarian at The College of New Jersey.

Manuscript submitted January 26, 2021; returned to authors for minor revision March 10, 2021; revised manuscript submitted March 13, 2021; accepted for publication March 18, 2021.

bibliographic databases. Although the primary focus of the following study is on LCGFT terms recorded in MARC field 655 subfield $a, multiple data points within the records are used for the authors' analysis. Filling a clear gap in the literature, such quantitative analysis will provide a broad overview as to the state of LCGFT usage within a shared cataloging environment, and will make significant contributions related to multiple library stakeholders. It will give catalogers a much better, empirical understanding of the extent to which LCGFT has been applied within MARC bibliographic records. Additionally, detailed analysis of the vocabulary usage will offer insights into future cataloging practices and training needs in the technical services community. This paper's findings will also offer useful insights for public services librarians, as they will benefit from learning in depth about patterns of LCGFT application in bibliographic databases for their work with users to help them navigate front end systems utilizing such data for improved resource discovery.

Literature Review

The question of providing access to genre and form information in library catalogs has not received much attention in the library literature, although it has been long recognized as one of the key intellectual foundations of information organization. In his influential *Rules for a Printed Dictionary Catalogue*, Cutter noted that a key objective of the catalog was the collocating function, that is, enabling users to discover all resources in a particular genre or form of material, and by author and subject.[3] Genre and form are also an essential part of the bibliographic universe as defined in the current IFLA Library Reference Model [e.g., LRM-E2-A1: *Category* attribute].[4] For many years, some limited access to the genres and forms found in library collections had been provided by LCSH, either as main headings or subdivisions, although their primary function was always to describe the content of the work (aboutness). By the end of the twentieth century, more recent developments brought increasing attention to the genre/form access question, with the creation of GSAFD (Guidelines on Subject Access to Individual Works of Fiction, Drama, Etc.) genre terms and the implementation of a new MARC subfield $v for "form" subdivisions in 6XX fields.[5] At the same time, LC announced its plan to develop new genre/form headings. And yet, it was not until 2007 that LC finally developed the new LCGFT thesaurus, starting with moving image materials and radio programs.[6] More than a decade after its inception, LCGFT has developed into a more fully fledged controlled vocabulary for genre and form access covering nine disciplines, including "general" materials.[7]

The steady development of the LCGFT thesaurus, however, has not yet yielded a new stream of scholarship on genre/form access in the cataloging literature, although there are several studies that have begun to look into the subject over the last decade. Perhaps the most important overview of the historical literature was provided by Lee and Zhang's 2013 paper in *Cataloging & Classification Quarterly*. The authors traced how genre and form terms had been conceptualized and treated in Anglo-American cataloging standards up to the implementation of RDA. Their comprehensive examination showed that genre had not been given the attention it deserved in the cataloging literature, despite the "expanding role genre plays in the current as well as future environments." Notably, the authors also concluded that the cataloging community had failed to establish clear definitions differentiating "genre" and "form."[8] This conceptual ambiguity is reflected in the current LCGFT manual, which defines both genres and forms as follows:

> Genres and forms may be broadly defined as categories of resources that share known conventions. More specifically, genre/form terms may describe the purpose, structure, content, and/or themes of resources.[9]

While other authors have also investigated issues relating to genre and form access in specific subject areas and specialist communities, such as audiovisual cataloging, there have been few published studies focusing on the LCGFT thesaurus itself.[10] Those few publications include Young and Mandelstam's 2013 paper in *Cataloging & Classification Quarterly*, in which they discussed, in addition to introducing the reader to its potential benefits and applications, how the LCGFT thesaurus was developed, often involving formal collaboration between LC cataloging policy specialists and outside library organizations.[11] Iseminger and others have also considered LCGFT development and applications in specialist communities, such as music cataloging.[12] As adding LCGFT headings to legacy metadata is clearly a very important step in fully realizing the benefits of the new vocabulary, Mullin examined the process for automatically assigning them for music resources retrospectively based on the presence of LCSH terms in their bibliographic records.[13]

Now that more than a decade has passed since the LCGFT thesaurus first became available for use in the library community, recent literature has finally started analyzing data on how the LCGFT thesaurus has been deployed in library catalogs and digital repositories. In 2018, Dragon contacted twenty-nine digital repositories in North American academic libraries and examined how they provided genre and form access for their digital collections, using such display labels as "Format," "Type," and "Genre."

For specific vocabularies being used, she found that DCMI (Dublin Core Metadata Initiative) Type Vocabulary and the Art and Architecture Thesaurus were both most widely used, while LCGFT was used at only two of the repositories that she contacted.[14] In contrast, Bitter and Tosaka decided to focus specifically on the usage of LCGFT headings in library catalogs and conducted a survey that revealed that the new thesaurus had gained wide, if somewhat uneven, adoption in the cataloging community. The survey data reported in their paper provided valuable insights into how the LCGFT thesaurus was currently used in copy and original cataloging practices and which types of resources were more likely to have their bibliographic records enriched with LCGFT terms.[15] Whereas these newer studies serve as good starting points for examining current LCGFT implementation, what is sorely needed in the literature is detailed research on how LCGFT terms have actually been deployed in bibliographic databases, such as local catalogs or WorldCat. This paper's intent is to make a significant contribution to studies on genre and form access by conducting a quantitative analysis of LCGFT usage patterns in selected MARC records retrieved from WorldCat.

Research Method and Data Retrieval

To fill the critical gap in the literature described above, this paper explores several areas of inquiry. Most broadly, dates within bibliographic records are used to investigate rates of LCGFT application over time. Second, format of material is examined to differentiate LCGFT use between different types of records—for example, does notated music contain more LCGFT than projected media? Third, records are grouped by LC classification (LCC) to examine patterns of LCGFT usage in terms of pre-existing classification—do certain areas of LCC see greater use of LCGFT? Finally, LCGFT terms applied within bibliographic records are delineated to analyze the extent to which broader or narrower headings have been used in terms of the hierarchies in LCGFT.

To explore the research questions outlined above, the authors examined MARC bibliographic record data retrieved from WorldCat. As WorldCat is a shared cataloging environment with close to half a billion records used by thousands of OCLC member libraries, analyzing portions of data from this database provided much needed empirical insights into the current state of LCGFT usage in the cataloging community. Although there were many possibilities for record selection, the approach selected for the current study was to examine WorldCat records based on the holdings of the authors' institution, The College of New Jersey, a mid-sized four-year public college in Ewing, New Jersey. The college is a comprehensive institution enrolling approximately 7,000 undergraduate students in a wide range of disciplines, and also offers master's and post-baccalaureate programs for over 600 students in a small number of graduate programs, such as Business, Counseling, Education, English & Humanities, Integrative STEM, and Nursing & Public Health. The authors' library is a typical academic library for a medium-sized institution. The only library serving the campus community, it holds over half a million titles in its physical collections, divided into seven main areas, including Archives, Children's/Young Adult, Curriculum & K-12, General, Music & Media, Periodicals, and Reference. The library also directly manages over 350,000 electronic titles, spread across various electronic collections. The vast majority of MARC records for both the physical and electronic collections are cataloged in WorldCat.

The authors believed that performing an analysis on this set of records selected from WorldCat would lead to a good snapshot of the current patterns of LCGFT usage within bibliographic records typically used by academic institutions. That is, overall patterns of LCGFT usage can be better inferred from this record set since the vast majority of these records are selected via copy cataloging from WorldCat and the authors have made efforts to include only high-quality best matches in their local catalog. That would contrast with analyzing the entire WorldCat database, which would contain a plethora of duplicates, to say nothing of lower quality bibliographic records that the authors feared would make their analysis much more complicated than necessary. Additionally, they decided to avoid analyzing bibliographic records in their local catalog for the obvious reason that those records do not include changes, including LCGFT headings added, since they were last copy-cataloged from WorldCat.

To obtain WorldCat master record data for their library's institutional holdings, the authors first turned to OCLC WorldShare Collection Manager, a cloud-based application designed to promote efficiencies in managing metadata for print and electronic collections held by OCLC member libraries. The feature used in Collection Manager was "query collection," which enabled the authors to retrieve master records for all of their library's local holdings. Using query collection was straightforward, as only a single criterion needed to be specified in the query, "li:NJT," which limited the resulting collection to holdings based on their library's OCLC symbol. Several files retrieved contained their library's entire institutional holdings, which totaled 846,862 records. It initially appeared as if this data could be used for the present study; however, authentic dates and times of latest transactions were not recorded in the MARC field 005. Each field 005 in the retrieved records in the query collection contained the same calendar date, "20200321 . . ." followed by hours, minutes,

seconds, and fractions of seconds, e.g., 20200321084945.4, that is, the download date of each record as WorldCat apparently considers this to be a record transaction date. This seemed to diminish the utility of the retrieved data for the authors' intended analysis because they had expected that the replace date of each record might be queried to expose varying rates of vocabulary application over time. Email communications with OCLC support representatives confirmed that the authentic replace date in the field 005 could not be retrieved via query collection. Although the retrieved data thus could not be used as originally planned, the authors set aside the 846,862 unique OCLC numbers contained in the collection as the basis for future data retrievals, as described below.

The authors decided to use the OCLC Bib API integration in MarcEdit, a freely available leading MARC data editing tool developed by Terry Reese, who is the Head of Digital Initiatives and Infrastructure Support at The Ohio State University Libraries. This tool enables users to retrieve WorldCat master records by OCLC number, ISBN, ISSN, or Title/Author. To use the OCLC Bib API, the authors first needed to contact OCLC to obtain API keys, which were then recorded in MarcEdit. Once the integration was established, it was then possible to use MarcEdit's OCLC Record Downloader and extract the needed MARC records using OCLC control numbers. For the present study, there were two major advantages to using the OCLC API integration in MarcEdit. First, the records delivered contained the authentic field 005, i.e., the last replace date in WorldCat. Second, much larger batches could be processed (the authors generally retrieved 50,000 records during a single session), thereby eliminating the ceiling of 9,999 records that would have been possible via batch searching in Connexion. Despite these advantages, the OCLC API also presented some drawbacks. First, it was highly error-prone—that is, the downloader would typically fail to retrieve every MARC record matching the OCLC number specified in the search. Therefore, it was necessary to cross-check the OCLC numbers in the resultant download file against the original query and then retrieve missing records in a quick follow-up session. The WorldCat master records matching all of the authors' institution's holdings were successfully retrieved in sets of 50,000 records each between April 24 and May 2, 2020. They were combined into a single file of 846,862 records (henceforth referred to as the base file), the contents of which are analyzed in the Analysis section that follows.

Beyond generating this base file, LCGFT terms from the vocabulary itself required organization for the present study. Two files of LCGFT terms were prepared, based on data compiled using *Classification Web* as of February 25, 2020, which were then brought up to date in early May with *Library of Congress Subject Headings Monthly List* 03 (March 16, 2020). The first file created, LCGFT-1, was a single list containing all unique LCGFT terms (2,357 terms). In compiling the LCGFT-1 file, the authors also divided all the LCGFT terms into four levels of hierarchy by applying numbers 1, 2, 3, and 4 to each term based on their hierarchical relationships. That is, 1 was the highest level assigned to the broadest terms (e.g., "Art"), while 4 was the lowest level assigned to more specific, narrower terms (e.g., "Pageants"). These scores were assigned in the LCGFT-1 file so that information about aggregate depth of indexing also could be garnered for LCGFT headings used in bibliographic records. The LCGFT Manual instructs catalogers to "assign terms that are as specific as the genres and forms exemplified in a resource" and some disciplines, such as music, have a well-developed hierarchy of LCGFT headings.[16] The authors recognized that a broader term may be used instead under certain situations (e.g., when a given term may subsume several narrower genre and form terms). They were thus interested as part of their data analysis in identifying the extent to which narrower, specific terms had been assigned in WorldCat records as they evaluated the overall patterns of LCGFT application. Four levels of hierarchy were chosen for the current study as deeper levels of specificity (e.g., fifth and sixth levels) did not seem too productive for analysis. Additionally, as LCGFT is polyhierarchical (i.e., some terms belong to more than one broader discipline, sometimes at different levels), it was necessary to find a consistent way of applying hierarchy levels to terms occurring in multiple LCGFT disciplines and/or at multiple levels of hierarchy. For the purposes of this study, LCGFT terms were coded in the highest possible hierarchy for each discipline. "Loose-leaf services" is an apt example as it occurs at different levels, under both top terms "Law Materials" and "Informational Works." Under "Law Materials," "Loose-leaf services" would be coded 2 since it is a second-level term. Under "Informational Works," it would be coded 3 since it is a third-level term. In the combined LCGFT-1 file, "Loose-leaf services" was coded 2.

The authors also created the second file, LCGFT-2, containing twenty-one separate lists for each of the LCGFT subject categories (art, cartographic materials, commemorative works, creative nonfiction, derivative works, discursive works, ephemera, illustrated works, informational works, instructional and educational works, law materials, literature, motion pictures, music, recreational works, religious materials, sound recordings, tactile works, television programs, video recordings, and visual works). LCGFT terms in these separate lists were also given annotations for depth of indexing respectively, with 1, 2, 3, and 4 assigned in the same fashion as in LCGFT-1. LCGFT-2 was used to examine prevalence and depth of indexing of the vocabulary used in each category, as will be discussed below in the Analysis section.

Findings and Analysis

Date

During the planning phase of the present study, the authors had expected that the MARC field 005 (Date and Time of Latest Transaction) would prove to be a useful data point in analyzing LCGFT usage in WorldCat. That is, as this field functions as a replace date in WorldCat, examining LCGFT usage against field 005 might provide interesting insights into changing rates of vocabulary application. Though they recognized that LCGFT headings for various disciplines were introduced over different years, the year 2007 was chosen as the point of demarcation for this simple exploratory analysis on the grounds that the LCGFT thesaurus was first established in that year. However, the base file used for analysis (864,862 records, as described in Research Methods and Data Retrieval) revealed that all records had been replaced within the last seven years. The oldest field 005 was dated June 6, 2013. Although many of these records would have been upgraded manually by catalogers within Connexion or by various OCLC member libraries via automated means (such as datasync, which automatically generates a new field 005), other records would also have been updated by WorldCat's internal automated processes, such as addition of RDA 33X fields or FAST subject headings.[17] Indeed, as the field 005 did not extend beyond the past seven years, the field was found to be effectively unusable for the intended analysis.

An alternative to replace date that was identified for the authors' analysis was the "Date 1" fixed field, available in the field 008 positions 07-10. Nearly all of the records in the base file had usable Date 1 data. However, some records had to be expunged due to incompleteness (for example, uuuu, ||||, 0002, and similar non-usable data values). After eliminating these records, 838,875 records (99.1 percent) remained and were used for this area of the analysis. Whereas the exact meaning of Date 1 data can vary based on the coding of the DtSt fixed field (Type of Date/Publication Status—008 position 06), the vast majority can be accurately linked to the manifestation being cataloged, be it in form of the year of production, publication, distribution, release, manufacture, or copyright as specified in the code in DtSt.

Using Date 1 values in the base file, the records were divided into two groups: before 2007 and 2007 to the present. Records prior to 2007 numbered 640,449 (76.3 percent of the base file); records from 2007 to the present numbered 198,426 (23.7 percent). First, examining these two sets of records for LCGFT application showed some increase in the latter group, which is not surprising given that LCGFT was not available for use before 2007. As seen in figure 1, 144,045 (22.5 percent) pre-2007 records and 58,489 (29.5 percent) records from 2007–present contained one or more LCGFT terms. Additionally, the average number of LCGFT terms for records containing LCGFT increased slightly, from 1.34 to 1.50. Date 1 values will continue to serve as a point of illumination in the sections that follow.

Format

Format of material was also examined to find disparities in LCGFT application, if any, between various types of resources. From the entire base file, 205,879 records had one or more 655 fields containing subfield $2 lcgft, representing 24.3 percent of all the records under examination. From this set of records, type of record (Leader position 06) was retrieved to examine the format of material described by each record. Table 1 illustrates the proportion of records containing LCGFT based on type. (As there were few resources coded as kit, manuscript cartographic material, manuscript notated music, mixed materials, and three-dimensional artifact or naturally occurring object, these formats are omitted in table 1 as they are not substantively significant for the purpose of this analysis.) Here it is worth noting the high rates of LCGFT application for a handful of format types. Indeed, over half of the records for five types contained one or more LCGFT terms: manuscript language materials (96.2 percent), projected media (88.3 percent), cartographic materials (65.8 percent), notated music (53.5 percent), and two-dimensional nonprojectable graphics (50.7 percent). In contrast, less than half of the records contained LCGFT for musical sound recordings (40.8 percent), nonmusical sound recordings (19.5 percent), language materials (18.8 percent), and computer files (13.1 percent).

Comparing the pre-2007 and 2007–present record sets revealed some other interesting data on changes in LCGFT application across format types. Of 202,534 records containing LCGFT and a valid Date 1 value (as described earlier in the *Date* section), 144,045 (71.1 percent) were pre-2007 and 58,489 (28.9 percent) were 2007–present. These two sets of records were compared against all the records in the base file containing valid dates (divided into two files, pre-2007 and 2007–present) to measure changes in LCGFT application over time. With the exception of musical sound recordings, all types showed an increase in LCGFT application in the 2007–present set, as evidenced in figure 2. The most significant increases were found in notated music (a 34.1 percent increase, from 52.4 percent to 86.5 percent), two-dimensional nonprojectable graphics (21.6 percent, from 44.3 percent to 65.9 percent), nonmusical sound recordings (20.7 percent, from 18.1 percent to 38.8 percent), projected media (15.8 percent, from 83.8 percent to 99.6 percent), and cartographic materials (12.9 percent, from 58.9 percent to 71.8 percent). A less noticeable change was apparent in computer files (8.2 percent,

from 12.3 percent to 20.5 percent) and language materials (6.0 percent, from 17.5 percent to 23.5 percent).

While these increases may be expected, the data also revealed an unexpected *decrease* in LCGFT usage for musical sound recordings (11 percent, from 46.2 percent to 35.2 percent). One possibility here is that pre-2007 materials had received LCGFT terms via retrospective application. Of course, not all materials are cataloged contemporaneously—manifestations predating 2007 could easily have been cataloged well past the initial implementation of LCGFT, although the year 2007 may be a rather arbitrary point of demarcation for this format in particular because LCGFT for musical works were implemented in 2015. Additionally, it might be possible that increased use of batch loading from external providers in the set of records from 2007 to the present may have increased the number of records lacking LCGFT—for example, newer records for streaming sound recordings. Regardless, this surprising result obviously seems to warrant a separate future inquiry. Lastly, it should be noted that kits, manuscript notated music, manuscript cartographic materials, and mixed materials did not present significant changes between the two periods (not presented in figure 2).

Library of Congress Classification

The authors also decided to take a close look at LC classification (LCC) in the data file to see if it might render different insights into patterns of LCGFT application in WorldCat records. Of the records containing one or more 655 fields with $2 lcgft (205,879 records), 158,125 records (76.8 percent) contained one or more LC call numbers. For the set of LCGFT records containing LCC, call numbers were extracted from fields 050 and 090 to perform classification analysis. The records were checked for internal duplication of classes and subclasses. For example, a record containing two instances of ML was only counted once toward ML. Also, for the purpose of the current study, records containing differing LC subclasses or classes were counted in each area; for example, if a record contained subclasses DS and PN, the record counted toward both subclasses and both overall D and P classes. Additionally, invalid call numbers were removed from the data set. For example, the authors found that many 050/090 fields contained Dewey or SuDocs numbers, or textual phrases such as "ISSN RECORD." These types of records were removed, and remaining LC classes could then be trimmed to their first letter alone for the analysis.

Extracting, cleaning, and deduplicating the call numbers from the set of 158,125 records containing LCGFT resulted in 163,067 valid instances of LCC classes. Figure 3 contains the entire distribution of LCC within records containing one or more LCGFT terms. P (language and literature, 28.2 percent) and M (music and books on music, 21.8 percent) represented half of the LCC classes in the authors'

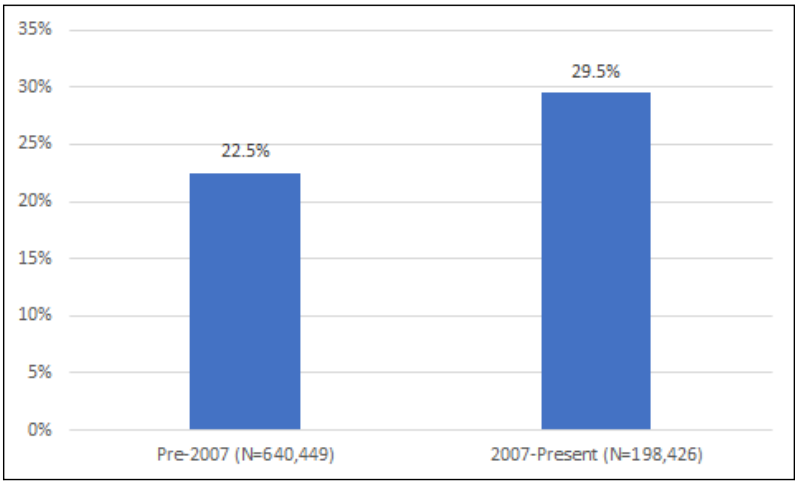

Figure 1. Percentage of Records with One or More LCGFT Term, by Date

Table 1. Percentage of Records with One or More LCGFT Terms by Type of Record

Type	Description	No. of Records with LCGFT	Total No, of Records in File	Percent
t	Manuscript language material	1,433	1,489	96.2
g	Projected medium	14,705	16,661	88.3
e	Cartographic material	4,595	6,979	65.8
c	Notated music	4,327	8,093	53.5
k	Two-dimensional nonprojectable graphic	205	404	50.7
j	Musical sound recording	51,402	125,898	40.8
i	Nonmusical sound recording	347	1,781	19.5
a	Language material	128,713	684,396	18.8
m	Computer file	133	1,013	13.1

base file; K (law, 9.9 percent) and H (social sciences, 7.4 percent) also revealed a moderate amount of representation in the file, followed by Q (science, 4.4 percent), D (world history and history of Europe, Asia, Africa, Australia, New Zealand, etc., 3.6 percent), and E (history: America, 3.2 percent). The remaining fourteen LC classes (each amounting to less than 3 percent) added up to just over 20 percent of the records containing LCGFT headings.

While this provides a broad picture of LCC distribution within the set of records containing LCGFT, a proportional analysis of this data against the entire base file provides a more accurate indication of the rate at which LCGFT has been applied within each class. For this analysis, records containing LCGFT were measured in each LC class against 622,777 records with 684,540 occurrences of valid LC classes from the base file. As seen in figure 4, M (music and books on music, 44.5 percent), P (language and literature, 41.1 percent) and K (law, 33.1 percent) still have high representation of records containing LCGFT; however, Z (bibliography/library science, 41.9 percent) has moved to second place, showing high levels of gene/form application for these resources. Although H (social sciences, 12.7 percent) ranked fourth in the earlier pure distribution, it dropped to the bottom half in the proportional analysis. C (auxiliary sciences of history, 26.4 percent), E (history: America, 26.3 percent), N (fine arts, 24.1 percent), and G (geography/anthropology/recreation, 22.4 percent) also showed moderate levels of LCGFT application.

Examining the number of terms applied by class per record also revealed interesting LCGFT application patterns, as shown in figure 5. For the 163,067 valid instances of LCC, there were a total of 220,668 fields 655 with $2 lcgft, yielding an average of 1.35 terms per record. There was some variability observed within this set; classes P (1.51) and M (1.48) show slightly higher levels of application (about 10 percent higher than the average), while K (1.02)—the LC class with the lowest level of LCGFT application—averaged only marginally higher than a single term assigned per record (about 25 percent lower than the average).

Another relevant area of analysis with regard to LCC was the distribution of LCGFT by Date 1. As described in the previous section on *Date*, the base file was divided into pre-2007 records and records from 2007 to the present. These two files of records were analyzed for LCC and Date 1; any record not containing a valid LCC class or Date 1 was omitted from this area of analysis. This resulted in 683,187 records total (80.7 percent of the base file). Out of this subset, 566,562 (82.9 percent) were in the pre-2007 group and 116,625 (17.1 percent) were in the 2007–present group. These two files were then examined for LCGFT; in the pre-2007 file, 125,630 records (22.2 percent) contained one or more LCGFT terms, while 36,754 records (31.5

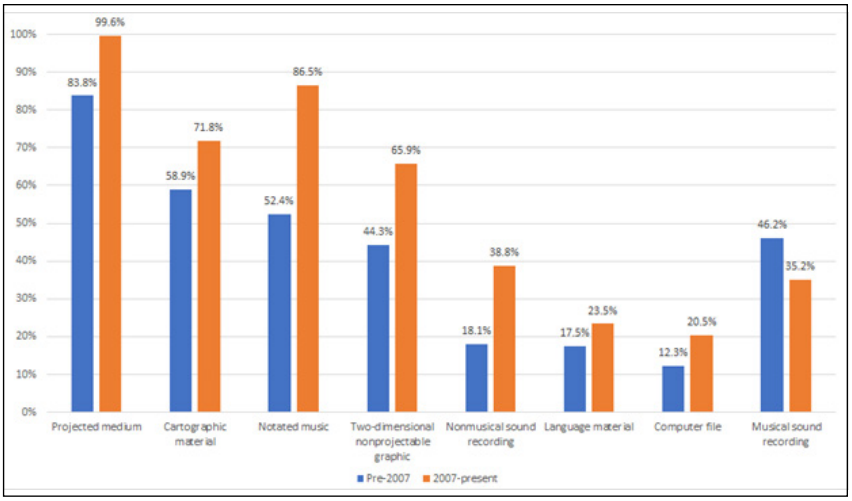

Figure 2. Percentage of Records with One or More LCGFT Terms by Type of Record, Grouped by Date

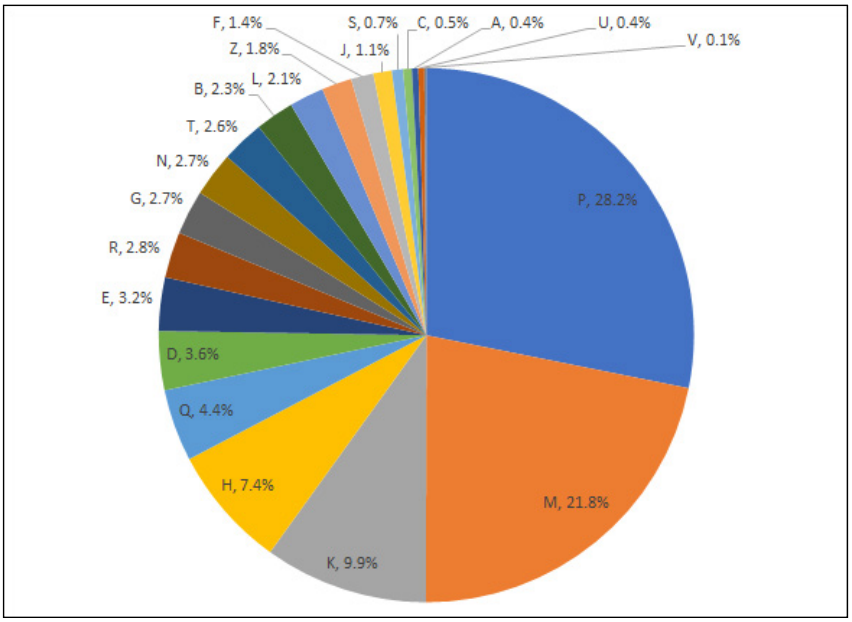

Figure 3. Distribution of LCC in Records Containing LCGFT, by Class (N = 163,067)

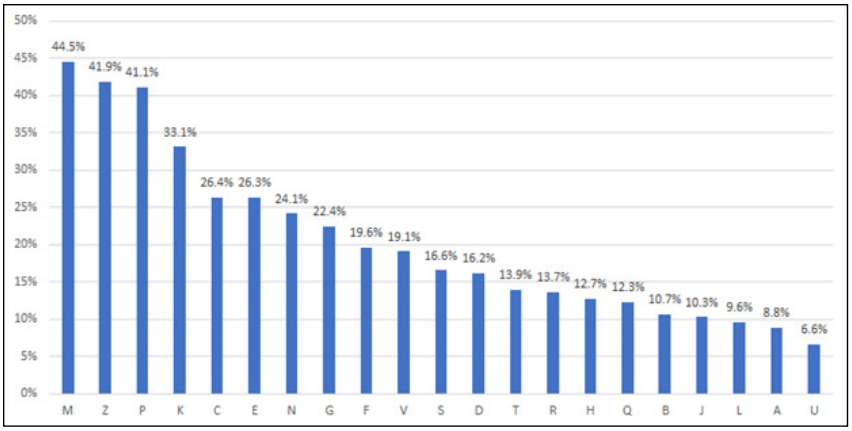

Figure 4. Proportion of Records Containing LCGFT, by Class (N = 684,540)

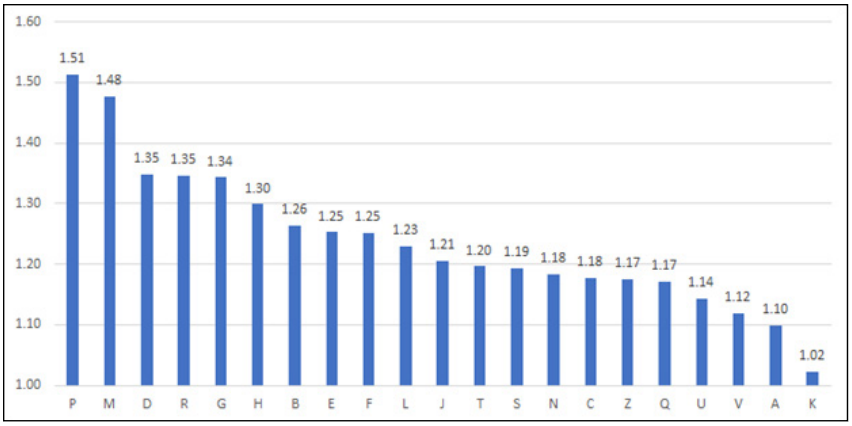

Figure 5. Number of LCGFT Terms Per Record by LCC Class, (N = 220,668)

percent) had LCGFT in the 2007–present file. Figure 6 shows the proportion of records containing LCGFT by class within each group of records, divided into pre-2007 and 2007–present records.

To further illuminate the data provided in figure 6, the authors also examined changes in the rate of LCGFT application in each LC class over time (figure 7). Two LC classes showed significantly increased rates of LCGFT application in more recent, 2007–present records, i.e., A (general works, 41.2 percent) and K (law, 38.2 percent). N (fine arts, 18.1 percent) also showed a moderate increase. As shown in the above analysis of LCGFT application by format, M (music, -13.2 percent) showed a moderate decrease in LCGFT application in the 2007–present group. LCGFT headings in Z (bibliography/library science, -10.6 percent) also decreased in the records representing more recent resources, despite its high representation of LCGFT against the entire base file (41.9 percent, see figure 4).

LCGFT Terms Assigned

Analyzing the individual LCGFT terms contained in the file was another relevant area of inquiry for the present study, as it revealed how the thesaurus had been used in WorldCat records. From the 205,879 records with one or more 655 fields containing subfield $2 lcgft (24.3 percent of the base file), terms in field 655 subfield $a were extracted to perform analysis. There were 284,964 655 fields with subfield $2 lcgft across this subset of records, with an average of 1.38 terms per record with LCGFT. After crossing each individual field 655 subfield $a against the master file of LCGFT (2,357 total terms), the authors found that 10,346 fields did not contain a valid LCGFT term. That is, 274,618 fields contained authentic genre and form terms, resulting in a 3.6 percent error rate in the file. For the 655 fields containing authentic LCGFT, 1,362 unique terms were present in the file, meaning that 57.8 percent of all LCGFT terms had been used in WorldCat records matched to the authors' local library holdings.

Analysis of the invalid LCGFT terms revealed a number of different types of errors. Many were simple typograpical errors (e.g., Stuides (Music), llustrated works), while others were missing qualifiers (e.g., Vespers, Rhapsodies, Thrillers). However, the majority of invalid LCGFT headings found were incorrectly assigned terms. Top offenders included "Electronic government information" (2,656 occurrences), "History" (806 occurrences), "Electronic Journals" (588 occurrences), "Juvenile works" (378 occurrences), and "Picture books for children" (217 occurrences). Table 2 contains every invalid LCGFT term that had more than 100 occurrences in the file.

Despite the fact that LCGFT terms can be easily controlled within the Connexion Client, the authors' data thus make it abundantly clear that invalid terms are still being deposited in field 655. One could easily infer any number of sources through which these invalid LCGFT terms had been introduced into WorldCat records. Some terms could have been simply misapplied by catalogers or there may be a deeper misunderstanding of the vocabulary. Conversely, terms may have been inadvertently added through improper authority control. For example, some authority systems might have flipped LCSH to LCGFT even though an equivalent term does not exist, i.e., 650 _0 $a Piano music.

changed to 655 _7 $a Piano music. $2 lcgft.[18] These headings could have easily ended up in WorldCat master records, particularly in light of ongoing data sync projects. As 3.6 percent is a relatively small portion of the file, one could argue that the problem is not so severe. However, given the ease of correcting many of these headings (for example, Sonatas (Piano) could easily be flipped to Sonatas), it seems regrettable that so many improper terms coded as LCGFT headings exist in WorldCat. Of course, many of these terms likely exist downstream in local library catalogs relying on WorldCat copy records, so the problem collectively has enormous cascading effects on the integrity of bibliographic databases across the wider library community.

Individual terms were further analyzed with a focus on the depth of indexing, that is, the extent to which broader and narrower terms have been assigned in terms of the hierarchies in LCGFT. As seen in figure 8, the overwhelming majority of LCGFT terms found in the authors' base file were coded in the second and third levels of hierarchy (45.7 percent and 42.5 percent, respectively). As expected, more specific, lower-level terms (level 4 in figure 8) were rarely used overall either because they are suited to describing few specialized resources, or because catalogers have applied broader terms for said resources instead. This result seems to reflect the basic guideline in the introduction to LCGFT Manual: "The preference is for broader, rather than narrower, terms. Most literary and artistic resources provide only a broad indication of their genres and forms. Broader terms can therefore expedite cataloging and also serve the users, who do not have to search several very narrow sub-genres or forms to find materials of interest to them."[19] (Note, however, that this guideline does have some conflict with the other guideline found in instruction sheet J 110: "Assign terms that are as specific as the genres and forms exemplified in a resource."[20]) Additionally, broadest, top-level terms (level 1) not surprisingly saw less use (7.9 percent) because these terms are intended more for collocation in each discipline; indeed, in many disciplines, top terms were rarely applied, if at all.[21] Based on the authors' analysis of LCGFT terms used in the base file records, some exceptions included "Sound recordings," "Literature," "Illustrated works," "Music," and "Video recordings," as illustrated by figure 9.

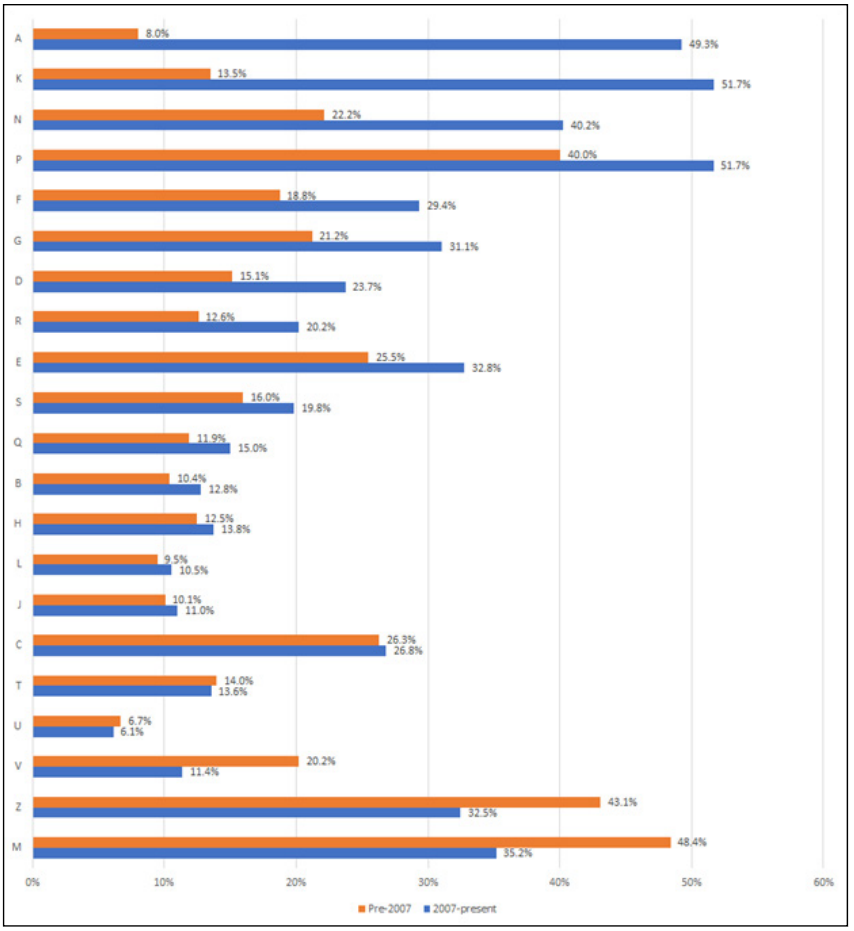

Figure 6. Proportion of Records Containing LCGFT, by Class and Year (N = 683,187)

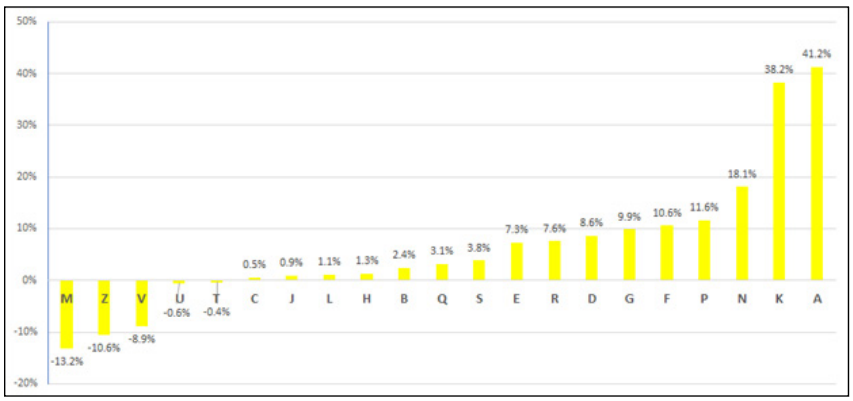

Figure 7. Percent Change in Number of Records Containing LCGFT, pre-2007 to 2007–present (N = 683,187)

In addition to the hierarchical distribution of LCGFT across the base file, records were also analyzed similarly

Table 2. Most Prevalent Invalid LCGFT Terms

Invalid 655 $a with $2 lcgft	No. of Occurrences
Electronic government information	2,656
History	806
Electronic journals	588
Juvenile works	378
Picture books for children	217
Detective and mystery stories	187
Sonatas (Piano)	174
High interest-low vocabulary books	170
Piano music	161
Electronic books	157
Criticism, interpretation, etc	154
Streaming audio	154
Photography, Artistic	147
Children's poetry	122
Compact discs	120
Young adult fiction	109
Children's stories	101

within each LCGFT category. With some exceptions, the most popular level of LCGFT application was level 2, or second-level terms. Thirteen of twenty-one LCGFT categories—commemorative materials, creative nonfiction, derivative works, ephemera, illustrated works, instructional and educational works, literature, motion pictures, religious works, tactile works, television programs, video recordings, and visual works—all followed this pattern. Some other vocabulary categories, by contrast, were applied more at level 3, or third-level LCGFT. These included cartographic materials, discursive works, informational works, law materials, music, and recreational works. The distribution of LCGFT terms in these categories favoring third-level LCGFT can be seen in figure 10. The two remaining categories, art and sound recordings, showed the greatest proportion of terms at level 1, first-level LCGFT. 72.3 percent of art terms were first level; this is not surprising given the relative sparse nature of art compared with other LCGFT categories—indeed, even at the second level only twelve terms are available as of this writing. As for sound recordings, 64.0 percent of terms were first level, which is understandable given the relatively broad applicability of the term. It should be noted, however, that the Music Library Association (MLA) states that "The term 'Sound recordings' is effectively a heading of last resort, i.e., it is a broad term that may be used to capture the sound recording aspect of a resource in cases where a narrower term is not available."[22] Despite these recommendations, the term "Sound recordings" had been available

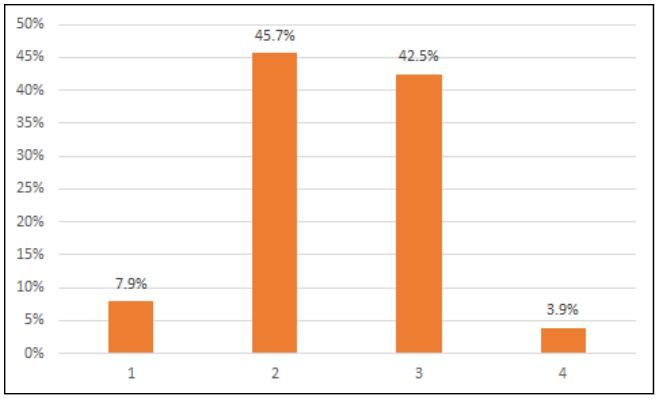

Figure 8. Hierarchical Level of LCGFT Terms as a Proportion of all LCGFT Terms (N = 274,618)

long before many of its narrower terms, which might explain why the top term was applied at greater levels. For example, "Sound recordings" was available as early as 2011, while "Studio recordings" first entered the vocabulary in 2019.

Lastly, two categories showed relatively high usage of fourth-level LCGFT, as evidenced in figure 10. Both cartographic materials (9.7 percent) and music (18.2 percent) exhibited somewhat heightened use of level 4; indeed, of the 21 LCGFT categories, only five showed application of fourth-level LCGFT at rates higher than five percent (cartographic materials, literature, motion pictures, music, and recreational works), with more than half of the twenty-one categories yielding less than one percent. Regarding cartographic materials, the position of both "Bathymetric maps" and "World atlases" within level 4 accounted for the majority of terms contributing to the rate of 9.7 percent in the authors' data. Examining the corresponding records reveals that the vast majority of these materials were for online government documents. As for music, relatively high fourth-level application is somewhat not surprising given the size and nature of the discipline; indeed, at 847 terms, music accounted for over a third of the entire LCGFT vocabulary. What is more, over half of music LCGFT terms (436 terms) in the vocabulary occurred within fourth level, indicating that all those specific terms were clearly regarded as necessary in describing musical resources when LC originally partnered with MLA to develop genre and form terms for music. Thus, it is far more likely for level 4 terms to be applied in this discipline than religious materials, for example, in which only 21.6 percent of terms occur at the lowest level of hierarchy.

Conclusion

The purpose of this study was to provide exploratory analysis of LCGFT within a large set of MARC bibliographic data. The authors retrieved their institutional holdings

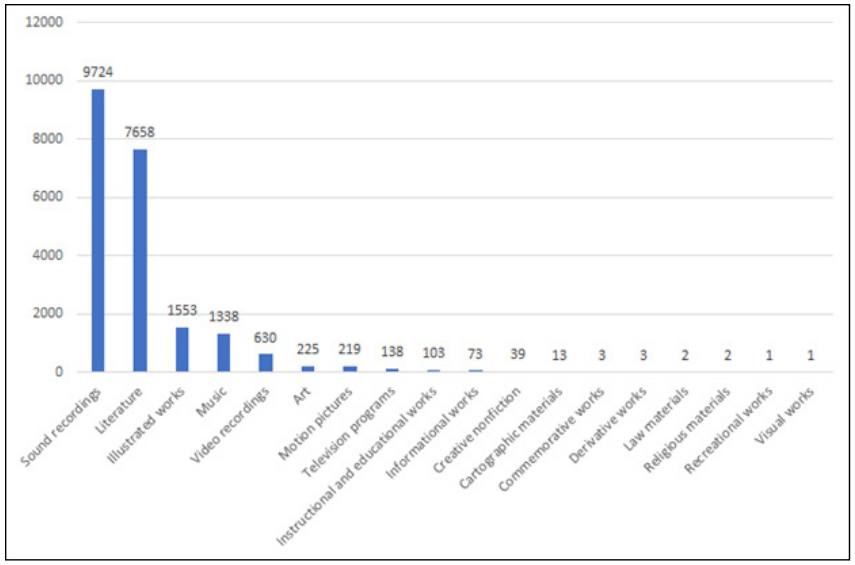

Figure 9. Number of Top Level Terms Used

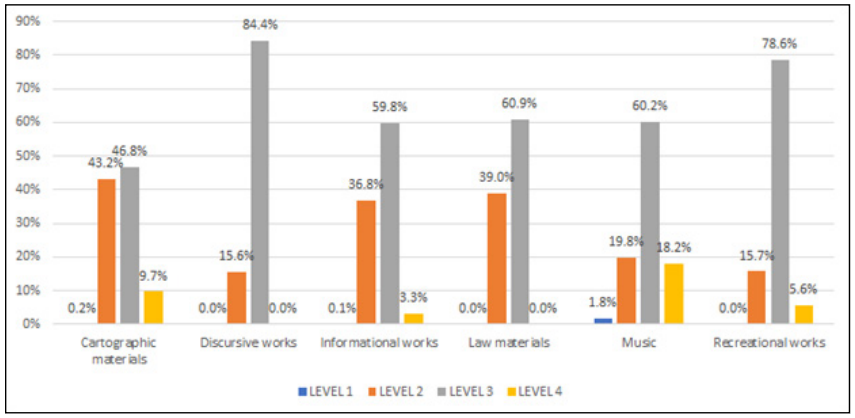

Figure 10. LCGFT Categories Favoring Level 3 Application

in WorldCat, using more than 800,000 master WorldCat records as the basis for analysis. From this base file, various data, such as date, format, call numbers, and LCGFT, were extracted and analyzed to explore a series of research questions related to the current status of LCGFT usage in WorldCat. With regard to changes in LCGFT application over time, there was an increase of seven percent between pre-2007 records and recent records from 2007–present (22.5 percent to 29.5 percent). Additionally, the average number of LCGFT terms increased in records containing them, from 1.34 to 1.50. When analyzing the data by format (e.g., type of record), most formats saw an increase in LCGFT application over time, with the exception of musical sound recordings. These findings are also supported by further analysis based on LCC; indeed, while many classes showed an increase in application between the pre-2007 and 2007–present sets, unexpected decreases were found in both M (music, -13.2 percent) and Z (bibliography/library science, -10.6 percent). The reason for such decreases could be that pre-2007 music materials may have received higher levels of retrospective application of LCGFT, or they could have been originally cataloged after music terms were added to LCGFT in 2015. Alternatively, the decrease might be explained by increased levels of batch loading of newer records for streaming sound recordings by external providers. A separate inquiry into these results and prospects for retrospective application would be warranted in view of the varied LCGFT application between formats and LC classes. Furthermore, this exploratory study used the year 2007—when LC first released the LCGFT thesaurus for moving image materials—as the point of demarcation to shed some light on changes in LCGFT application over time. Because LCGFT terms have been added in different disciplines over multiple years, it will be worthwhile to pursue further research on how LCGFT usage changed respectively when the LCGFT project was completed for a given discipline.

When examining the entire LCGFT vocabulary in terms of hierarchy, the authors found that second and third-level headings were assigned most frequently (45.7 and 42.5 percent, respectively). This was also evident for the overwhelming majority of individual LCGFT disciplines examined, such as motion pictures (favoring second-level) and music (favoring third-level). Perhaps the preponderance of second and third-level LCGFT headings used suggests that the hierarchical design of the vocabulary is working; it is reasonable to assume that they are specific enough, compared with the broadest, top-level terms, to describe the genres and forms exemplified in resources being cataloged, but not too narrow to impede efficient cataloging or confound the users as they try to find materials of interest to them. Further, the most specific, fourth-level LCGFT (which included fourth-level terms and below in the current paper) saw the least usage as these terms would naturally only be used for more specialized or unique resources; for example, cartographic materials and music, which had higher fourth-level usage than other disciplines. These results suggest that future efforts to add new terms to the vocabulary should aim to strike a balance

between specific and broad terms. Additionally, it should be noted that the application of first-level terms within certain LCGFT disciplines may warrant further analysis. For example, headings such as "art" and "sound recordings," particularly if the only genre form term recorded in the record, may not necessarily provide users with altogether helpful information, and further analysis might yield new insights that will be essential for any individual constituencies that wish to develop best LCGFT practices guidelines in these disciplines. Lastly, the number of erroneous terms in fields 655 subfield $a with $2 lcgft (10,346 total fields in the base file) points to some much needed data cleanup in WorldCat, as well as potential training and documentation for applying LCGFT terms correctly.

While the data reported in this study point to a moderate increase in LCGFT use over time, the amount of LCGFT within the base file suggests that the vocabulary has not been applied to the fullest extent possible in WorldCat. The results of the present study indicate that it is highly important that newly cataloged materials receive LCGFT application within records from the outset, so as to ensure that a more sizable portion of new bibliographic records include appropriate genre and form terms and lessen the need for retrospective application over time. As such, training needs to be increased in both libraries and library schools to facilitate broader LCGFT application. Increased communication with vendors may also be warranted, as LCGFT may be lacking (or incorrect) in vendor-supplied metadata. While these actions may further improve end-user retrieval based on genre and form, catalogers and other technical services librarians may need to begin to investigate more sophisticated methods in applying the vocabulary retrospectively to appropriate legacy records as well. Indeed, as the data contained in the base file show somewhat uneven application of LCGFT, and with nearly half a billion records in WorldCat as of the 2020 OCLC report, it remains a certainty that much of LCGFT's full potentials for genre/form access and retrieval will remain untapped until innovative solutions are introduced to increase vocabulary usage in bibliographic databases.[23]

References

1. Library of Congress, "Introduction to Library of Congress Genre/Form Terms for Library and Archival Materials," 2019, https://www.loc.gov/aba/publications/FreeLCGFT/2019%20LCGFT%20intro.pdf.
2. Library of Congress, "Report on the Moving Image Genre/Form Project," 2008, https://www.loc.gov/catdir/cpso/movimgenre.pdf.
3. Charles A. Cutter, *Rules for a Printed Dictionary Catalog* (Washington, DC: US Government Printing Office, 1876); Elaine Svenonius, *The Intellectual Foundation of Information Organization* (Cambridge: MIT Press, 2000).
4. Pat Riva, Patrick Le Bœuf, and Maja Žumer, *IFLA Library Reference Model* (International Federation of Library Associations and Institutions, 2017).
5. American Library Association. Subcommittee on the Revision of the Guidelines on Subject Access to Individual Works of Fiction, *Guidelines on Subject Access to Individual Works of Fiction, Drama, etc.* (Chicago: American Library Association, 2000); David P. Miller, "Out From Under: Form/Genre Access in LCSH." *Cataloging & Classification Quarterly* 29, no. 1-2 (2000): 169–88; Edward T. O'Neill et al, "Form Subdivisions: Their Identification and Use in LCSH." *Library Resources & Technical Services* 45, no. 4 (2001): 187–97.
6. Library of Congress, "Report on the Moving Image Genre/Form Project"; Janis L. Young and Yael Mandelstam, "It Takes a Village: Developing Library of Congress Genre/Form Terms," *Cataloging & Classification Quarterly* 51, no. 1–3 (2013): 6–24.
7. Library of Congress, "Introduction to Library of Congress Genre/Form Terms for Library and Archival Materials."
8. Hur-Li Lee and Lei Zhang, "Tracing the Conceptions and Treatment of Genre in Anglo-American Cataloging," *Cataloging & Classification Quarterly* 51, no. 8 (2013): 891–912.
9. Library of Congress, "Introduction to Library of Congress Genre/Form Terms for Library and Archival Materials."
10. Martha Yee, "Two Genre and Form Lists for Moving Image and Broadcast Materials: A Comparison," *Cataloging & Classification Quarterly* 31, no. 3–4 (2001): 237–95; Faye Leibowitz, "Form and Genre Headings in Serials Cataloging," *Cataloging & Classification Quarterly* 20, no. 3 (1995): 19–41; Spillane Wilson, "The Relationship Between Subject Headings for Works of Fiction and Circulation in an Academic Library," *Library Collections, Acquisitions, & Technical Services* 24, no. 4 (2000): 459–65; Carrie Newsom, Jimmie Lundgren, and Nancy Mitchell Poehlmann, "Genre Terms for Chemistry and Engineering: Not Just for Literature Anymore," *Cataloging & Classification Quarterly* 46, no. 4 (2008): 412–24.
11. Young and Mandelstam, "It Takes a Village."
12. Beth Iseminger et al., "Faceted Vocabularies for Music: A New Era in Resource Discovery," *Notes*, 73, no. 3 (2017): 409–31; Mark McKnight, "Are We There Yet? Toward a Workable Controlled Vocabulary for Music," *Fontes Artis Musicae* 59, no. 3 (2012): 286–92.
13. Casey A. Mullin, "An Amicable Divorce: Programmatic Derivation of Faceted Data from Library of Congress Subject Headings for Music," *Cataloging & Classification*

Quarterly 56, no. 7 (2018): 607–27.
14. Patricia M. Dragon, "Form and Genre Access to Academic Library Digital Collections," *Journal of Library Metadata*, 20, no. 1 (2020): 29–49.
15. Colin Bitter and Yuji Tosaka, "Genre/Form Access in Library Catalogs: A Survey on the Current State of LCGFT Usage," *Library Resources & Technical Services* 64, no. 2 (2020): 44–61.
16. Library of Congress, "Assigning Genre/Form Terms," 2016, https://www.loc.gov/aba/publications/FreeLCGFT/J110.pdf.
17. OCLC, "FAST Frequently Asked Questions," 2019, https://www.oclc.org/content/dam/oclc/fast/FAST-FAQ-Nov2019.pdf.
18. Bitter and Tosaka, "Genre/Form Access in Library Catalogs."
19. Library of Congress, "Introduction to Library of Congress Genre/Form Terms for Library and Archival Materials."
20. Library of Congress, "Assigning Genre/Form Terms."
21. Library of Congress, "Frequently Asked Questions about Library of Congress Genre/Form Terms for Library and Archival Materials (LCGFT)," 2011, https://www.loc.gov/catdir/cpso/genre_form_faq.pdf.
22. Music Library Association. Cataloging and Metadata Committee Subcommittee, "Best Practices for Using LCGFT for Music Resources," 2019, http://cmc.blog.musiclibraryassoc.org/wp-content/uploads/sites/5/2019/07/BestPracticesforUsingLCGFT_Music_1.2_20190708_revURLs.pdf.
23. OCLC, "OCLC Annual Report 2019–2020," 2020, https://www.oclc.org/en/annual-report/2020/home.html.