Form Subdivisions

Their Identification and Use in LCSH

Edward T. O'Neill, Lois Mai Chan, Eric Childress, Rebecca Dean, Lynn M. El-Hoshy, and Diane Vizine-Goetz

Form subdivisions have always been an important part of the Library of Congress Subject Headings. However, when the MARC format was developed, no separate subfield code to identify form subdivisions was defined. Form and topical subdivisions were both included within a general subdivision category. In 1995, the USMARC Advisory Group approved a proposal defining subfield $v for form subdivisions, and in 1999 the Library of Congress (LC) began identifying form subdivisions with the new code.

However, there are millions of older bibliographic records lacking the explicit form subdivision coding. Identifying form subdivisions retrospectively is not a simple task. An algorithmic method was developed to identify form subdivisions coded as general subdivisions. The algorithm was used to identify 2,563 unique form subdivisions or combinations of form subdivisions in OCLC's WorldCat. The algorithm proved to be highly accurate with an error rate estimated to be less than 0.1%. The observed usage of the form subdivisions was highly skewed with the 100 most used form subdivisions or combinations of subdivisions accounting for 90% of the assignments.

Edward T. O'Neill (oneill@oclc.org) is Research Scientist, Office of Research, Online Computer Library Center (OCLC), Dublin, Ohio.

Lois Mai Chan (loischan@pop.uky.edu) is Professor, School of Library and Information Science, University of Kentucky, Lexington.

Eric Childress (eric_childress@oclc.org) is Consulting Product Support Specialist, Metadata Services Division, OCLC.

Rebecca Dean (rebecca_dean@oclc.org) is Manager, Metadata Analysis and Investigation Section, OCLC.

Lynn M. El-Hoshy (lelh@loc.gov) is Senior Cataloging Policy Specialist, Cataloging Policy and Support Office, Library of Congress, Washington, D.C.

Diane Vizine-Goetz (vizine@oclc.org) is Research Scientist, Office of Research, OCLC.

Manuscript received March 8, 2001; accepted for publication June 25, 2001.

Recent efforts to distinguish between topical and form data are moving Library of Congress Subject Headings (LCSH) closer to a truly faceted subject vocabulary. While form data in LCSH are represented in both form headings and form subdivisions, under the current LC application rules, form data appear in most cases as subdivisions under topical or name headings.

In implementing the $v subfield code for form subdivision in the MARC 21 (formerly USMARC) format, a number of issues have come to the fore:

- distinction between form and topical subdivisions
- combinations of two or more form subdivisions in the same heading string

In this article, a method is developed to algorithmically identify form subdivisions lacking explicit form subfield coding.

Explicit Coding for Form Subdivisions

Form subdivisions have been a part of LCSH since its inception. Beginning in 1906, the Library of Congress issued auxiliary lists of subdivisions that included a section of "General form divisions under subjects." Guidelines on the use of subdivisions, such as those published in the introduction to the eighth edition of

Library of Congress Subject Headings (Library of Congress 1975), instructed catalogers to use individual subdivisions either "as a topical subdivision," "as a form subdivision," or "as a form or topical subdivision" under specified types of headings for particular types of materials. Yet when the MARC format for encoding and communicating bibliographic data was developed in the late 1960s, a separate subfield code to identify form subdivisions in subject heading strings was not defined. Form subdivisions were included along with topical subdivisions in a general subdivision category to be coded as $x.

In 1991, a conference was convened at Airlie, Va., to consider the role of subdivisions in LCSH. One of the conference's six recommendations was: "The question of whether subdivisions should be coded specifically to improve online displays for end users should be considered . . . In particular, the Library of Congress should investigate implementing a separate subfield code for form subdivisions" (O'Hara Conway 1992). In response, the Library of Congress requested that the ALA Association for Library Collections and Technical Services (ALCTS) Cataloging and Classification Section (CCS) Subject Analysis Committee (SAC) investigate form subdivision coding. Hemmasi, Miller, and Lasater (1999) report on the issues that SAC identified and studied, including "retrospective conversion, varying cataloging practices and user needs across disciplines, no distinct list of form headings, cataloger training, and the redundancy of content in USMARC record elements" (unnumbered). In 1993, SAC recommended that a separate subfield code for form subdivisions be implemented. Subsequently, two discussion papers defining a new subfield code and posing questions on retrospective conversion, the use of a form subdivision subfield by online systems, authority control, implementation options, and general user opinions were considered by the USMARC Advisory Group before it approved a proposal to define subfield $v for form subdivisions in 1995. The proponents argued that a separate subfield code would make it possible to retrieve form data more predictably, improve online displays for users, and separate LCSH elements into their facets of topic, place, chronology, and form.

Guidelines for Assignment

In applying form subdivisions, the question is: Where does the cataloger look for guidance? There are several sources and methods of information:

- *Subject Cataloging Manual: Subject Headings (SCM)* (Library of Congress 1996)
- *Free-Floating Subdivisions: An Alphabetical Index (FFS)* (Library of Congress 2000)
- Patterns discerned in assigned heading strings in LC MARC records
- Subdivision authority records
- The test of what the work "is" versus what the work is "about" to determine the appropriate category of subdivision
- The "reading backwards" or "from right to left" test to determine the proper order of subdivisions within the string

When a question arises, the first place for a cataloger to look for an answer is *Subject Cataloging Manual: Subject Headings.* The manual gives numerous instructions and examples on the application of many of the subdivisions, although they are scattered throughout the publication. The publication *Free-Floating Subdivisions: An Alphabetical Index* provides a quick reference for precombined subdivisions. Nevertheless, there are still situations not fully covered; many multiple free-floating subdivisions that appear in LC MARC authority records are not shown in *SCM* or *FFS.* For these, one must rely on other means. One possible approach is to examine patterns in assigned heading strings in LC MARC bibliographic records, which can serve as examples but hardly provide definite answers. The test that a form subdivision "represents what the book is, rather than what it is about" (Haykin 1951) may also be used to help in the distinction between form and topic. Finally, another test that has been suggested is to read the heading string backwards, i.e., from right to left, to see if the string fits the context of the item being cataloged. For example:

Art—Bibliography—Periodicals
(a serially issued art bibliography)
Art—Periodicals—Bibliography
(a bibliography of journals on art)

Distinction between Form and Topical Subdivisions

Virtually all efforts to revise or improve LCSH, including the Airlie Conference (O'Hara Conway 1992), ALCTS/SAC/Subcommittee on Metadata and Subject Analysis (*Subject data in the metadata record 1999*), and OCLC's FAST (Faceted Application of Subject Terminology) project (Chan et al. 2001), consider form subdivisions as a distinct type and treat form subdivisions differently from general ($x) subdivisions. All of these efforts assume that form subdivisions can be identified. However, until recently, the Library of Congress coded form subdivisions the same as general subdivisions ($x). Only in 1999 did the Library of Congress begin explicitly identifying forms with the $v subfield code.

In coding form subdivisions, the first issue to be resolved is how to determine whether a particular subdivision in a subject string represents a topic or form. Although many terms clearly belong to one or the other category, many others are ambiguous. While subdivisions such as **—Education** or **—Quality control** can only be considered topical, others are not so obvious. For example, subdivisions such as **—Texts** and **—Translations into French [German, etc.]** may be used as either a topical or form subdivision, depending on the context. Even subdivisions such as **—Periodicals** are sometimes used as topical subdivisions. For example, in the heading:

Academic achievement $xPeriodicals
$vIndexes
(an index to a journal on academic achievement)

the subdivision **—Periodicals** is topical; but in the heading:

Universities and colleges $xFinance
$vPeriodicals
(a journal on higher education finance)

it is a form since it is assigned to represent a publication issued in serial form.

Currently, the subfield code for each free-floating subdivision is shown in *SCM* and *FFS*. The Library of Congress is in the process of creating authority records for free-floating subdivisions with specific information regarding subfield codes. When completed, the specific instruction will contribute greatly to consistency in application.

Combinations of Two or More Form Subdivisions

The use of two or more subdivisions involving form data within the same heading raises at least three problems:

- When can a form subdivision be further subdivided by another form, geographic, or topical subdivision?
- In what order should the subdivisions appear?
- How does one code each subdivision, that is, how does one choose between $v, $x, and $z?

To answer the first question, *SCM* and *FFS* list many precombined multiple form subdivisions as an aid to catalogers. Examples include:

—Biography—Dictionaries (v-v)
—Biography—Sermons (v-v)
—Maps—Facsimiles (v-v)

In many cases, a form subdivision may be further subdivided by a topical subdivision.

—Concordances, English—Authorized, [Living Bible, Revised Standard, etc.] (v-x)
—Dictionaries—Polyglot (v-x)

In limited cases, a form subdivision may also be further subdivided by a geographic subdivision as in:

School buildings $vSpecifications $zIowa

However, it is not practical to list all possible combinations in *SCM* or *FFS*, and many such combinations not enumerated in these publications have been assigned to bibliographic records. For example:

—Biography—Sources (v-v)
—Catalogs—Periodicals (v-v)
—Indexes—Periodicals (v-v)
—Observations—Periodicals (v-v)
—Statistics—Periodicals (v-v)

Again, in each case, the cataloger is called upon to exercise judgment.

There are situations where LC instructions specifically prohibit certain combinations of form subdivisions. For example, **—Abstracts** should not be used after **—Congresses** (cf. *SCM* H1460). H1927 in *SCM* contains a list of form subdivisions that cannot be further subdivided by the subdivision **—Periodicals**. It is important that the cataloger be aware of the prohibition when using these subdivisions.

The second question relating to the use of two or more form subdivisions is: In what order should the individual form subdivisions appear within the string? The first place to seek guidance is in *SCM* or *FFS*. The lists of free-floating subdivisions enumerate many precombined subdivisions, for example, **—Bibliography—Catalogs**.

For combinations not listed in *SCM* or *FFS*, other methods must be employed. In most subject headings, the form subdivision appears as the last element, following the general pattern of subdivision order, **Topic—Topic—Place—Time—Form.** However, there are exceptions such as: **—Conversation and phrase books—Polyglot** (v-x).

When the desired combination is not enumerated, the cataloger must exercise judgment based on the context. One suggestion made earlier is to "read backwards," or from right to left, to see if the string fits the context of the document.

—Periodicals—Indexes
(for an index to periodicals)
—Indexes—Periodicals
(for a serially issued index)

For guidance on the third question, how to code subdivisions in each case, the Library of Congress has provided a most valuable service in indicating subfield coding for each free-floating subdivision in recent updates of *SCM* and *FFS*. Newly created authority records of subdivisions also indicate the appropriate coding information. Nevertheless, lists in these publications are not exhaustive. For example, while **—Biography—Anecdotes** (v-v) and **—Biography —Dictionaries** (v-v) are enumerated, the combination **—Biography—Bibliography** is not, even though it has been used in bibliographic records. The difficulty lies in the fact that one cannot assume that in all cases, when two or more form subdivisions appear under the same heading, the coding is always v-v. When an apparent form subdivision is followed by another form subdivision or another topical subdivision, the subfield code can change. For example,

—Bibliography (v)
—Bibliography—Exhibitions (v-v)
—Bibliography—Methodology (x-x)

—Hymns (v)
—Hymns—History and criticism (x-x)
—Hymns—Texts (v-v)

—Maps—Early works to 1800 (v-v)
—Maps—Facsimiles (v-v)
—Maps—Symbols (x-x)

The specific guidance given in *SCM* and *FFS* is of enormous help, but what if one combines **—Abstracts** with **—Periodicals,** a combination not listed in *FFS*?

The advice often given for distinguishing between form and topical subdivisions is to ask whether the subdivision in question represents what the document "is" or what it "is about." This test can usually resolve the question of content versus form.

In certain cases, a trailing form subdivision may affect the coding of the preceding form subdivision, for example:

—Maps (v)
(Map(s) of . . .)
—Maps—Bibliography (x-v)
(list(s) of maps of . . .)
—Periodicals (v)
(serial(s) or periodical(s) on . . .)
—Periodicals—Abbreviations of titles (x-v)
(abbreviations of titles of serials or periodicals on . . .)
—Periodicals—Bibliography (x-v)
(list(s) of serials or periodicals on . . .)
—Periodicals—Bibliography—Catalogs (x-v-v)
(list(s) of serials or periodicals held by one organization or library)
—Periodicals—Bibliography—Union lists (x-v-v)
(catalog(s) of serials or periodicals on those subjects held by two or more libraries)

In some cases, a subject heading may include two or more form subdivisions, which further compound the problem in order and in coding, for example:

Alcoholism $xPrevention $xPeriodicals $vAbstracts $vDatabases
Jews $zPoland $zRadom (Voivodeship) $xHistory $xSources $vBibliography $vCatalogs

The Subdivision —History

The application of the subdivision **—History** is particularly problematic. Currently, it is coded as a topical ($x) subdivision in *SCM* and *FFS*. In effect, when it appears in a subject heading string, it usually represents what the document "is" rather than what it is "about." For example, the heading **Education—History** is assigned to a work that "is" a history of education, not a work "about" the history of education. The problem is compounded when the subdivision **—History** is combined with another form subdivision. For example:

Science $xHistory $vPeriodicals
(a serial or periodical on scientific history)
Science $xPeriodicals $xHistory
(a history of scientific serials or periodicals)

Here, the method of judging by what it "is" versus what it is "about" fails to work.

A similar subdivision is **—History and criticism,** which is also coded as a general ($x) subdivision. The heading **Literature—History and criticism** is normally assigned to a history of literature rather than a work about literary history. The use of **—History** and **—History and criticism** also results in combinations such as:

—Biography (v)
(biography of . . .)
—Biography—History and criticism (x-x)
(a history or criticism of biography of . . .)
—Music (v)
(music of an ethnic group)
—Music—History and criticism (x-x)
(a history or criticism of the music of an ethnic group)

Algorithmic Identification of Form Subdivisions

Identifying and coding form subdivisions is not a simple task. OCLC's WorldCat contains more than eight million unique Library of Congress topical and geographic subject headings—less than 4% contain explicitly coded form subdivisions. The other headings either do not contain any forms or have forms coded as general subdivisions. Identifying forms is difficult due to the complexity of forms structure and the fact that many subdivisions can be either topical (general) or form depending on the context of the heading.

The sheer number of headings demands that an automated procedure be developed to identify and recode form subdivisions. For this purpose, research staff at OCLC developed an algorithmic method based on a table-driven procedure. After extended review and analysis, the approach adopted in this project for identification is first to deal with the special forms, that is, form subdivisions with special or unique application rules, and then to use a table-driven procedure to identify the remaining forms.

Step One: Identifying Special Forms

The following subdivisions are governed by special rules when they are used as the last subdivision in a heading string: **—Periodicals, —Juvenile, —Juvenile literature, —Juvenile films, —Juvenile sound recordings, —Databases, —Early works to 1800,** and **—Facsimiles.** Any of these forms can be removed from the heading and the remainder of the heading can be treated as if these forms were never part of the heading. For the purpose of identifying form subdivisions, the heading:

Land value taxation $zIreland $xTables $xEarly works to 1800

can be reduced to:

Land value taxation $zIreland $xTables.

After removing **—Early works to 1800,** any remaining forms in the heading can be identified using the table-driven procedure.

There are some additional restrictions on removing these forms. The restrictions on what can precede **—Periodicals** are specified in *SCM* (H1927). To prevent invalid combinations of form subdivisions from being identified, if any of the subdivisions specified in H1927 or the subdivisions **—Exhibitions** or **—Newspapers** immediately precedes **—Periodicals**, the subdivision is not removed from the heading. The "Juvenile" forms are restricted to headings not otherwise identified as juvenile. These are not removed when they begin with the word "Juvenile" or "Children's."

In headings involving this group of forms, the last subdivision in the string would be recoded as $v, and the rest of the heading would be analyzed with the last subdivision removed from the heading. For example, the heading **Cities and towns $zUnited States $xMaps $xDatabases** (before recoding) would be treated as **Cities and towns $zUnited States $xMaps** in the remainder of the analysis. The following are some examples where the last (underlined) general subdivision would be removed:

Medical care $zArab countries $xEarly works to 1800
Photography $xCatalogs $xPeriodicals
Fuelwood consumption $zPrince Edward Island $xStatistics $xPeriodicals
Lesbian teenagers $zUnited States $xCase studies $xJuvenile literature

However, the following would not be removed since they are exceptions to the general rule:

African Americans $zNew York (State) $xGenealogy $xPeriodicals
Christmas $xJuvenile fiction $xJuvenile sound recordings
Art, German $zGermany (East) $xExhibitions $xPeriodicals

Note that the remaining general ($x) subdivisions are not necessarily correct—only that they are not valid form subdivisions. Regardless of whether or not any forms are removed, the headings continue to be analyzed.

The forms **—Bibliography, —Congresses,** and **—Indexes** are also given special treatment. Any heading that ends with either of these subdivisions is recoded with **—Bibliography, —Congresses,** or **—Indexes** as $v, but none of the other subdivisions will be considered to be forms. The following headings are shown with the revised subfield codes (assuming the $v were originally coded as $x):

Scottish poetry $y20th century $vBibliography
Nahuas $xPeriodicals $vIndexes
Urbanization $zNigeria $xStatistics $vCongresses

There are four form subdivisions that can be geographically subdivided: **—Catalogs and collections, —Job descriptions, —Specifications,** and **—Registers of dead.** The following are examples of recoded form subdivisions followed by geographic subdivisions:

Medicinal plants $vCatalogs and collections $zThailand $zSala Ya
Sewage disposal plants $vSpecifications $zTexas $zEl Paso

The subdivision **—Readers,** which is used for reading texts, is another special case that can be followed by any topic. **—Readers** should be recoded as a $v but the following topic is retained as $x. Some examples of recoding **—Readers** are:

Spanish language $vReaders $xCivilization
German language $vReaders $xScience
Russian language $vReaders $xSoviet Union

Note that even under subdivision **—Readers** when the *topic* is a geographic name such as the **Soviet Union,** the subdivision containing the geographic name is coded as a topical ($x) subdivision rather than a geographic ($z) subdivision, because in this case the place name represents a topic rather than a location.

Step Two: Identifying Forms

A table of form subdivisions was created by supplementing the list of forms identified in *Free-Floating Subdivisions* with other forms identified through various sources. All headings not subject to the special treatment described above are then checked to determine if they have terminating subdivision(s) matching those in the augmented list. This expanded list contains form patterns and their preferred subdivision coding. Included in the list are 639 entries containing from one to three subdivisions.

Since there can be no more than three form subdivisions after removing the special forms, the list is searched in three steps. The first search is for the last three (if they exist) general subdivisions. If it matches an entry in the list, the heading is recoded using the preferred coding. If no match is found, the last two general subdivisions are searched. If still no match is found, a final search is made for the last subdivision. For example, in the heading:

American literature $xAfrican American authors $xHistory and criticism $xTheory, etc.

the last three general subdivisions, **—African American authors—History and criticism—Theory, etc.,** would be checked against the table. When no match was found, the last two subdivisions, **—History and criticism—Theory, etc.,** would be checked. If, again, no match was found, the final subdivision, **—Theory, etc.,** would be checked. If all matches failed, the conclusion would be that all of the subdivisions were topical and that the original coding was assumed to be correct.

The following form subdivisions from the list serve as patterns for other national, ethnic, or language terms:

—Concordances, *English*
—Films for *English* speakers
—Harmonies, *English*
—Interlinear translations, *English*
—Liturgical lessons, *English*
—Parallel versions, *English*
—Paraphrases, *English*
—Personal narratives, *English*
—Sound recordings for *English* speakers
—Textbooks for *English* speakers
—Translations into *English*
—Video recordings for *English* speakers
—Catechisms—*English*
—Conversation and phrase books—*English*
—Dictionaries, Juvenile—*English*
—Dictionaries—*English*
—Prayer-books and devotions—*English*
—Textbooks for foreign speakers—*English*
—Bio-bibliography—Dictionaries—*English*
—Biography—Dictionaries—*English*

In these combinations, *English* serves as the pattern and can be replaced by any national, ethnic, or language terms.

Evaluation of the Identification Algorithm

For the algorithm to be usable, it had to be highly reliable. With complex processes of this type, developing an error-free process is an unrealistic goal. Neither manual recoding by skilled professionals nor machine algorithms can be expected to produce perfect results. Even highly skilled professionals make mistakes—typically errors of oversight. Such an error is illustrated in the following heading:

English language $vDictionaries $vChinese

The form subdivision **—Dictionaries** can be subdivided by language, but the language subdivision **—Chinese** should be coded as a general subdivision ($x). This type of error is relatively common in spite of the fact that most professionals understand that language should be coded as a general subdivision. It is not that the cataloger didn't know how to code it but rather that it was overlooked. The only way errors of this type can be eliminated, or at least dramatically reduced, is to have at least two people recode each heading and to recheck each heading where the cod-

ing differs. The use of multiple coders, however, is very expensive, and would be difficult to justify in a production environment.

By contrast, algorithms produce very consistent results: they do not overlook anything. Algorithms, however, have very limited ability to understand the context. For example, the heading

Executives $vQuotations

is valid since **—Quotations** is authorized in the pattern heading for *Classes of Persons*. Based on general knowledge of language, most catalogers understand that executives are a class of people and, therefore, the pattern heading is appropriate. Algorithmic procedures have a much more difficult time with this type of contextual information. Unless the algorithm has been explicitly *told* or has previously *learned* that executives are a class of people, it has no way to validate this heading. As a result, the type of errors resulting from the algorithmic coding tend to be different from those made by people.

While recognizing that comparing manual and algorithmic error rates is a little like comparing apples to oranges, it nevertheless seems to be the best approach to evaluating the algorithm. It was assumed that algorithmic error rates that were as good or better than those observed in manual assignment would be acceptable. A methodology to estimate the algorithm's accuracy was required. Fortunately, since the Library of Congress currently is explicitly coding form subdivisions, there are a large number of records in WorldCat with the form subdivisions explicitly identified with the $v subfield code. For testing, all topical (650) and geographic (651) subject headings with explicit form coding were extracted to create a test file. Presumably, all of these records follow current coding practice.

To test the algorithm, all $v subfield codes in the heading were replaced with $x codes and then the heading was algorithmically recoded to explicitly identify the form subdivisions. For example, the heading **Agriculture $vIndexes** was changed to **Agriculture $xIndexes** in the test file. That heading then was algorithmically recoded as **Agriculture $vIndexes.** The resulting heading was then compared to the original to identify any headings for which the algorithmic form coding was different from the original heading. Presumably, all of the recoded headings that matched the original were correct. All headings that did not match the original were manually reviewed. As an example, headings pairs from that list are shown below:

Artists $zGermany $vInterviews $vBibliography

Artists $zGermany $xInterviews $vBibliography

France $xPolitics and government $y1789- $vHistoriography
France $xPolitics and government $y1789- $xHistoriography

The first heading of each pair is the original heading as it appeared in the MARC record. The second heading is the same heading after being algorithmically recoded. Each of these heading pairs was reviewed by at least two of the authors to determine the correct coding. When the initial reviewers did not agree on the coding, the headings were reviewed by all of the authors to ensure that the results were as accurate as possible.

During the review, it was found that some headings contained errors that could not be corrected by changing the subfield coding only. For example, in the heading

Onondaga Indians $vProtraits

the subdivision **—Portraits** is misspelled. As a result, the heading can only be corrected by changing the text of the subdivision. All headings with errors that could not be fully corrected by changing the subfield coding were removed from the test file.

The resulting test file contained 20,970 headings: 17,208 topical and 3,762 geographic. Of these headings, 662 contained manual coding errors resulting in a manual error rate of 3.15%. The coding of **—Dictionaries—English** as v-v rather than v-x was typical of the manual miscoding observed. The algorithm miscoded 15 headings, resulting in an algorithmic error rate of 0.07%, significantly better than in the case of manually coded records.

Caution is required in interpreting these results. First, the subject headings used in the test were assigned or recoded in early 1999 and, therefore, include the first attempts to explicitly code form subdivisions. As the catalogers gain experience in assigning subfield code $v, the accuracy of the coding can be expected to improve significantly. Second, the test headings were created in a *production* environment at the Library of Congress. In such an environment, accuracy must be balanced with productivity.

However, even recognizing that the current manual error rate is likely to be significantly less than the 3.15% observed, it appears to be impossible to achieve in a production environment a manual error rate as low as the algorithmic rate. To obtain manual error rates less than 1% would probably require at least two people independently assigning the subfield codes. Certainly, when compared to the manual error rate, an algorithmic error rate of less than 0.1% appears to be very acceptable.

Usage Patterns

A large number of valid forms were identified but many were rarely assigned; 2,412 unique form subdivisions or combinations of form subdivisions were identified in topical and geographic headings from WorldCat. The geographically subdivisible forms **—Catalogs and collections, —Job descriptions, —Registers of dead,** and **—Specifications** were considered without their geographic subdivisions. For example, the combination **—Catalogs and collections —Japan** was treated simply as **—Catalogs and collections.** Collectively, these four forms would have resulted in 1,674 additional unique form subdivisions if they had been included with their geographic subdivisions. General ($x) subdivisions were included so that the forms identified consisted of a combination of one to four $v and/or $x subdivisions. The 100 most frequently assigned form subdivisions are shown in table 1. The complete table is available on OCLC's Web site (http://wcp.oclc.org/fast).

As shown in table 1, the most frequently used form is **—Congresses,** which has been assigned a total of 1,109,724 times in WorldCat, including the 317,800 times it had been assigned by the Library of Congress. The "Relative use by the Library of Congress" column indicates the relative frequency that the form was assigned by the Library of Congress compared to its use in contributed records. For example, the relative use of 50% for **—Periodicals** means that it is assigned by the Library of Congress about half as often as it is in contributed records. By contrast, **—Biography** is assigned more than twice as frequently by the Library of Congress. The wide variation in relative use by the Library of Congress reflects both a difference in cataloging practice and in the types of materials cataloged.

The forms identified contain 1 to 4 subdivisions, excluding any geographic subdivisions. The majority (63%) of the forms contain 2 subdivisions each, and only a quarter of all forms contain a single subdivision; 11% contain 3 subdivisions each. The only combination identified with 4 subdivisions was **—Biography—Dictionaries—Arabic—Early works to 1800** (v-v-x-v)

However, forms with 2 or more subdivisions were rarely assigned. Forms consisting of a single subdivision accounted for almost 95% of all assignments. The longer forms tend to be very specific, greatly limiting their applicability. Forms with a single subdivision were assigned an average of 12,587 times, those with 2 subdivisions were assigned an average of 311 times and forms with 3 or more subdivisions were assigned only an average of 14 times.

In general, the use of forms is very skewed; the 10 most assigned forms, **—Congresses, —Periodicals, —Biography, —Bibliography, —Directories, —Statistics, —Maps, —Handbooks, manuals, etc., —Catalogs,** and **—Fiction,** account for more than half of all assignments. The 100 most used forms account for more than 90% of all assignments. The remaining 2,463 forms account for less than 10% of all uses. The complete usage distribution is shown in figure 1.

Conclusion

Form subdivisions, which describe what the document "is" rather than what it is "about," represent an aspect of the subject distinct from the topical aspect. However, to effectively utilize the form information requires that the form subdivisions be explicitly identified. Until recently, when the Library of Congress started explicitly assigning the *$v* subfield code, form subdivisions were coded as general subdivisions making them indistinguishable from topical subdivisions. OCLC's WorldCat contains more than eight million unique subject headings that potentially could contain form subdivisions. Identifying these potential forms is difficult since many subdivisions can be either topical or form depending on the context. Manual efforts to recode form subdivisions are slow and error-prone due to the complexity of the coding guidelines.

An algorithm was developed to identify and recode form subdivisions in Library of Congress topical and geographic subject headings. The algorithm proved to be highly reliable with an error rate estimated to be less than 0.1%, significantly less than the observed error rate for manual coding. The algorithm identified 2,563 unique forms or combinations of form subdivisions in WorldCat. The usage of these forms was very uneven; the 10 most frequently assigned form subdivisions accounted for more than half of all assignments. Perhaps the greatest advantages of the algorithmic approach are the high accuracy rate and the ability to handle a large number of operations efficiently.

Table 1. Usage of Common Form Subdivisions

Total WorldCat Usage	Library of Congress Usage	Relative Use by the Library of Congress	Subfield Coding	Form Subdivision
1,109,724	317,800	161	v	Congresses
994,462	110,689	50	v	Periodicals
522,795	200,163	249	v	Biography
342,555	63,605	92	v	Bibliography
337,366	44,244	61	v	Directories
315,553	49,704	75	v	Statistics
315,532	91,673	165	v	Maps
300,303	44,642	70	v	Handbooks, manuals, etc.
280,777	45,163	77	v	Catalogs
275,552	81,549	169	v	Fiction
264,891	74,228	156	v	Exhibitions
229,330	67,552	168	v	Juvenile literature
189,742	11,682	26	v	Scores
181,334	18,542	46	v	Early works to 1800
122,016	42,750	217	v	Case studies
119,529	27,044	118	v—v	Statistics—Periodicals
113,277	24,665	112	v	Dictionaries
109,822	4,712	18	v	Scores and parts
77,729	16,968	112	v	Sources
70,872	9,178	60	v	Texts
67,174	16,941	136	v	Pictorial works
65,201	2,395	15	v	Excerpts
59,779	5,521	41	v	Juvenile fiction
58,980	20,452	213	v	Guidebooks
58,002	1,564	11	v	Vocal scores with piano
57,301	11,229	98	v—v	Bibliography—Catalogs
56,722	7,465	61	v	Problems, exercises, etc.
55,947	9,926	87	v	Indexes
54,698	5,119	41	v	Examinations, questions, etc.
53,669	2,371	19	v	Sermons
52,035	10,590	103	v	Poetry
47,310	3,120	28	v	Curricula
46,842	3,952	37	v	Software
46,436	920	8	v	Parts
40,967	154	2	v	Juvenile films
39,323	7,966	102	v	Drama
35,738	10,143	159	v	Cases
35,218	11,377	192	v	Popular works
34,810	1,950	24	v	Readers
34,026	831	10	v	Collections
33,077	4,394	62	v	Specimens
31,685	4,342	64	v	Tables
31,644	7,797	131	v	Classification
31,471	5,236	80	v—x	Dictionaries—English
30,681	990	13	v	Songs and music
30,469	11,123	231	v	Correspondence
29,369	4,800	79	v	Facsimiles
28,831	8,892	179	v	Interviews
28,261	10,939	254	v	Miscellanea
25,650	1,443	24	v	Librettos
25,548	4,863	94	v	Outlines, syllabi, etc.
25,254	6,609	142	v	Identification
23,340	1,207	22	v	Music
23,062	1,667	31	v	Designs and plans
22,330	3,731	81	v	Abstracts
20,660	1,122	23	v	Controversial literature
20,463	46	1	v	Slides
20,259	530	11	v	Excerpts, arranged

Table 1. Usage of Common Form Subdivisions, continued

Total WorldCat Usage	Library of Congress Usage	Relative Use by the Library of Congress	Subfield Coding	Form Subdivision
20,191	6,311	183	v	Translations into English
20,049	7,083	220	v	Folklore
19,271	252	5	v	Newspapers
18,741	4,554	129	v	Terminology
18,468	9,976	472	v—v	Biography—Juvenile literature
17,856	4,639	141	v	Forms
17,535	56	1	v	Photographs
17,226	1,822	48	v	Laboratory manuals
16,982	265	6	v	Studies and exercises
15,848	2,139	63	v	Personal narratives
15,547	1,692	49	v	Programmed instruction
15,391	2,839	91	v	Glossaries, vocabularies, etc.
15,378	2,205	67	v—v	Bibliography—Periodicals
15,182	127	3	v	Hymns
14,712	3,643	132	v—x	Dictionaries—German
14,664	3,238	114	v	Yearbooks
13,628	3,074	117	v—x	Dictionaries—French
13,427	4,430	198	v	Anecdotes
13,303	696	22	v	Solo with piano
12,740	5,019	261	v	Humor
12,611	794	27	v	Observations
12,269	2,468	101	v	Bio-bibliography
12,011	1,720	67	v	Textbooks for foreign speakers
11,961	405	14	v	Juvenile
11,909	3,585	173	v	Patterns
11,303	2,681	125	v	Atlases
11,078	1,069	43	v	Databases
10,978	2,672	129	v	Registers
10,790	4,494	287	v	Diaries
10,567	3,450	195	v	Amateurs' manuals
10,541	300	12	v	Textbooks
10,360	3,727	226	v	Digests
10,313	412	17	v	Methods
10,242	1,902	92	v	Study guides
10,064	473	20	v	Instrumental settings
9,887	3,002	175	v—x	Dictionaries—Chinese
9,813	507	22	v	Conversation and phrase books
9,735	2,859	167	v	Genealogy
9,622	2,063	110	v	Charts, diagrams, etc.
9,514	2,499	143	v—x	Dictionaries—Polyglot
8,994	2,519	156	v—x	Dictionaries—Japanese
8,983	21,46	125	v	Portraits

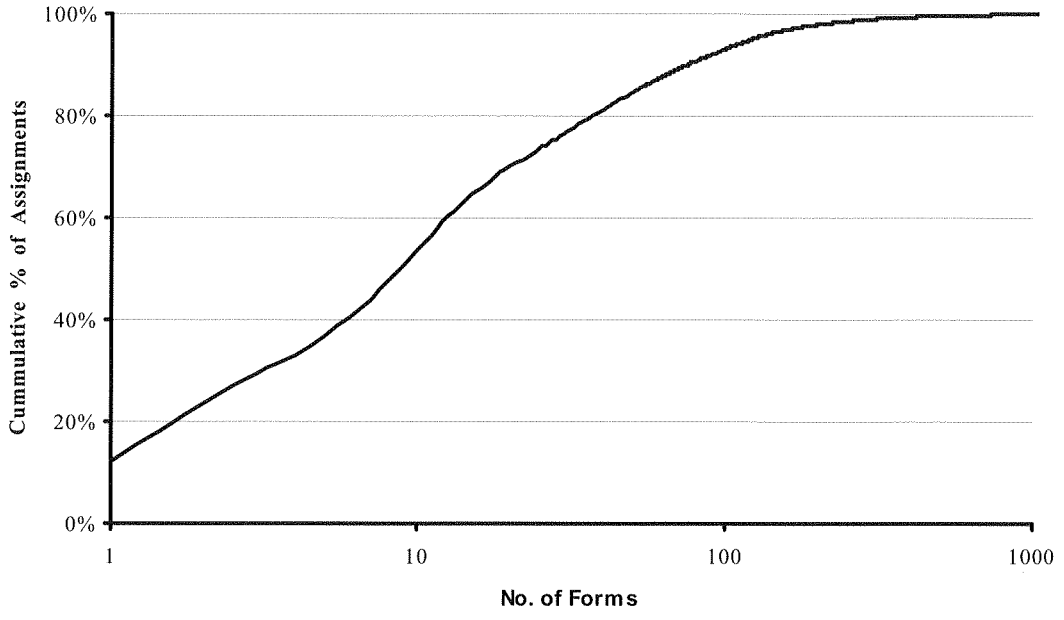

Figure 1. Form Subdivision Usage

Works Cited

Chan, Lois Mai, Eric Childress, Rebecca Dean, Edward T. O'Neill, and Diane Vizine-Goetz. 2001. A faceted approach to subject data in the Dublin Core metadata record. *Journal of Internet Cataloging* 4, no. 1/2: 35–47.

Haykin, David Judson. 1951. *Subject headings: A practical guide.* Washington, D.C.: Govt. Print. Off.

Hemmasi, Harriette, David Miller, and Mary Charles Lasater. 1999. Access to form data in online catalogs. ALCTS Online Newsletter 10, no. 4. Accessed Aug. 31, 2001, www.ala.org/alcts/alcts_ news/v10n4/formdat2.html.

Library of Congress. 1906. *Preliminary list of subject subdivisions.* Washington, D.C.: Govt. Print. Off.

———. 1975. *Library of Congress subject headings.* Washington, D.C.: Library of Congress.

Library of Congress. Cataloging Policy and Support Office. 1996–. *Subject cataloging manual: Subject headings.* Washington, D.C.: Cataloging Distribution Service, Library of Congress.

———. 2000. *Free-floating subdivisions: An alphabetical index.* 12th ed. Washington, D.C.: Cataloging Distribution Service, Library of Congress.

OCLC Online Computer Library Center, Inc., Office of Research. *Usage of form subdivisions.* Forthcoming, http://wcp.oclc.org/fast/.

O'Hara Conway, Martha, ed. 1992. *The future of subdivisions in the Library of Congress subject headings system: Report from the Subject Subdivisions Conference, May 9–12, 1991.* Washington, D.C.: Library of Congress, Cataloging Distribution Service.

Subject data in the metadata record recommendations and rationale: A report from the ALCTS/SAC/Subcommittee on Metadata and Subject Analysis. 1999. Accessed Aug. 31, 2001. www.govst.edu/users/gddcasey/sac/MetadataReport.html.