Year of Cataloging Research

The Distributions of MARC Fields in Bibliographic Records

A Power Law Analysis

By Matthew Mayernik

Library catalog systems worldwide are based on collections of MARC records. New kinds of Functional Requirements for Bibliographic Records (FRBR)*–based catalog retrieval systems, displays, and cataloging rules will build on ever-growing MARC record collections. Characterizing the kinds of information held in MARC records is thus an important step in developing new systems and rules. This study examined the incidence and prevalence rates of MARC fields in two different sets of library catalog records: a random selection of bibliographic records from the Library of Congress online catalog and a selection of records for two specific works,* Lord of the Flies *and Plato's* Republic*. Analysis showed that most fields were used in only a small percentage of records, while a small number of fields were used in almost all records. Power law functions proved to be a good model for the observed distribution of MARC fields. The results of this study have implications for the design of new cataloging procedures as well as for the design of catalog interfaces that are based on the FRBR entity-relationship model.*

MARC records are at the center of library cataloging processes. The MARC format, developed in the 1960s, is unlikely to be replaced in the foreseeable future, both because of its proven utility and because of the legacy volumes of existing MARC records held by libraries around the world. While the MARC format may not be going anywhere soon, the ways that MARC records are created and used are in a state of transition. *Functional Requirements for Bibliographic Records (FRBR)* outlined a conceptual model to describe the bibliographic universe.[1] Implementing the *FRBR* model in MARC-based cataloging practices and information retrieval systems has proven challenging. Numerous technical, structural, and institutional challenges must be overcome for libraries to shift to *FRBR*-based cataloging schemes and online catalog displays. Any new methods for cataloging and displaying library resources will be built from the existing MARC record databases. Thus understanding the state of the current data stored in MARC records is essential to the process of moving forward with new display systems and cataloging schemes.

This study serves the effort to better understand these challenges by more fully characterizing the kinds of information that can be found in MARC records. Specifically, this study aims to identify and characterize the patterns in the ways catalogers use MARC fields in bibliographic records by quantifying which fields are most commonly present in library catalog records. The author used two different samples of bibliographic records in this study. First, a random selection of

Matthew Mayernik (mattmayernik@ucla.edu) is a PhD student in the Department of Information Studies, University of California, Los Angeles.

Submitted May 10, 2009; returned to author June 19 with request to revise and resubmit; revised and resubmitted August 20; reviewed and accepted for publication September 26.

The author thanks Gregory Leazer and Jonathan Furner of the University of California, Los Angeles, as well as the editor and anonymous reviewers, for reading and commenting on earlier drafts of this paper.

bibliographic records from the Library of Congress (LC) online catalog were collected and examined. Second, a case study approach was used to analyze smaller samples of records from two specific works: William Golding's *Lord of the Flies* and Plato's *Republic*. This study tests whether a power law approach is useful in characterizing the distributions of MARC field in each sample and, if a power law distribution exists, what the implications are for the design of new *FRBR*–influenced cataloging schemes and catalog displays.

Background

This section describes the theoretical background for the analysis of MARC field use patterns reported in this study. First, the motivation and importance of the study is discussed, then power laws are introduced and described with the goal of illustrating how they can be used to model many phenomena both inside and outside of the library and information science domains.

Motivation for This Study

The motivation for a study of MARC fields in bibliographic records stems from the desire to understand the kinds of information that are available to build more informative displays into online library catalog interfaces. The deficiencies of online catalogs have been well documented. In separate studies, Calhoun and Markey pointed out how online catalogs have been slow to implement features that would greatly increase their utility, such as advanced retrieval techniques for subject searching, the inclusion of tables of contents, expanding the use of full-text searching, and leveraging classifications schemes as finding aids.[2] Certainly libraries faced many impediments in producing advanced catalogs, including financial limitations and the reliance on integrated library system vendors who were unable or unwilling to produce these additional functionalities.

Online catalogs have largely lacked the ability to identify and display relationships between different works and between representations of the same work.[3] Many different kinds of bibliographic relationships exist between library resources. Tillett identified seven types of relationships: equivalence, derivative, descriptive, whole-part, accompanying, sequential or chronological, and shared characteristic.[4] Smiraglia created a taxonomy that expanded on Tillett's derivative relationship and included seven categories: simultaneous derivations, successive derivations, translations, amplifications, extractions, adaptations, and performances.[5] Bibliographic relationships between library resources are common. Smiraglia and Leazer found that approximately 30 percent of bibliographic works in the OCLC's WorldCat have associated derivative works.[6] These relationships are manifested in a number of ways in MARC records, including through uniform titles, series statements, and added entries.[7] Despite this, most conventional library catalogs provide little in the way of collocation based on bibliographic relationships. Integrating these relationships into catalog displays would provide users with a significantly more powerful way to navigate through library resources.[8]

FRBR is the most visible effort to give bibliographic relationships a more central role in modeling the bibliographic universe. *FRBR* describes a conceptual model that identifies four main bibliographic entities: works, expressions, manifestations, and items. The first three entities are abstract concepts while the fourth entity, the item, represents the physical resource that exists on a library shelf. The *FRBR* model has been criticized for having a lack of conceptual clarity in the distinctions between the abstract work, expression, and manifestation entities, and for glossing over important differences between books and nonbook materials.[9] Despite these criticisms, the next generation cataloging code, Resource Description and Access (RDA), is integrating the *FRBR* entity-relationship model into the arrangement and implementation of the new cataloging rules.[10]

Power Laws in Library and Information Science

This study uses power law functions to characterize the patterns of MARC field use in bibliographic records. A power law function is a mathematical expression that describes an inverse exponential relationship between two phenomena. Power laws are commonly illustrated through the "80/20 rule" of wealth and power, that is, 80 percent of the world's resources are held by 20 percent of the world's countries, or by the "long-tail" phenomena of marketing and consumption, where very few music or book titles sell a large number of copies and a great many titles sell very few copies.[11] Power law functions have been used extensively in the library and information science literature. A study in 1995 showed that the individuals behind two of the classic bibliometric power laws, George Kingsley Zipf and Alfred J. Lotka, were at that time among the most cited people in the history of the discipline.[12]

Zipf's and Lotka's power laws provide similar formulations for different kinds of bibliometric phenomena.[13] Zipf derived his law from a study of word counts in a selection of English language texts. He showed an inverse relationship between the number of times a word is used and its use rank with the set of all words used. So, if the most frequently used word was used one hundred times, the second most frequently used word was used roughly fifty times (one-half as many), the third most frequently used word was used roughly thirty-three times (one-third as many), and so on down the word list. Lotka, on the other hand, derived his law from a study of the publication productivity of individual authors within a corpus of chemistry and physics journals. Lotka found an inverse-square relationship between the number of publications by

each author and the number of authors with a given number of publications. In other words, if one hundred authors produced one published paper, the number of authors that produced two published papers was roughly twenty-five (one-fourth as many), the number of authors who produced three published papers was roughly eleven (one-ninth as many), and so on. The authorship and publication patterns of many disciplines, including the library and information sciences, have been shown to follow power law distributions.[14]

Zipfian and Lotkan distributions have been observed in a number of other library and information science settings. Power laws have been used to describe the forms of names on bibliographic records, the frequency of name headings in the library catalog, library resource circulation patterns, and the use of descriptor term co-occurrences in a bibliographic hypertext system.[15] In 1990, Blair proposed that Zipf's distribution of word use might be used as an indicator of indexing effectiveness.[16] He suggested that the distribution of index term use should match distribution of word usage in documents. According to Blair, a match in term usage distributions would indicate that the indexers and users were using language in a similar fashion and thus bring the conceptions of document representation between the two groups closer together.

The scope of power law functions extend beyond the study of word counts and author productivity, however; power law relationships can be used to describe many natural and human phenomena, such as the size of cities, earthquakes, and forest fires. Newman described a number of theories proposed to explain the existence of power law forms in such a wide variety of phenomena.[17] One of the most important physical mechanisms Newman identified as explaining the occurrence of power laws is the Yule process, better known as the "rich get richer" phenomena. The Yule process is named for the developer of the first mathematical description of this process, G. Udny Yule. The description of the Yule process for the population size of cities, which follows a power law distribution, is that new instances, in this case new people moving to a city, occur in proportion to the number of people already living in each city. In other words, large cities are much more likely to add new members than small cities or towns, leading to the situation that currently exists where there are very few cities with a very large population and a large number of cities with a small population. Similarly for book sales figures, book titles that have already sold large numbers of copies are much more likely to sell more copies, while titles with lower sales figures are less likely to sell additional copies. Newman described how it has been shown mathematically that this "rich get richer" process leads to power law distributions.

The present study extends the application of power law functions to a new area: the distribution of MARC fields in catalog records. The author was only able to find one work that explored the distribution of field use in MARC records. Markey and Calhoun studied the prevalence of fields that provided "subject-rich" words, which they define as words that would be useful when performing subject searches in online catalogs.[18] They found that less than 5 percent of bibliographic records they studied contained the MARC 505 (Formatted Contents Note) and 520 (Summary, etc.) notes fields. No other work on this topic was found. The next sections describe the method used to sample and analyze bibliographic records from the LC website and outline the major results. This is followed by a discussion of the methods used to collect a smaller targeted sample of records for the two case study works, *Lord of the Flies* and *Republic*.

Research Method—Random Sample

The first set of records included in this study came from the LC online catalog (http://catalog.loc.gov). The author randomly sampled and examined 1,500 MARC records and recorded their data fields. Following the examination of each record, an analysis identified the most widely used MARC fields across the sample and determined if power laws would provide a good model for the statistical distribution of MARC field use. This section describes the methods used to collect a random sample of MARC records from the LC online catalog as well as the method used to count the fields in each record.

The random sample analyzed in this study consisted of 1,504 LC MARC records that were collected in early March 2009. The author collected the records from the LC online catalog using a script written in the Python programming language. The process used to collect the records was based on the system of LC Control Numbers (LCCNs) used by the LC. Since 1898, each item cataloged by the LC has been assigned an LCCN. These numbers are either eight or ten digits in length and consist of two concatenated segments. The first segment (either two or four digits) indicates the year the LCCN was assigned. For records created prior to January 1, 2001, the year segment is the last two digits in the year; for example, the LCCN year for 1987 is "87." For records created after January 1, 2001, all four digits of the year are used in the LCCN, for example, "2007." Because the change to a full four-digit year segment did not occur until 2001, records that begin with the digits "98" indicate records from 1898 and 1998, "99" indicates records from 1899 and 1999, and "00" indicates records 1900 and 2000. For these records, the years can be distinguished by the second segment, a six-digit assigned control number. For most years, the control numbers begin with the "000001" and are assigned incrementally as six digits: 000001, 000002, and so on. The records from 1998, 1999, and 2000, however, have been assigned control numbers that follow those from 1898, 1899, and 1900, respectively. Additionally, the

LCCNs for the years 1969–72 use a different numbering scheme, described as follows on the LC website, "During the 1969–1972 period, a 7-series year number was assigned. In these numbers the initial digit of 7 was followed by a modulus-ll check digit."[19] The importance of this different scheme for this study is that the LCCN numbers during those years do not follow the same pattern as all other years. This is discussed further in the description of the sampling algorithm.

The author collected the MARC field data for this study from the LC online catalog using the MARCXML LCCN Permalink pages. The Permalink page provides an XML representation of the MARC catalog record for a given item held by the LC. For example, the LCCN Permalink for the first records from 1905 and 2005 can be found at http://lccn.loc.gov/05000001/marcxml and http://lccn.loc.gov/2005000001/marcxml, respectively.

To illustrate more clearly, the first few lines from the MARCXML page for one LCCN that was included in this study (2008448698) is given here:

```
<record>
   <leader>00979cam a22002774a 4500</leader>
   <controlfield tag="001">15355205</controlfield>
   <controlfield tag="005">20081030122910.0</
      controlfield>
   <controlfield tag="008">080617s2008 fr b 000 0 fre </
      controlfield>
   <datafield tag="906" ind1=" " ind2=" ">
      <subfield code="a">7</subfield>
      <subfield code="b">cbc</subfield>
      <subfield code="c">origres</subfield>
      <subfield code="d">3</subfield>
      <subfield code="e">ncip</subfield>
      <subfield code="f">20</subfield>
      <subfield code="g">y-gencatlg</subfield>
   </datafield>
   <datafield tag="925" ind1="0" ind2=" ">
      <subfield code="a">acquire</subfield>
      <subfield code="b">1 shelf copy</subfield>
      <subfield code="x">policy default</subfield>
   </datafield>
```

The MARC fields are given in the tag that follows each <controlfield> or <datafield> tag. In this truncated example, the control field tags are 001, 003, and 008, and the data field tags are 906 and 925. The MARCXML pages also give the subfields and indicators for each data field, as shown above, but the subfield tags and indicators were not collected or analyzed in this study.

The record sampling and collection algorithm followed these steps: First, the Python script generated a random year between 1898 and 2009. Depending on the year generated, the number of digits was adjusted to the appropriate two- or four-digit year length. A random six-digit control number was then generated. For years prior to 1969, the random number was generated between 000001 and 200000, while for the years between 1969 and 2009 the random number was generated between 000001 and 899999. This difference in the number range reduced the bias toward more recent records that stems from the dramatic increase in publishing volumes over the past thirty years (and the corresponding increase in the number of LCCNs assigned per year). Changing the range at 1969 also was an attempt to reduce problems relating to the idiosyncratic pattern of LCCN assignment for the years 1969–72 discussed above. The six-digit control number was then concatenated onto the year digits to create the full eight- or ten-digit LCCN. With this number in hand, the Python script sent an HTTP request to the LC website for the corresponding MARCXML LCCN Permalink page.

After receiving a response from the LC server, the script checked for an <error> tag in the XML. The presence of an <error> tag indicated that the randomly generated LCCN did not have an associated record. This was common, as many randomly generated numbers were not assigned to any record. For example, requesting the page http://lccn.loc.gov/50100000/marcxml returns an error because fewer than 100,000 LCCNs were assigned in 1950. When this occurred, the script generated a new random LCCN. This process of generating random LCCNs and requesting Permalink pages was repeated until a good record was found, as indicated by the absence of the <error> tag. When an error-free MARCXML Permalink page was downloaded, the script recorded the <controlfield> and <datafield> tags, which, when compiled, provide the list of MARC fields occurring in that record.

In performing an initial analysis, the author found the data from four records to be faulty because of problems that occurred during the data collection process, and those records were thus removed from the data set. For example, in one case what appeared to be one record was actually two records, 98171673 and 99111995, that were collected under a <marcCollection> tag on the XML page for 98171673. Because this misrepresented the field counts for that record, it was excluded. The other excluded records had similar issues. Excluding these four records left a total of 1,500 records for analysis.

Figure 1 shows the distribution of sampled records by year. This distribution looks mostly random, with the exception of 1965–71, 1898–1900, and 1998–2000. The number of samples taken from each year tends to increase through the twentieth century and into the twenty-first century, which is to be expected from a random sample, as the publication of titles has increased over the century. The totals for 1965–68 were artificially high and the totals

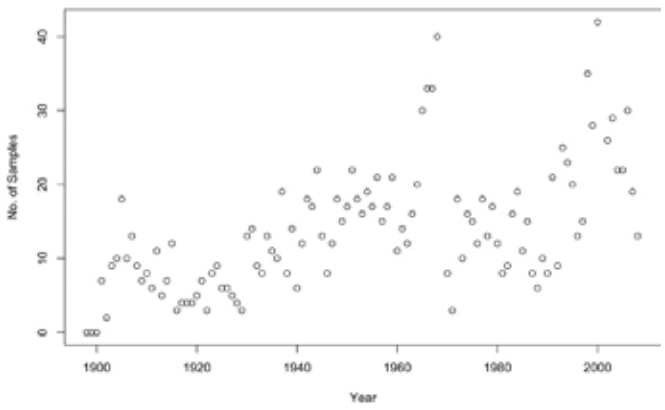

Figure 1. Distribution of Records Sampled from the Library of Congress Online Catalog by Record Year

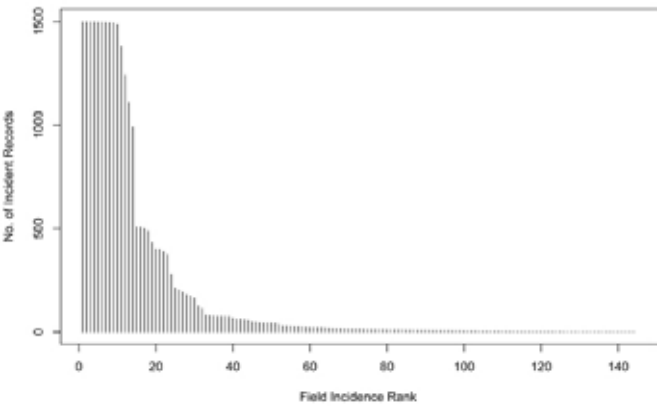

Figure 2. Rank-Frequency Plot of the Field Incidence for the Library of Congress Records

for 1969–71 were artificially low because of the algorithm used to generate LCCNs described above. No samples were records from 1898–1900 because the sampling algorithm was biased to take into account the LCCN inconsistencies described in the first paragraph of this section. Thus all sampled records with LCCNs starting with 98, 99, and 00 were from 1998–2000. Even with these sampling anomalies, the highest number of records sampled from a given year was 42 in 2000, which constitutes less than 3 percent of the total sample population.

The author used two metrics to count fields in records: field incidence, defined to be the presence or absence of a field in a record, and occurrence, defined to be the total number of times a record uses a field. In other words, when counting field incidence, each field is only counted once per record, even if it is used more than once in a given record, whereas occurrence counts all instances of that field. Thus a record that contained three 700 fields received an incidence value of one and an occurrence value of three for the 700 field.

Results and Analysis of Random Sample Research

The 1,500 sampled records contained 29,689 fields. The mean and median numbers of fields per record were 19.8 and 19, respectively, with a maximum of 80, a minimum of 10, and a standard deviation of 4.95. The vast majority of the sampled records (91 percent) contained between 13 and 26 fields. Only 1 percent of the sample (two records) contained less than 13 fields, and 8 percent of the sample contained more than 26 fields. One record (53035190, a record for a German periodical) contained 80 fields, placing it far outside the main group. Of the 80 fields in this record, 31 were 991 fields, which is a preprocessing location/conversion field. Even with the 991 fields removed, however, this record would still have the most fields of any record in the sample.

Field Incidence and Occurrence Rates

The author observed 144 MARC fields in the 1,500 record sample, and the incidence rates ranged widely for different fields. Each field was incident in an average of 190 records. The median incidence rate, however, was far lower, at 14 records per field. The difference in the mean and median values indicates a large skew in the field incidence rate. Four fields were incident in all 1,500 records in the sample, and the minimum number of incidences was one record. Nineteen fields were only observed in a single record. Figure 2 shows a rank–frequency plot of the field incidence. In this figure, the y-axis shows the number of incident records that contained at least one instance of a given field, with the x-axis showing the rank in order of frequency of incidence. The vast majority of fields were incident in only a small number of records, but a small number of fields appear in nearly all of the sampled records. Appendix A lists the 23 most incident fields. All other fields were incident in less than 20 percent of the sampled records. The curve in figure 2 shows the obvious shape of a power law distribution, with the notable exception of the ten top ranked fields, which, as appendix A shows, were all incident in greater than 99 percent of the sampled records.

The next step was to test how the Zipf distribution fits the distribution of MARC fields in bibliographic records. The equation for Zipf's law is $f = c/r^a$, where f is the frequency of incidence, r is the rank, c is a constant, and following Zipf's original formulation, $a = 1$.[20] Figure 3 shows the same data with the ten highest ranked fields that were incident in 99 percent or more of the sampled records collapsed into a single entry with an assigned frequency of 100 percent. The two lines represent Zipf functions with c = 30 (solid line) and c = 45 (dotted line). The best fitting value lies somewhere

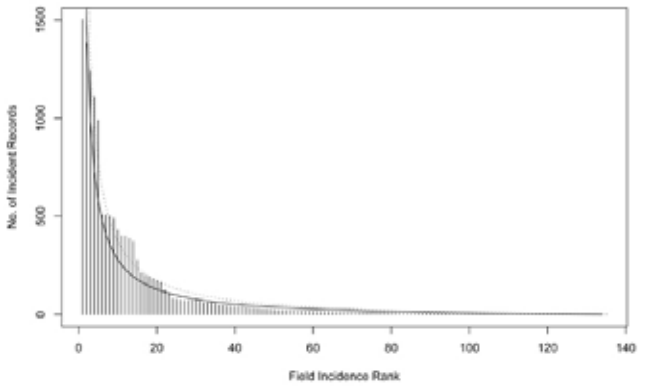

Figure 3. MARC Field Incidence Rates, with Zipf Functions f = 30/r (solid lnie) and f = 45/r (dotted line)

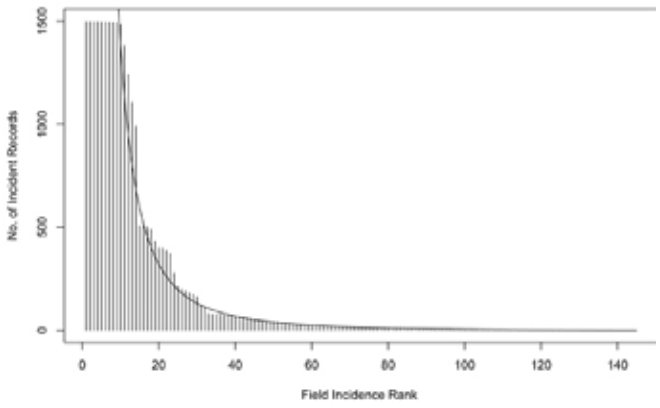

Figure 4. MARC Field Incidence Rates from Figure 2, with Lotka Function $y = 1500/x^{2.005301}$

between these two values. As this figure shows, the Zipf functions appear to provide a good model for the field incidence rates when the highest ranked fields are collapsed into a single entry.

But is this collapsing necessary? Can the incidence data be modeled even with the inclusion of the top ranked fields? The next step was to see if the Lotka function fit the data better than the Zipf function. Egghe provides a method for fitting Lotka functions to informetric data.[21] The formula for the Lotka function is similar to the Zipf function, $y = c/x^n$, where y = the number of records in which a field is incident, x = the rank in order of incidence rate, and c is a constant, as is n, which may or may not equal one. Following Egghe, the values of c and n can be found in the following manner:

> c = y(1) = the number of incident records for the top ranked field = 1,500. And n = (2*A-T)/(A-T), where A = 27,310 = the total number of incident records (sum of the third column in appendix A), and T = 144 = the number of unique MARC fields observed. Thus, n = 2.005301.

Plugging these values into the Lotka function and plotting it alongside the data created figure 4. In this figure, the line represents the function $y = 1500/x^{2.005301}$. As this figure shows, this function fits the data very well. Thus the Lotka function appears to give a good model for the incidence rates of MARC fields in the LC bibliographic records.

MARC Field Occurrence Rates

As explained above, incidence rates measure how many records contained a given field at least once, and occurrence rates measure how often each field is used in a given record. Many fields occurred more than once in an individual MARC record. Counting all occurrences of individual fields, the mean number of occurrences per field was 206 with a median of 14 occurrences (compare with the incidence mean of 190 and median of 14). The maximum number of occurrences of a single field was 1,817 for the 650 field, and the minimum number of occurrences was 1 for 19 different fields.

The total number of field occurrences across all records does not tell a great deal about the use of individual fields, however, as the incidence rate has a large effect on the occurrence rate. Thus the occurrence rates for each field were normalized by dividing them by their incidence rate. Appendix B shows the 20 highest ranked MARC fields in order of their normalized occurrence rates. The last column shows the total number of times individual fields occurred. The 880 field (Alternate Graphic Representation) had the highest normalized occurrence rate, an average of 4 per each incident record.

The average normalized occurrence rate for the observed fields was 1.10 field occurrences per record, and this rate again showed the power law shape. A Zipfian distribution line showed the best fit to the normalized occurrence rate, as illustrated in figure 5. In this plot, the equation for the solid line is f = 1+ 1/(20*r), and the equation for the dotted line is f = 1+ 1/(30*r). Thus c is approximately equal to 1/20 or 1/30.

Research Method—Case Studies

This section describes an analysis of fields in MARC records for two specific works, William Golding's *Lord of the Flies*, and Plato's *Republic*. Plato's *Republic* represents a canonical text in modern Western society with a publication history that extends back for hundreds of years and, correspondingly, an extensive bibliographic family. *Lord of the Flies* represents a contemporary work, but one that is popular enough to have been published in a number of editions, translations, and compilations. The case studies were chosen to represent works that typically would be found in library catalogs and have bibliographic families large enough to benefit from

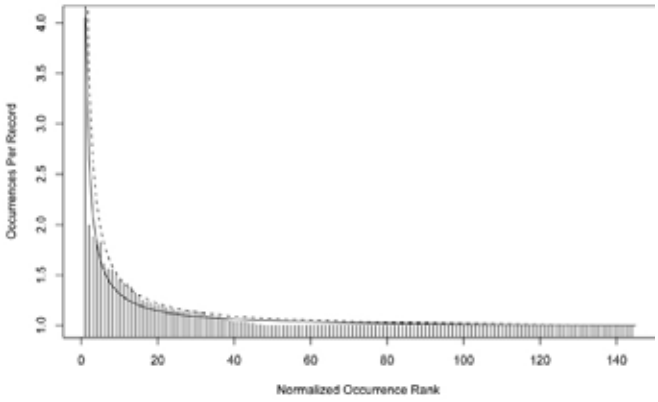

Figure 5. Distribution of Fields for the Library of Congress Records by Normalized Occurrence Rate, with Zipf Functions f = 1+1/(20*r) (solid line) and f = 1+1/30*r) (dotted line).

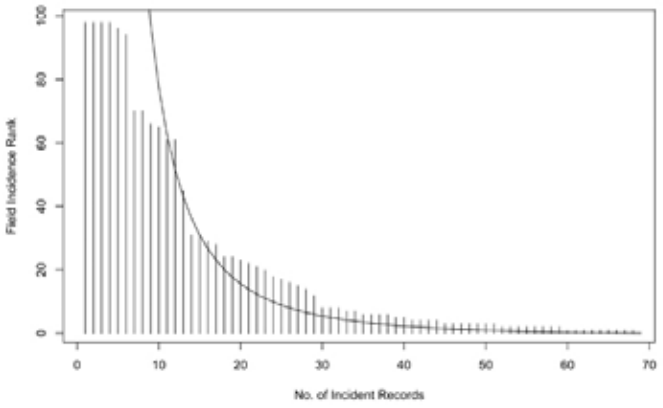

Figure 6. *Lord of the Flies* Frequency of MARC Field Incidence, with Lotka Function $y = 98/x^{2.047519}$

more "*FRBR*-ized" methods of display and access.

The author collected the MARC records for the case studies from the University of California, Los Angeles (UCLA) online catalog and the Research Libraries Group RLIN Union Catalog in early 2007. One hundred MARC records in the UCLA online catalog were related to Plato's *Republic*, but only 17 records were related to William Golding's *Lord of the Flies*. In light of the small number of records available through the UCLA catalog, a title search for "Lord of the Flies" was run in RLG's RLIN Union Catalog, and 98 records were pulled from the highest volume search result. This entry contained 106 records, but 8 unrelated records were excluded (an example is *The Best American Science Writing 2000*, which containing a work entitled "Lord of the Flies" by Jonathan Weiner, an essay on fruit fly breeding), leaving a 98 record sample. These samples did not include all manifestations of these two works, but rather should be taken as representative samples of their extended bibliographic families.

The author examined each record individually for both works and recorded the data fields present in each record. As before, MARC fields were examined at the number level only; presence or absence of subfields or indicators was not recorded. The MARC incidence data were then accumulated for each work separately and incidence rates were calculated.

Results and Analysis of Case Studies

The case study results generally followed the results of the random sample analysis. For the *Lord of the Flies* records pulled from the RLIN union catalog, there was an average of 19 fields per record (maximum = 53 fields, minimum = 8 fields), with a standard deviation of 8.8. For the *Republic* records pulled from the UCLA catalog, there was an average of 22.5 fields per record (maximum = 46 fields, minimum = 12 fields), with a standard deviation of 5.3. The two works exhibited 87 different fields. Appendix C shows the fields that appeared in more than 20 percent for the records of each work. For the *Lord of the Flies* records, 34 percent of the incident fields (23 of 68) had incidence rates greater than 20 percent, and similarly, for the *Republic* records, 39 percent of the incident fields (24 of 62) had incidence rate greater than 20 percent. These results echo what was found in the random sample discussed above, though the incidence rates of some fields varied between the two works and the random sample.

Fitting Lotka functions to the MARC data from these works was not as successful as for the random sample. Applying Egghe's method, one can try to fit Lotka functions for each work.[22] For the *Lord of the Flies* data, c = 98 (the number of incident records for the top ranked field), A = 1,499 (total number of incident fields), T = 68 (number of unique MARC fields), giving n = 2.047519. For the *Republic* data, c = 100, A = 1,918, and T = 62, giving n = 2.033405.

Figures 6 and 7 show rank-frequency plots of the MARC incidence data for both works. Figure 6 displays the *Lord of the Flies* data, with a line representing the Lotka function $y = 98/x^{2.047519}$. Figure 7 displays the *Republic* data, with a line representing the function Lotka function $y = 100/x^{2.033405}$.

The power law curves in these figures generally follow the trends of the data, though not as well as the data from the random sample of LC records. This is likely because of the smaller number of records sampled as well as the differences in cataloging practice that generated the records. For example, the most obvious difference between the two case studies is that only 6 fields were present in more than 95 percent of the *Lord of the Flies* records, but 11 fields were present in the *Republic* records at that rate. Looking more closely at appendix C, 3 of the fields found in 100 percent of the *Republic* records are processing fields that were either for local UCLA use (000 and 910) or were not included in

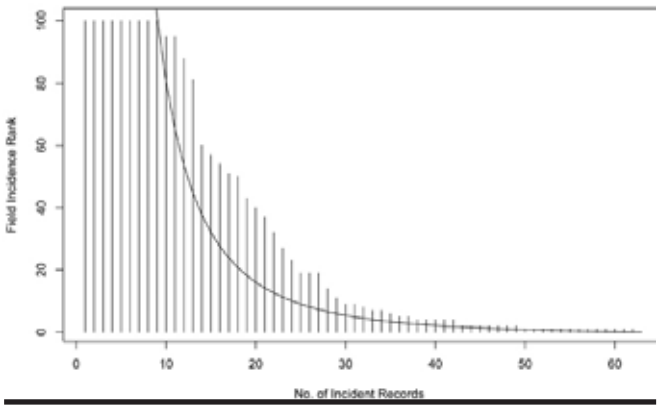

Figure 7. *Republic* Frequency of MARC Field Incidence, with Lotka Function $y = 100/x^{2.033405}$

the RLIN catalog (005). The presence of additional processing fields in the *Lord of the Flies* records would more closely align the results of the two case studies.

Discussion

What are the implications of Zipf and Lotka distributions of MARC fields in bibliographic records? As Tague noted, observing or quantifying these distributions in a given set of materials means very little by itself; instead the focus needs to be on explanations or ramifications of the analysis.[23]

First, if a prospective goal of catalog systems is to generate displays that link related records in a *FRBR*-like way, these field distributions may give some indication of the available data for doing so. Looking at the fields that are found in the most records (see appendix A), of the 12 fields incident in more than 80 percent of the sampled LC records, 9 of them are processing, control, or classification fields. The other 3 are the "workhorse" description fields of the MARC record (245, 260, and 300), which all occur in more than 99.5 percent of the random sample. After these 12 fields, the incidence rate drops off rapidly, with the 100 and 650 fields being the only other fields that were incident in more than 50 percent of the randomly sampled records. A similar pattern can be seen in the case studies, although some fields were more important to the case studies than the larger LC sample. For example, the 700 (Added Entry) field was found in 62 percent and 88 percent of the *Lord of the Flies* and *Republic* records respectively, which is not surprising given the number of translations, compilations, and annotated editions that can be found in the bibliographic families of each of these works, whereas the 700 field was found in 26.7 percent of the random LC sample. Additionally, the 500 (Notes) field was found more often in the case studies than in the random LC sample, being incident in 62 percent and 50 percent of the *Lord of the Flies* and *Republic* records respectively, as opposed to 33.8 percent in the LC records.

Looking closer at the incidence data, other patterns emerge. Table 1 shows the incidence rates for sets of fields. While individual fields may not occur very often—such as the 110 field, which was incident in 11.5 percent of the randomly sampled LC records—any occurrence of a 1XX field in a record provides important information. The cumulative incidence rates for these sets of fields are higher than for any individual field in each set. For example, the 5XX fields occurred in 55 percent of the randomly sampled LC records, but the highest value for any individual 5XX field was the 500 field, which was incident in 34 percent of the sample. Similarly, the 650 field was incident in 66 percent of the sampled records, but the set of 6XX fields were incident in 84 percent of the sample. These cumulative incidence values are perhaps more useful than the values for individual fields in estimating the types and quantity of data available for identifying and characterizing the bibliographic relationships between records.

Combining this table with appendix A, one can see that the main sources of information in MARC records are found in the 1XX, 245, 260, 300, and 6XX fields, with the 5XX, 71X–75X, and 4XX fields providing some additional information. The added entries fields, 71X–75X, are an important source of information in determining relationships between records, and at a nearly 40 percent combined incidence rate they are by far the most widely used of such fields. Table 1 illustrates how the 76X–78X fields, which allow the cataloger to indicate direct relationships between records (conventionally used in describing

Table 1. Cumulative Incidence Rates for the Randomly Sampled Library of Congress Records

Fields	Field Description	Incident Records	Incidence Rate (%)
1XX	Main Entry	1303	86.87
4XX	Series Statement	306	20.40
5XX	Notes	824	54.93
6XX	Subject Access	1266	84.40
71X–75X	Added Entry	596	39.73
76X–78X	Linking Entry	21	1.40
80X–83X	Series Added Entry	114	7.60
841–88X	Holdings, Location, Alternate Graphics, etc.	131	8.73

relationships between serial titles though not specifically reserved for that purpose), are rarely used. Thus, even if these fields can be mapped to the *FRBR* model, they will be of little use when actually creating a "*FRBR*-ized" catalog.[24] Uniform title and edition statement fields, which provide additional record linking information, also had small incidence rates. The 130 (Main Entry—Uniform Title) field occurred in 17 records, a 1 percent incidence rate, and the 240 (Uniform Title) field occurred in 57 records, a 4 percent incidence rate. The 250 (Edition Statement field occurred in 192 records, a 12.8 percent incidence rate.

Why Do MARC Fields Exhibit Power Laws?

Moving now to the observed power law relationships themselves, one may ask why these relationships were observed. Can these power law relationships in MARC field use by catalogers be accounted for in some way? Earlier, the "rich get richer" process was discussed as one possible mechanism that can lead to the observation of power law relationships in a wide variety of settings.[25] This "rich get richer" process is likely an important mechanism in relation to the findings of this study. The most widely used fields follow long established cataloging traditions. Charles Ammi Cutter, in his well-known 1904 work *Rules for a Dictionary Catalog,* describes the objects of the catalog as the following:

1. To enable a person to find a book of which either
   A. The author is known.
   B. The title is known.
   C. The subject is known.
2. To show what the library has
   A. By a given author.
   B. On a given subject.
   C. In a given kind of literature.
3. To assist in the choice of a book
   A. As to its edition (bibliographically).
   B. As to its character (literary or topical).[26]

To achieve these objectives, Cutter proposed the following means:

1. Author-entry with the necessary references (for A and D).
2. Title-entry or title-reference (for B).
3. Subject-entry, cross-references, and classed subject-table (for C and E).
4. Form-entry and language-entry (for F).
5. Giving edition and imprint with notes when necessary (for G).
6. Notes (for H).[27]

Cutter's catalog objectives and means for achieving them identified the kinds of information most useful to achieving what in his view were the objectives of the catalog. Cutter was highly influential to the subsequent development of cataloging practice, particularly to the code-developing work of Seymour Lubetzky, and his catalog objectives (and corresponding means to achieve them) were ultimately encoded into the *Anglo-American Cataloguing Rules*.[28]

Looking back at appendix A, one sees a close correspondence between the MARC fields with the highest incidence rates from the random sample of LC records and Cutter's list of means for achieving his catalog objectives. The 100 field (74 percent incidence) meets means 1, the 245 field (99.9 percent incidence) meets means 2, the 050 and 650 fields (99 percent and 66 percent incidence respectively) meet means 3, the 008 field (100 percent incidence) meets the language-entry aspect of means 4, and the 260 and 300 fields (both 99.7 percent incidence) meet means 4 and 5 on Cutter's list. Thus, in a "rich get richer" manner, certain kinds of information were identified to be important by scholars such as Cutter, and over time the practice of including these particular kinds of information became "richer" through wider incorporation into cataloging practices and cataloging code.

Power Laws and Cataloging Code

The relation of cataloging code to the power law distribution of MARC fields is another important implication of this study. It may seem tautological to say that there are important relations between the distributions observed in this study and the *Anglo-American Cataloging Rules*, given that a large portion of the records included in this study were cataloged to either the first or second edition of the *Anglo-American Cataloguing Rules*, but with the pending release of the RDA next generation cataloging rules, probing these relations may shed light on potential issues in converting to the new rules.[29] The *Anglo-American Cataloguing Rules*, 2nd edition (*AACR2*) is organized in such a way that the fields found to be most prevalent in MARC records are given prime importance. *AACR2* is broken into two main sections: Part 1 provides the rules for describing library resources, and Part 2 provides the rules for creating headings, uniform titles, and references. Focusing on Part 1, the first chapter of *AACR2* gives the general rules for description and is followed by eleven chapters that enumerate sets of rules for dealing with particular kinds of resources, such as books, pamphlets, and printed sheets (chapter 2); cartographic materials (chapter 3); and manuscripts (chapter 4). Each chapter contains section headings for the following description areas: title and statement of responsibility, edition, material specific details, publication and distribution, physical description, series, note, and standard number. These main headings encompass the nonprocessing fields that in this study had the highest incidence rates in catalog records (illustrated in appendix A), as well as fields that were

less prevalent, such as the edition statement, material specific details, and series statement. The rules specific to the lesser used fields are found by looking further into the chapters for each particular material type. The material-specific chapters in *AACR2* emphasize certain description areas in greater or lesser detail depending on the needs of each type of material. For example, the physical description area in chapter 5, "Music," has six subsections under the physical description area, whereas chapter 6, "Sound Recordings," has nineteen subsections under the physical description area.

The organization of the rules in RDA is notably different.[30] RDA's main organizational scheme draws on the conceptual models found in the *FRBR* and *Functional Requirements for Authority Data (FRAD)* reports.[31] RDA is broken into ten sections. Each section contains between one and five chapters devoted to recording information about particular entities in the *FRBR* and *FRAD* models. For example, the first four sections are "Recording Attributes of Manifestation and Item," "Recording Attributes of Work and Expression," "Recording Attributes of Person, Family, and Corporate Body," and "Recording Attributes of Concept, Object, Event, and Place." Subsequent chapters specify rules for recording information about relationships between entities. Looking closer at RDA, the location of the rules for recording high-incidence fields (title and statement of responsibility, publication and distribution, physical description, and notes) are not organized as linearly as they are in *AACR2*. The rules for recording the title, statement of responsibility, publication and distribution details (as would be recorded in the MARC 100, 245, and 260 fields) are found in chapter 2, "Identifying Manifestations and Items," and the rules for recording the physical description (MARC 300 field) are found in chapter 3, "Describing Carriers." Rules for creating notes are found throughout both chapters. The material-specific rules—such as those for music, recorded sound, video, etc.—are mixed into each chapter of RDA rather than the *AACR2* practice of giving them their own chapters. This can be illustrated by looking at five rules in RDA specific to printed music. In chapter 2, which contains 228 pages, the rule for type of musical composition, medium of performance, key, etc., is found on page 25; the rule for devised titles for music is on page 56; and the rule for publisher's number for music is on page 194. Similarly, in chapter 3, which contains 141 pages, the rule for extent of notated music is on page 30 and the rule for score and parts in a single physical unit is on page 134. These five RDA rules correspond to rules 5.1B1, 5.1B2, 5.4D3, 5.5B, and 5.5B1 in *AACR2*, all of which are found within the twenty pages of chapter 5, which is dedicated to printed music. This is only a small selection of the rules specific to printed music in RDA; many other music specific rules are spread around other chapters. The same illustration could be made for any other kind of material.

These differences between RDA and *AACR2* raise important questions about the accessibility of RDA to cataloging students as well as the ease of transition to the new rules by experienced catalogers. Whether the less direct correspondence between the organization of the rules in RDA and the most widely used MARC fields provide as clear a roadmap as *AACR2* for students learning to catalog is an open question. Additionally, regarding the co-location of material-specific rules in *AACR2*, and the lack thereof in RDA, few situations require that catalogers use rules specific to two different kinds of materials in cataloging a single item. Further research will be necessary to show whether the new rule organization scheme in RDA speeds up or slows down catalogers, both novice and experienced, as they find and apply the less-used but still important material-specific rules.

Conclusions

Library catalog systems worldwide are based on collections of MARC records. New kinds of catalog retrieval systems, displays, and cataloging rules, whether they are based on the *FRBR* entity-relationship model or another model, will build on these ever-growing MARC record collections. Characterizing the kinds of information held in MARC records is thus an important step in developing new systems and rules. This study examined the incidence and prevalence rates of MARC fields in two different sets of library catalog records. First, an analysis of a random sample of 1,500 MARC records from the LC online catalog found that most of these records contained between 13 and 25 fields. Ten fields occurred in more than 99 percent of all of the randomly sampled records. Further analysis showed that the rates of MARC field incidence fell off rapidly for less-used fields and that the rate of drop-off in use can be modeled very accurately by Zipf and Lotka power law functions. Second, a similar analysis on records for two specific works, William Golding's *Lord of the Flies* and Plato's *Republic*, tested whether trends found in a large random sample would hold for smaller subsets of records. Overall, the trends were similar. Most records consisted of an average of 19 MARC fields, and the majority of fields occurred in less than 20 percent of the records for each work. The incidence data from these two works followed the power law shape but did not fit power law curves as well as the random sample of LC records; this is most likely because of differences in the cataloging practices of the record sources and the smaller number of records in the case studies.

These results have important implications for the design of "*FRBR*-ized" catalog displays. The fields that explicitly create links between records, such as the 76X–78X fields and the 130 and 240 uniform titles fields, have low incidence rates. However, many implicit links, such as those created

by the 71X–75X added entry fields, are available in records. These implicit links enable work on algorithmic methods to pinpoint relationships between records and to cluster records for display to a catalog user. Aalberg, as well as Hickey and O'Neill, provides illustrations of current work in creating such algorithms.[32]

The results of this study highlight some issues that may arise in the transition from *AACR2* to RDA. Looking at the organization of the rules in RDA and *AACR2* in relation to the results of this study, two main differences are apparent. First, there is a less direct correspondence between the main chapter and section headings and the observed incidence rates of MARC fields in RDA than there is in *AACR2*. Second, the less-used material-specific rules are grouped together into chapters in *AACR2*, while in RDA they are not grouped together by material type but are spread around multiple chapters according to how they apply to the *FRBR* and *FRAD* entities. These differences may be difficult for novice catalogers to learn and for experienced catalogers to adapt to. The introduction to *AACR2* states, "The rules follow the sequence of cataloguers' operations in most present-day libraries and bibliographic agencies."[33] With the transition to RDA, this will no longer be the case, particularly in the early stages of the move. Catalogers, library administrators, and cataloging instructors all will have to adjust their practices and policies to the new rules.

This study has a number of potential extensions. Looking at the relationship between classification assignments and MARC field incidence might show whether materials from particular disciplines are cataloged more thoroughly than others. Similarly, analyzing records by the period in which they were cataloged might illustrate how field use rates have changed over time. Additionally, field incidence rates could be compared for different kinds of works (such as monographic, serial, cartographic, visual, etc.) or between digital and physical copies of the same works to estimate the relative cataloging workload that different materials might require. As the case study works in this analysis show, different kinds of materials likely will show different distributions of MARC incidence. Further, it might be useful to look in more detail at MARC record characteristics that this study did not examine, such as the prevalence of subfields and indicators, to provide a more nuanced characterization of the field incidence patterns. Finally, it will be interesting to revisit this study after libraries and other bibliographic organizations have transitioned to RDA, as successful application of the new rules will certainly affect the field incidence rates in future MARC records.

References

1. IFLA Study Group on the Functional Requirements for Bibliographic Records, *Functional Requirements for Bibliographic Records: Final Report* (Munich: K.G. Saur, 1998), www.ifla.org/VII/s13/*FRBR /FRBR*.pdf (accessed Aug. 10, 2009); Olivia M. A. Madison, "The Origins of the IFLA Study on Functional Requirements for Bibliographic Records," *Cataloging & Classification Quarterly* 39, no. 3/4 (2005): 15–34.
2. Karen Calhoun, *The Changing Nature of the Catalog and Its Integration with Other Discovery Tools* (final report, prepared for the LC, 2006), www.loc.gov/catdir/calhoun-report-final.pdf (accessed Sept. 26, 2009); Karen Markey, "The Online Library Catalog: Paradise Lost and Paradise Regained?" *D-Lib Magazine* 13, no. 1/2 (2007), www.dlib.org/dlib/january07/markey/01markey.html (accessed Aug. 10, 2009).
3. Martha Yee and Sara Shatford Layne, *Improving Online Public Access Catalogs* (Chicago: ALA, 1998).
4. Barbara B. Tillett, "Bibliographic Relationships: An Empirical Study of the LC Machine-Readable Records," *Library Resources & Technical Services* 36, no. 2 (1992): 162–88.
5. Richard P. Smiraglia, "Derivative Bibliographic Relationships: Linkages in the Bibliographic Universe," in *Navigating the Networks: Proceedings of the ASIS Mid-Year Meeting, Portland, Oregon, May 21–25, 1994*, ed. D. L. Andersen, T. J. Galvin, and M. D. Giguere (Medford, N.J.: Learned Information, 1994): 115–35.
6. Richard P. Smiraglia and Gregory H. Leazer, "Derivative Bibliographic Relationships: The Work Relationship in a Global Bibliographic Database," *Journal of the American Society for Information Science* 50, no. 6 (1999): 493–504.
7. Gregory H. Leazer and Richard P. Smiraglia, "Bibliographic Families in the Library Catalog: A Qualitative Analysis and Grounded Theory," *Library Resources & Technical Services* 43, no. 4 (1999): 191–212.
8. Zorana Ercegovac, "Multiple-Version Resources in Digital Libraries: Towards User-Centered Displays," *Journal of the American Society for Information Science and Technology* 57, no. 8 (2006): 1023–32.
9. Patrick Le Boeuf, "*FRBR*: Hype of Cure-All? Introduction," *Cataloging & Classification Quarterly* 39, no. 3/4 (2005): 1–13; Martha M. Yee, "*FRBR* and Moving Image Materials," in *Understanding FRBR : What It Is and How It Will Affect Our Retrieval Tools*, ed. Arlene G. Taylor (Westport, Conn.: Libraries Unlimited, 2007): 117–29.
10. Joint Steering Committee for Development of RDA, RDA: Resource Description and Access, Full draft of RDA (July 1, 2009), www.rda-jsc.org/rdafulldraft.html (accessed Sept. 26, 2009).
11. Chris Anderson, *The Long Tail: Why the Future of Business Is Selling Less of More* (New York: Hyperion, 2006).
12. Howard D. White and Katherine W. McCain, "Visualizing a Discipline: An Author Co-Citation Analysis of Information Science, 1972–1995," *Journal of the American Society for Information Science* 49, no. 4 (1998): 327–55.
13. George Kingsley Zipf, *Human Behaviour and the Principle of Least Effort* (Reading, Mass.: Addison-Wesley, 1949); A. J. should be Alfred James Lotka, "The Frequency Distribution of Scientific Productivity," *Journal of the Washington Academy of Sciences* 16, no. 12 (1926): 317–23.
14. Ellen Bonnevie, "A Multifaceted Portrait of a Library and Information Science Journal: The Case of the Journal

of Information Science," *Journal of Information Science* 29, no. 1 (2003): 11–23; Ravinder Nath and Wade M. Jackson, "Productivity of Management Information Systems Researchers: Does Lotka's Law Apply?" *Information Processing & Management* 27, no. 2/3 (1991): 203–9; Henry Voos, "Lotka and Information Science," *Journal of the American Society for Information Science* 25, no. 4 (1974): 270–72; Judit Bar-Ilan, "Informetrics at the Beginning of the 21st Century: A Review," *Journal of Informetrics* 2, no. 1 (2008): 1–52.

15. E. E. Fuller, "Variation in Personal Names in Works Represented in the Catalog," *Cataloging & Classification Quarterly* 9, no. 3 (1989): 75–96; Arlene Taylor-Dowell, *AACR2 Headings: A Five-Year Projection of Their Impact on Catalogs* (Littleton, Colo.: Libraries Unlimited, 1982); *Richard P. Smiraglia,* "The History of the 'Work' in the Modern Catalog," *Cataloging & Classification Quarterly* 35, no. 3/4 (2003): 553–67; Thierry Lafouge and Sylvie Laine-Cruzel, "A New Explanation of the Geometric Law in the Case of Library Circulation Data," *Information Processing & Management* 33, no. 4 (1997): 523–27; and Dietmar Wolfram, "Inter-Record Linkage Structure in a Hypertext Bibliographic Retrieval System," *Journal of the American Society for Information Science* 47, no. 10 (1996): 765–74.
16. David C. Blair, *Language and Representation in Information Retrieval* (Amsterdam: Elsevier, 1990).
17. M. E. J. Newman, "Power Laws, Pareto Distributions and Zipf's Law," *Contemporary Physics* 46, no. 5 (2005): 323–51.
18. Karen Markey and Karen Calhoun, "Unique Words Contributed by MARC Records with Summary and or Contents Notes," *Proceedings of the ASIS Annual Meeting* 24 (1987): 153–62.
19. Library of Congress, Network Development and MARC Standards Office, Structure of the LC Control Number, www.loc.gov/marc/lccn_structure.html (accessed Aug. 10, 2009).
20. Ronald Rousseau and Sandra Rousseau, "Informetric Distributions: A Tutorial Review," *Canadian Journal of Information and Library Science* 18, no. 2 (1993): 51–63.
21. Leo Egghe, *Power Laws in the Information Production Process: Lotkaian Informetrics* (Amsterdam; New York: Elsevier/Academic Pr., 2005), 387.
22. Ibid.
23. Jean Tague, "What's the Use of Bibliometrics?" in *Informetrics 87/88*, ed. Leo Egghe and Ronald Rousseau (New York: Elsevier, 1988): 271–78.
24. Pat Riva, "Mapping MARC 21 Linking Entry Fields to *FRBR* and Tillett's Taxonomy of Bibliographic Relationships," *Library Resources & Technical Services* 48, no. 2 (2004): 130–43.
25. Newman, "Power Laws, Pareto Distributions and Zipf's Law."
26. Charles Ammi Cutter, *Rules for a Dictionary Catalog*, 4th ed. (Washington, D.C.: GPO, 1904): 12.
27. Ibid.
28. Michael Gorman, "Seymour Lubetzky, Man of Principles," in *The Future of Cataloging: Insights from the Lubetzky Symposium*, ed. Tschera Harkness Connell and Robert L. Maxwell (Chicago: ALA, 2000): 12–21.
29. *Anglo-American Cataloging Rules* (Chicago: ALA; London: Library Assn., 1967); *Anglo-American Cataloguing Rules*, 2nd ed., 2002 rev. (Ottawa: Canadian Library Assn.; London: Library Assn. Publishing; Chicago: ALA, 2002).
30. Joint Steering Committee for Development of RDA, RDA: Resource Description and Access, Full draft of RDA.
31. IFLA Working Group on Functional Requirements and Numbering of Authority Records (FRANAR), *Functional Requirements for Authority Data: A Conceptual Mode* (Munich: K.G. Saur, 2009).
32. Trond Aalberg, "A Process and Tool for the Conversion of MARC Records to a Normalized *FRBR* Implementation," *Lecture Notes in Computer Science* 4312 (2006): 283–92; Thomas B. Hickey and Edward T. O'Neill, "*FRBR*-izing OCLC's WorldCat," *Cataloging & Classification Quarterly* 39, no. 3/4 (2005): 239–51.
33. *Anglo-American Cataloguing Rules*, 2nd ed., 1.

Appendix A. The Most Frequently Incident Fields in the Library of Congress Records

Rank by Frequency of Incidence	MARC Field	Field Description	Number of Records that Contained a Given Field	% of Records with Field
1	906	Local Processing Field	1500	100.00
-	001	Control Number	1500	100.00
-	008	Date Entered on File	1500	100.00
-	010	LCCN	1500	100.00
5	005	Date and Time of Latest Transaction	1499	99.93
-	040	Cataloging Source	1499	99.93
7	245	Title Statement	1498	99.87
8	260	Publication, Distribution, etc.	1496	99.73
9	300	Physical Description	1495	99.67

Appendix A. The Most Frequently Incident Fields in the Library of Congress Records (cont.)

Rank by Frequency of Incidence	MARC Field	Field Description	Number of Records that Contained a Given Field	% of Records with Field
10	050	LC Call Number	1487	99.13
11	035	System Control Number	1381	92.07
12	991	Preprocessing Location/ Conversion	1242	82.80
13	100	Main Entry—Personal Name	1109	73.93
14	650	Subject Added Entry—Topical Term	990	66.00
15	500	General Note	507	33.80
16	985	Record Source or Project History	506	33.73
17	043	Geographic Area Code	501	33.40
18	042	Authentication Code	489	32.60
19	082	Dewey Decimal Classification Number	433	28.87
20	700	Added Entry—Personal Name	401	26.73
21	955	Local Tracking Field	400	26.67
22	020	International Standard Book Number	388	25.87
23	504	Bibliography, etc. Note	372	24.80

Appendix B. Highest Ranked MARC Fields for the Library of Congress Records by the Frequency of Occurrence, Normalized by the Number of Records in Which They Were Incident

Rank by Frequency of Occurrence	MARC Field	Field Description	Number of Occurrences per Record Observed to Have That Field	Total Field Occurrences
1	880	Alternate Graphic Representation	4.05	174
2	074	GPO Item Number	2.00	4
3	249	Local Title Field	1.88	30
4	650	Subject Added Entry—Topical Term	1.84	1817
5	592	Local Notes	1.83	11
6	856	Electronic Location and Access	1.61	100
7	510	Citation/References Note	1.56	14
8	655	Index Term—Genre/Form	1.55	65
9	538	System Details Note	1.50	3
10	500	General Note	1.45	735
11	700	Added Entry—Personal Name	1.42	569

Appendix B. Highest Ranked MARC Fields for the Library of Congress Records by the Frequency of Occurrence, Normalized by the Number of Records in Which They Were Incident (cont.)

Rank by Frequency of Occurrence	MARC Field	Field Description	Number of Occurrences per Record Observed to Have That Field	Total Field Occurrences
-	740	Added Entry—Uncontrolled Related/Analytical Title	1.42	105
13	651	Subject Added Entry—Geographic Name	1.38	384
14	246	Varying Form of Title	1.32	98
15	035	System Control Number	1.30	1,799
16	052	Geographic Classification	1.25	15
17	952	Cataloger's Notes	1.23	53
18	600	Subject Added Entry—Personal Name	1.22	200
19	016	National Bibliographic Agency Control Number	1.20	6
-	710	Added Entry—Corporate Name	1.20	254

Appendix C. MARC Fields with Incidence Rates Greater than 20 Percent for the Case Studies

	Lord of the Files (*n* = 98)	
MARC Field	Field Description	Incidence Rate (%)
001	Control Number	100
008	Date Entered on File	100
245	Title Statement	100
260	Publication, Distribution, etc.	100
300	Physical Description	98
040	Cataloguing Source	96
035	System Control Number	71
100	Main Entry—Personal Name	71
852	Shelving Location	67
020	International Standard Book Number	66
500	General Note	62
700	Added Entry—Personal Name	62
650	Subject Added Entry—Topical Term	46
082	Dewey Decimal Classification Number	32
600	Subject Added Entry—Personal Name	32
050	Library of Congress Call Number	30
440	Series Statement/Added Entry—Title	28
250	Edition Statement	24

Appendix C. MARC Fields with Incidence Rates Greater than 20 Percent for the Case Studies (cont.)

Lord of the Files (*n* = 98)		
MARC Field	**Field Description**	**Incidence Rate (%)**
511	Participant or Performer Note	24
007	Physical Description Fixed Field	23
520	Summary Note	22
010	Library of Congress Control Number	21
710	Added Entry— Corporate Name	20

Republic (*n* = 100)		
MARC Field	**Field Description**	**Incidence Rate (%)**
000	Local UCLA Field—Type of Record	100
001	Control Number	100
005	Date and Time of Latest Transaction	100
008	Date Entered on File	100
035	System Control Number	100
245	Title Statement	100
260	Publication Information	100
300	Physical Description	100
910	Local UCLA Field—Cataloger and Date	100
040	Cataloguing Source	95
100	Main Entry—Personal Name	95
700	Added Entry —Personal Name	88
935	Local UCLA field—ID from Old System	81
650	Subject Added Entry—Topical Term	60
010	Library of Congress Control Number	57
041	Language Code	54
050	Library of Congress Call Number	51
500	General Note	50
240	Uniform Title	43
504	Bibliography, etc. Note	40
020	International Standard Book Number	37
490	Series Statement	32
082	Dewey Decimal Classification Number	27
505	Formatted Contents Note	23