From the Ubiquitous to the Nonexistent

A Demographic Study of OCLC WorldCat

Jay H. Bernstein

Analysis of a random sample of bibliographic records from OCLC WorldCat finds that the great majority of items in WorldCat are held by very few participating libraries, and that an inverse geometric relationship exists between the number of libraries holding an item and the number of items with a given level of shared holdings. The findings provide a context for interpreting holding levels in WorldCat with regard to the proportion of widely shared items and the characteristics of items at various ranges of holdings. Used with other quantitative and evaluative measures, these findings will assist libraries in assessing their collections.

OCLC WorldCat is arguably one of the most valuable tools available to librarians, as it provides online, global access to the shared records of thousands of libraries around the world. This enables librarians, at a minimum, to verify the existence of an item mentioned by a patron and to provide correct bibliographical facts of publication for it. More generally, it helps librarians identify and locate items that may not be held in their own collections.

WorldCat is also a potentially powerful research tool for collection analysis because each of its bibliographic records indicates the number of member libraries holding that item. These data are tantalizingly provocative to scholars interested in patterns in the distribution of books, knowledge, and information.

The theoretical basis for considering WorldCat holding levels in library research is that the number of libraries holding a given title provides a score that can measure that title's influence or impact. Several studies conducted over the years on specific library material domains—adult fiction, books from small publishers, scientific journals, and award-winning monographic titles in the social sciences and humanities—have cited figures on holdings in WorldCat or its predecessor, the OCLC Online Union Catalog.[1] Similarly, Mirwis has included WorldCat holdings as one of several factors used to rate and rank encyclopedias.[2]

The difficulty with citing these holdings data lies in knowing how to interpret them. Without a clear picture of what exactly is in WorldCat and what the spectrum of distribution is, the raw holdings numbers mean little. Only by examining a sample of items from WorldCat itself, rather than beginning with particular

Jay H. Bernstein (jbernstein@kingsborough.edu) is Assistant Professor and Reader Services Librarian in the Robert J. Kibbee Library, Kingsborough Community College, City University of New York.

The author thanks Jennifer Coggon, Edward Martin, Allan Mirwis, Roberta Pike, Elizabeth Tompkins, and Stephen E. Wiberley Jr. for their helpful comments on earlier drafts of this paper, and Anna H. Perrault for her advice and encouragement.

known items or a particular library collection, can one develop an interpretive framework that could maximize the analytical value of these data.

In the interest of providing a context for interpreting WorldCat holding levels, I undertook what might be called a demographic study of WorldCat, focusing on its composition and major characteristics as a population by analyzing a random sample of 500 records. The goal in investigating the items of WorldCat as representatives of a population was to determine the categories of materials that are represented in WorldCat and their proportions, the proportion of widely held items, and the characteristics of those items.

Precedents for this study may be found in publications by White and Perrault.[3] White demonstrated the connection between Research Libraries Group (RLG) Conspectus levels and OCLC holding levels. He found that titles at the research level had holding levels up to 150, and that the midpoint of holdings of titles in library collections at all levels was about 400.[4] Perrault studied a systematic random sample of nearly 3.4 million WorldCat records to gauge trends in library collection building over time and assess WorldCat's potential for promoting shared access to scarce resources.[5] While her study did not concern levels of shared holding in a manner comparable to White's work, she did uncover a finding that is striking when considered alongside White's results: more than 53 percent of the records in Perrault's sample were held by only one member library.[6]

The present study attempts to show the big picture of what is in WorldCat by considering its widely shared items against the background of a much larger number of extremely scarce items. Like Perrault's study, it employs an inductive approach to assess WorldCat as a whole, though it is based on a sample far smaller than hers. Like White's work, it relates its findings to those items in WorldCat that are frequently found in library collections. This study finds that 39.8 percent of items in WorldCat are held by only one library. This is significantly less than the 53.1 percent found by Perrault, though it is still a very considerable fraction. The study also finds that only 9.1 percent of items are held by more than 50 participating libraries. Perhaps most surprising, the study shows that 1.2 percent of items in WorldCat are not held by *any* libraries.

OCLC WorldCat in a Nutshell

Provided by OCLC Online Computer Library Center, WorldCat is an online bibliographic database containing records created and maintained collectively by more than 9,000 member institutions (the exact number is unavailable). Described by OCLC as "the largest and most comprehensive database of its kind," its resources "span thousands of years and nearly every form of human expression. Records exist for everything from stone tablets to electronic books, wax recordings to MP3s, DVDs and Web sites."[7] A million records were added in the first three-and-a-half months of 2004, bringing the total inventory to 55 million. On average, WorldCat adds a new record every 12 seconds, a phenomenally rapid rate of growth.[8] The database has doubled in size in just more than eleven years.[9]

Not long after launching its Online Union Catalog in 1971, OCLC set forth initiatives throughout the world promoting input from other countries that would make its database global. In doing so, OCLC had to incorporate international variations in MARC format, authority rules, record quality, and library traditions, not to mention non-Roman alphabets and characters.[10]

The many libraries around the world contributing to WorldCat make it an unsurpassed data bank for research on the contents of libraries. It is probably a good measure of library holdings, especially in the English-speaking world. However, this does not necessarily mean that it accurately reflects the totality of world library holdings, much less the entire store of recorded human knowledge.

Most institutions that contribute to WorldCat are governing members of OCLC, meaning that they contribute all current cataloging to an OCLC-affiliated database. Two lower grades of OCLC membership involve less-than-full participation in WorldCat. More than 77 percent of governing members are located in the United States. WorldCat seems to represent academic libraries more so than other kinds of library. According to data in OCLC's 2003 annual report, 31 percent of governing member libraries are college and university libraries, compared to 22.2 percent of governing members that are public libraries, 12.8 percent that are federal, state, or municipal government libraries, and 10.2 percent that are corporate or business libraries.[11] Other categories of libraries making up OCLC's governing membership are community college and vocational (8.3 percent), school (7.4 percent), associations and foundations (4.0 percent), state and national (1.0 percent), and other (3.0 percent).[12]

Data Collection Procedures

The first step in data collection was to select 500 random numbers from 1 to 54 million using an online random number generating service.[13] The maximum number of 54 million was chosen because the database contained "54 million quality records and counting" at the time research for the study began in July 2004, according to OCLC's WorldCat home page.[14] Each randomly chosen number obtained using this method was matched with the bibliographic record for the item with that OCLC record number. Given the size of the chosen population, the sample of 500 provides a confi-

dence level of 95 percent and a confidence interval of 4.4. Therefore, conclusions about this sample are expected to be accurate plus or minus 4.4 percent, with 95 percent certainty, as statements about the entire WorldCat database. Such a sample is adequate for drawing general conclusions about the major outlines and proportions of items in WorldCat. It is not sufficiently large enough to include many outstanding but extremely rare types of items, such as incunabula, paleographic writing boards, or items in hieroglyphics.

Findings

Types of Documents and Materials

In order to comprehend WorldCat as a database, one must sort out the various kinds of items it contains. The *Anglo-American Cataloguing Rules,* 2nd edition, 2002 revision, differentiates in its table of contents the following categories of items for cataloging: books, pamphlets, and printed sheets; cartographic materials; manuscripts (including manuscript collections); music [printed notated music]; sound recordings; motion pictures and videorecordings; graphic materials; electronic resources; three-dimensional artifacts and realia; microforms; continuing resources [serials]; and analysis [monographic or journal analytics].[15]

Such a synopsis is typical. However, it is inadequate as a categorical scheme, as the individual categories are not mutually exclusive. Books, for example, can be manifested not only on printed paper but also in sound recordings, microforms, and electronic resources. Indeed, most document types can appear in a number of material formats. To avoid confusion in collection analysis, a system that creates nonoverlapping categories by distinguishing explicitly between document type and material type must be established.

The use of the terms *document type* and *material type* is intended to emphasize a sharp conceptual distinction that is absent from most discussions of the categories of items in catalogs and bibliographic databases. Without this distinction, any discussion of the categories of items in libraries and catalogs is bound to be muddled.

Document type is primary, and material type is subsidiary. Document type sums up the general, overall character of a work in a basic, salient, overriding category. The major document types, such as book, serial, manuscript, sound recording, and motion picture, are sociocultural categories characterized by prototypical cores but fuzzy and ambiguous boundaries.[16] Material type denotes not only the physical medium (e.g., paper, magnetic tape, celluloid film), but also the mode of communication—the medium in which the content is presented. Thus, printed language material, printed cartographic material, and printed musical material are distinct material types. Similarly, atlases and scores on microfilm are separate material type categories from microfilm books. Material type differentiates among various specific microformats (e.g., microfilm, microfiche, microopaque); among formats for reproducing sound, visual, or graphical data; and among various electronic formats and means of access to computer files (Web, file transfer protocol, Usenet, diskette, CD-ROM, and so on.).

Analyzing the makeup of OCLC WorldCat in terms of document type and material type involves both the aggregation and separation of categories. Thirteen document types and 22 material types may be identified in the sample. Combining document types and material types, bibliographic records from the sample of 500 can be placed into 30 mutually exclusive categories, plus a null category consisting of 5 numbers matching no records. Items are by no means distributed evenly among categories, but are heavily clustered in a few categories, with a few scattered entries among all the other categories (see table 1). Printed language books (hereafter print books) alone account for more than two-thirds of all items.

For both document type and material type, the predominating type is responsible for more than 70 percent of records and is about ten times more prevalent than the second most common type. The 2 leading types in both categories are closely related to each other: the leading document

Table 1. Document types and material types in WorldCat sample (n=495)

	No.	%
Document type		
Books	364	73.5
Manuscripts (text)	40	8.1
Serials	25	5.1
Scores	15	3.0
Musical recordings	12	2.4
All others	39	7.9
Material type		
Printed language	355	71.7
Original language	36	7.3
Microfilm language	21	4.2
Printed notated music	14	2.8
Microfiche language	10	2.0
All others	59	11.9
Combined document: material type		
Books: Printed language	335	67.6
Manuscripts: Original language	33	6.7
Serials: Printed language	18	3.6
Books: Microfilm language	15	3.0
Scores: Printed notated music	14	2.8
All others	80	16.2

type is the book, and the leading material type is printed language material, while in second place are manuscript and original language material respectively. The same ratio holds for the 2 leading combined document-material types: there are ten times as many print books as original language material manuscripts. Together, these 2 kinds of items account for almost 75 percent of all items in the database.

Categories of Holdings Levels

This study aims to determine the proportion of items in the catalog that are widely held and to analyze the characteristics of these widely held items. Determining a cut-off point for high holdings to define a category of widely held items is an arbitrary procedure that can lead to circular reasoning. One cannot know in advance what a high level of holdings is. Because this paper is a first attempt to determine the occurrence of widely held items in context of the totality of merged online catalogs, predefining the category seems unwise. It is somewhat easier on an intuitive level to grasp the meaning of *scarce* items than *common* items (though here too the cut-off point is arbitrary). Therefore, rather than attempt to provide a parameter for widely held items, I will set these items aside for the time being by dividing the sample into four categories:

- Nonexistent: items with 0 holdings
- Unique: items with 1 holding
- Scarce: items with 2 to 50 holdings
- Non-scarce: items with more than 50 holdings

In this typology, non-scarce is a residual category and will be addressed later in this paper.

The reader may be surprised at these categories, particularly the first. However, the typology arises from the data themselves, so the distribution of holding levels itself is the real surprise:

- Nonexistent items: 6 (1.2 percent)
- Unique items: 197 (39.8 percent)
- Scarce items: 247 (49.9 percent)
- Non-scarce items: 45 (9.1 percent)

Non-scarce items account for a relatively small fraction of the sample, and this finding justifies not subdividing the non-scarce portion into thinner layers at the outset.

These distributions vary by both document type and material type. Table 2 shows the distribution for all items and for the top five document types, which make up more than 92 percent of the sample. Government publications, which account for 7.3 percent of the sample, break down in approximately the same proportions as other items.

Based on the sample, 1.2 percent of items cataloged in WorldCat have absolutely no holdings; they are not held by the libraries that had cataloged the items and input the records. These appear to be items that once were held but that were deaccessioned, as none of the records are based on prepublication data or unexamined material. However, the bibliographic records that should have been deleted or reported as errors persist, and the items must be considered nonexistent in terms of library collections. For 5 of the 6 items with 0 holdings, other records with holdings for items with the same title are present; all items are extremely scarce, except for one that is cataloged in the sample as a book but that is held as a microform serial by 17 libraries. In addition to the 6 records that have 0 holdings, 5 numbers used to obtain sample records do not match any records. The percentages given here are percentages of the sample of 495 records, not of all the 500 random numbers used to generate the sample, unless otherwise noted.

In this paper, *unique* means that only one library that is a participating member of OCLC owns the item. *Unique* does not necessarily mean that no other library in the world owns the item, much less that it is an absolutely unique document. In analyzing the percent of unique items, the differentiation of original or archival materials, which one would expect to be unique from other categories of materials is advantageous. Original and archival materials in the sample include language and music manuscripts, mixed material collections, and a graphic collection. Forty-four items in the sample are archival, and 41 of these are unique. All but one of the language manuscripts are academic theses, and the exception appears to be an incorrectly cataloged book. The great majority of theses are for degrees from United States institutions and are in English. Non-archival materials account for 451 items in the sample and, of these, 156 (34.6 percent) are unique. Archival materials account for only 8.9 percent of all materials in the sample but 20.9 percent of the

Table 2. Distribution of holding levels categories by document type (n=495)

Document type	Nonexistent	Unique	Scarce	Non-scarce	Total
Books	4	118	200	42	364
Manuscripts	0	36	4	0	40
Serials	0	11	13	1	25
Scores	0	7	8	0	15
Musical recordings	1	3	7	1	12
All others	1	22	15	1	39
Total	6	197	247	45	495

unique content. Besides archival materials, the category of unique items includes many of the items outside the print book category as well as many technical reports and items of only local significance.

Table 3 shows the distribution of frequency by material type categories. It contrasts print language material, which is the predominant category, making up almost three-quarters of the total, against all archival materials and the residual category of all other materials, which includes microforms, computer files, maps, sound recordings, and so on.

Unique items are more likely than items in other holding level categories to be short in length. Of unique print books, 36.2 percent have fewer than 50 pages (or leaves), and 40.9 percent are more than 100 pages long. This compares to 27.2 percent and 56.1 percent, respectively, in all holding level categories. Table 4 shows the distribution of print books in each holding level category according to page length.

In a study of nearly 3.4 million items in WorldCat, Perrault found that 53.1 percent of all records for *monographs* were unique.[17] Presumably, Perrault's *monographs* are identical to my *print books,* though the term also could refer to anything not cataloged as a serial. Only 39.8 percent of all items in the present research sample, including 31.3 percent of print books, are unique.

Scarce is defined as having greater than 1 but no more than 50 holdings; 49.9 percent of items are in the scarce category. Although a few items cataloged as archival materials have multiple holdings in very low numbers and thus may be considered scarce, most scarce items are non-archival.

Holding levels are strongly correlated with the presence of call numbers. One hundred fifty-eight records (31.9 percent of the total) have no call numbers of any kind, but only 2 are non-scarce. Of unique items, 46.2 percent lack call numbers (very close to the 47.6 percent of unique monographic titles with no call number found by Perrault), and for scarce items, the figure is 25.5 percent.[18] Half of all items have a Library of Congress Classification (LCC) number assigned either locally or by the Library of Congress (LC). Adding National Library of Canada numbers, 52.5 percent of items have at least one LCC-type call number. Dewey Decimal Classification (DDC) numbers, either LC-assigned or locally assigned, are present in 30.7 percent of items. For all items, the ratio of LC-assigned to locally assigned LCC numbers is approximately 1 : 1, but for DDC numbers the ratio is greater than 2 : 1 (see table 5). For unique items, however, locally assigned numbers greatly outnumber centrally assigned numbers in both classification systems.

Whereas scarce items frequently have incomplete cataloging, the records for non-scarce items tend to be the fullest, most complete records. They are far more likely than others to have been cataloged at the national level, meaning they have been both cataloged and transcribed by national bibliographic agencies, as indicated in the MARC 040 field (cataloging source) by codes for national cataloging agencies—the Library of Congress (DLC), the British Library (UKM), the National Library of Canada (NLC), the National Library of Medicine (NLM), and others. For example, an item cataloged and transcribed by LC would contain the code "DLC ǂc DLC" in the MARC 040 field. Compared to 18.6 percent of scarce titles and just 7.7 percent of unique titles, 58.7 percent of non-scarce titles are so cataloged. For titles with holdings of more than 100, the proportion is 81.8 percent. Excluding archival materials, 20 percent of records are cataloged and transcribed by national bibliographic agencies. Print books account for 90.9 percent of items cataloged at the national level. Table 5 shows that call numbers for non-scarce items are far more often created by LC than locally; 78 percent of items have LC-created numbers (LCC or DDC), as compared to 9 percent with locally created numbers. Non-scarce items often have specialized call numbers in addition to ordinary call numbers. Reflecting the dominance of Anglo-American institutions in WorldCat participation, Universal Decimal

Table 3. Distribution of holding levels categories by material type (n=495)

Material type	Nonexistent	Unique	Scarce	Non-scarce	Total
Printed language	4	118	197	40	359
Archival	0	41	3	0	44
Other	2	38	47	5	92
Total	6	197	247	45	495

Table 4. Length in pages of print books categorized by holdings level category

Holdings level	1–50 pages	51–100 pages	More than 100 pages	Unknown page length	Total
Nonexistent	2	0	1	1	4
Unique	38	14	43	10	105
Scarce	50	21	109	7	187
Non-scarce	1	2	35	1	39
Total	91	37	188	19	335

Table 5. Distribution of types of call numbers in frequency categories

	No call number		050 (LCC)	055 (Can. LCC)	060 (NLM)	070 (NAL)	080 (UDC)	082 (DDC)	084 (other)	086 (Gov. doc.)	090 (local LCC)	092 (local DDC)
	No.	%										
Nonexistent	2	33	0	0	0	0	0	0	0	1	2	1
Unique	91	46	18	2	3	0	1	8	3	1	57	18
Scarce	63	26	70	9	3	0	0	63	0	7	64	23
Non-scarce	2	4	35	1	3	2	0	35	0	4	4	4

Note: Numbers in top row refer to MARC fields. Not all categories are mutually exclusive.

Classification (MARC 080 field) or "other" (MARC 084 field) numbers are found only in unique items.

Non-scarce items make up only 9.1 percent of the total; 93 percent of these are in the book document type, and 89 percent are print books. Besides print books, non-scarce items include 2 microfilm reproductions of original books dating from 1665 and 1894, a map, a musical recording on a compact disc, and an online journal. Considering that these document and material types occur far less frequently than print books, one cannot conclude from the low numbers that items of these kinds are scarce in greater proportions than are print books.

Countries and Regions of Origin

Given OCLC's stated objective of the global networking of information, United States dominance as the place of origin of items in WorldCat (despite the representation of many countries in all regions of the world) is noteworthy.[19] Items originate in 54 countries, an impressive number for a sample this size. Twenty-four countries are represented by one item apiece, and 19 countries by between 2 and 5 items. The United States alone is responsible for 195 items, or 39.4 percent.

In second place, with 60 items, or 12.1 percent, are materials with no place or an unknown or undetermined place of origin, as indicated by "Ctry: xx" in the fixed field of the MARC record. Most of these are archival or unpublished materials, especially manuscripts, which as a rule do not name a country of origin. However, other items for which the place of publication or origin is not recorded or cannot be determined are coded the same way. Apart from manuscripts (language and music), 4.2 percent of records do not name a country of origin.

Limiting the materials to print books, 52 countries are represented in all (see figure 1). Twenty-three of these are responsible for only 1 book each, and another 16 are responsible for between 2 and 5 books each. The United States, England, and 8 other countries (mostly in Western Europe) are responsible for more than three-quarters of all print books in the sample. In the absence of manuscripts, archival materials, and computer files, only 3.0 percent have no known or determined country of origin, as compared to 30.6 percent in all other categories.

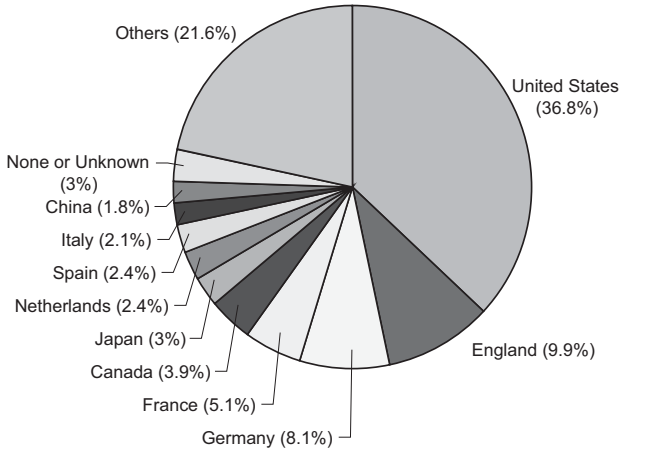

Figure 1. Distribution of print books in WorldCat sample by country of publication (n=335)

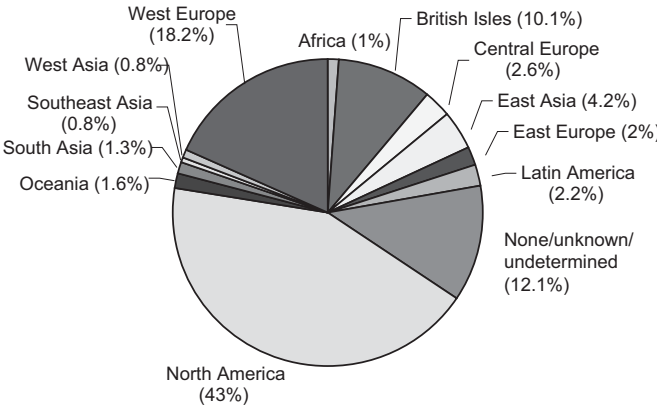

Figure 2. Distribution of items in WorldCat sample by region of origin (n=495)

Looking at WorldCat in terms of regional representation (see figure 2), North America (defined as the United States and Canada), Western Europe, and the British Isles (including the Republic of Ireland) are dominant, while Africa and all of Asia except for the Far East are only very slightly represented.

Microfilm and other reproductions usually state the country of origin of the original document, not the copy, at least in the fixed field of the MARC record, which is the preferred source of information on countries for this study. Bearing in mind this proviso, 31 of 45 non-scarce items (69 percent) originated in the United States, followed by eight (18 percent) from England; the remaining items (1 each) originated in Canada, France, Germany, Japan, Spain, and Scotland. All nonexistent items originated the United States.

Languages

Official information on the number of items in various languages in WorldCat, as of July 30, 2004, is available on OCLC's Internet page, "WorldCat facts and statistics."[20] Based on figures for the total number of items in the database presented therein, one can calculate the fraction of items in each language. Comparison of these statistics to those for the top 9 languages in the research sample (see table 6) lends assurance that the sample is representative of the whole, though noticeable differences exist concerning the proportions of items in French and Japanese. The reader is referred to the Web site for figures for the top 53 languages (Serbo-Croatian is counted twice, differentiating between Romanic and Cyrillic forms). The sample of 495 includes items in 31 languages (including both forms of Serbo-Croatian) and 2 items in multiple languages. Three percent of the sample, chiefly musical scores and recordings, are nonlingual. Of the 27 items in those categories, 13 (48 percent) have no language.

The last column of table 6 provides percentages for print books. It shows that the percentage of print books in English is significantly smaller than the percentage of all items in English while the percentages in French and German are considerably higher for print books than for all formats. This appears to be the effect of the large number of records contributed by academic libraries in the United States and other Anglophone countries for academic theses that are overwhelmingly in English and excluded from the print book category as these are cataloged as manuscripts.

Intensity of language representation and holding levels are strongly correlated, with the most frequently occurring languages having the highest holdings. The 305 English-language items can be divided into the frequency categories as follows: 6 (2 percent) nonexistent, 127 (42 percent) unique, 130 (43 percent) scarce, and 42 (14 percent) non-scarce. By comparison, of 190 items not in English, 69 (36 percent) are unique, 117 (62 percent) are scarce, and only 4 (2 percent) are non-scarce, including titles in French, German, and Spanish, as well as one nonlingual musical recording.

Material Age

To analyze the age structure of WorldCat, all serial or continuing items as well as undated items are removed from the sample, resulting in 460 items. The sample is divided by century, further breaking down the twentieth century into time periods for the first half, third quarter, and fourth quarter. Separate categories are created for items that have been reproduced in periods different from the original period of publication. The results appear in table 7. The period of 1976 through 2000 is the largest category, with 208 items, which is 45.2 percent of dated, non-serial and non-continuing items in the sample. This time period accounts for 69.6 percent of non-scarce items.

Holdings Distribution

The simplest description of the holdings distribution of sample records is that the greater the number of holdings,

Table 6. Representation of languages in WorldCat and in research sample by percentage

Language	OCLC statistics (%)	Research sample (%)	Print books (%)
English	61.2	61.6	56.7
French	6.4	7.5	8.7
German	6.3	5.7	8.4
Spanish	4.6	4.2	4.2
Japanese	2.7	1.8	2.7
Chinese	2.3	2.0	3.0
Russian	1.9	1.6	1.8
Italian	1.8	1.4	1.8
Latin	0.9	1.2	1.2

Table 7. Date of publication of non-serial and non-continuing items (n=460)

Year of Publication	Originals	Reproductions
1401–1600	0	3
1601–1700	2	1
1701–1800	7	1
1801–1900	37	7
1901–1950	54	6
1951–1975	104	2
1976–2000	208	0
2001–2003	28	0

the fewer the number of items having that many holdings. There are many records with low holdings, with the greatest number having a single holding, and decreasing numbers of records with higher numbers of holdings, followed by a few records with a large number of holdings. The first part of the curve can be seen as a reverse J-shaped curve with a long tail, as shown in figure 3, which provides the frequency of records with 1 to 40 holdings. But to say there is a long tail is insufficient, as this does not begin to suggest the relatively small but significant number of records that have much higher levels of holding.

Although any cut-off point is arbitrary, one may visualize the sample in two sections, as in figures 4 and 5. The first group is the large number of items with few holdings, and the second is the group with few members, but with elevated numbers of holdings. Two-thirds of all the records (328 out of 495) have 0 to 5 holdings, while the remaining one-third have more than 5 holdings. These categories are analyzed separately.

For the first group, one can apply a modified version of Lotka's law, which states that "the number of authors who have published a specific number of papers [is] approximately equal to the inverse square of that number multiplied by the number of authors who had published one paper only, that is, [if one sets] f (y) as the number of authors publishing y papers, . . . [then] $f(y) \approx 1/y^2 \times f(1)$."[21] One can adapt this law by substituting records for authors and holdings for papers. Given the size of the sample, this formula provides a reasonably close approximation for the fraction of items with 2 to 5 holdings (see figure 6).

For the second group, with elevated numbers of holdings, a different form of analysis is necessary. Here, one may divide the group into zones containing equal numbers of records to see what the range of holding levels are for each zone. Keeping in mind the geometric growth seen in the curve in figure 5, and following Price's law, which states that "the number of prolific authors in a subject area (i.e., producing about half the publications in the field) [is] approximately equal to the square root of the total number of authors," a zone is defined as a group containing a number of items equal to the square root of the entire sample (500), which is slightly more than 22.[22] Not all zones contain exactly 22 items, as the size is adjusted to fit all items with the same number of holdings. In all, there are 8 zones in all of the items with holdings of more than four, as seen in table 8. Unexpectedly, zone 1, the highest zone, defined as the zone with the 22 items with the highest level of holdings, consists of items (all print books) with more than 100 hold-

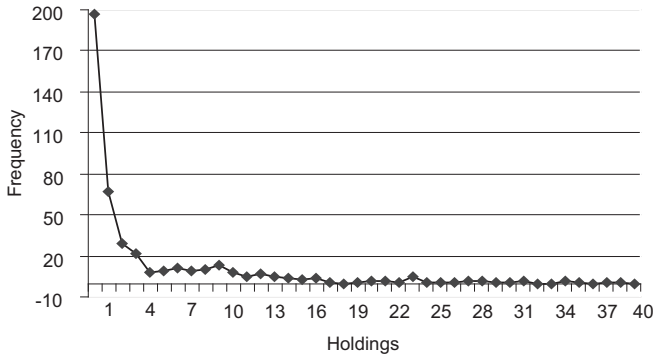

Figure 3. Frequency of records with 0 to 40 holdings (n=445)

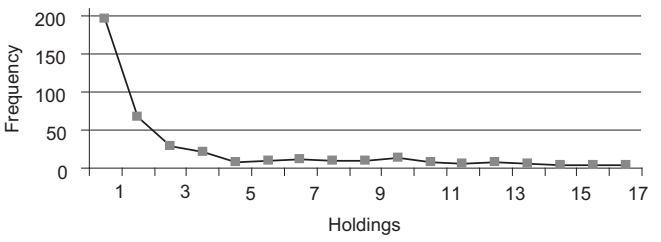

Figure 4. Frequency of records with 0 to 17 holdings (n=417)

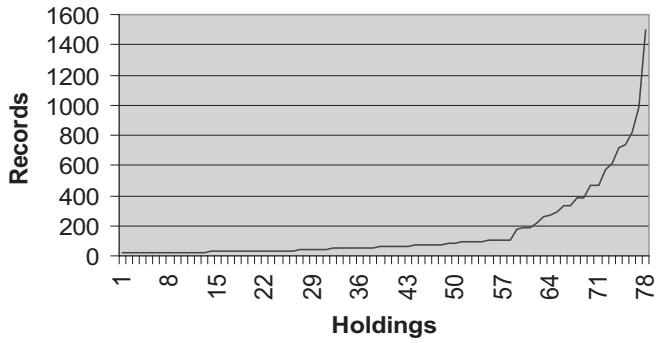

Figure 5. Occurrence of records with 18 or more holdings (n=78)

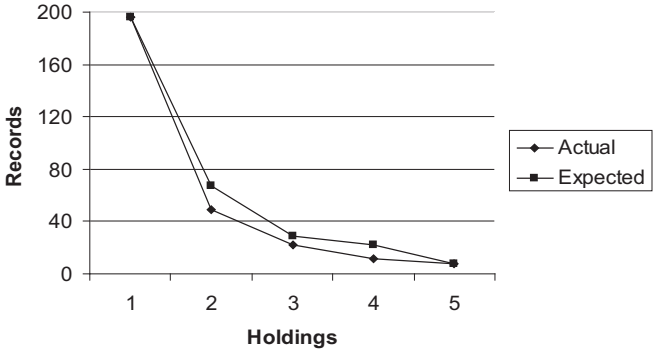

Figure 6. Distribution of items with 1 to 5 holdings (n=287)

ings. Zone 2 contains items with 52 to 100 holdings; zone 3 contains items with 25 to 51 holdings; and so on. In all zones, the majority of items are print books.

Table 8 shows the range of holdings levels in each zone, a range that becomes wider as the number of items increases. This analysis of zones helps segment the large and heretofore undifferentiated category of scarce items.

The number of widely shared items in OCLC WorldCat as a whole cannot be estimated from the square root of total number of items in the database. Given a database size of 54 million, this would mean that only 7,550 items have more than 100 holdings, an absurdly small number. On the contrary, the fraction of items in the sample with more than 100 holdings, 4.4 percent, is a much more reliable figure for projecting the number of items with more than 100. Using this percentage, one can estimate 2,376,000 records with more than 100 in a database of 54 million. The modified equivalent of Price's law is used simply to suggest that the number of items with more than 100 holdings is approximately the same as the number with 5 to 6 holdings, 7 to 8 holdings, and so on, up to about 52 to 100 holdings. The highest zone is limited only by the number of participating libraries; the second-highest zone has an interval of 49, and the third-highest zone has an interval of 27. As zones encompass materials with higher holdings, intervals of ranges grow exponentially. Excluding the first zone, whose wide range dwarfs all other zones, figure 7 graphs the internal sizes for 8 to 2.

Table 8. Zones of holding levels

Zone	Range	Interval	Total	Print books	Other formats
8	5–6	2	17	10	7
7	7–8	2	17	15	2
6	9–10	2	23	16	7
5	11–14	4	25	20	5
4	15–25	11	22	15	7
3	25–51	27	22	19	3
2	52–100	49	22	17	5
1	101–1,491	1,391	22	22	0

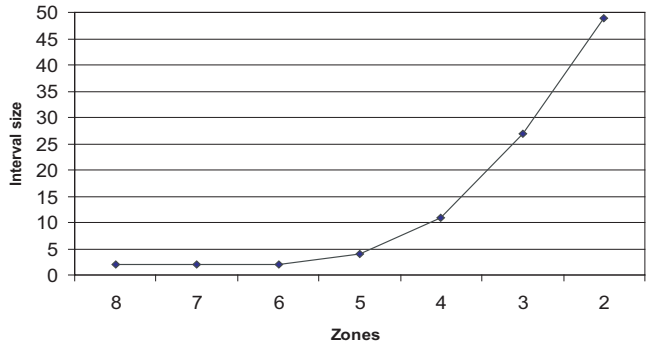

Figure 7. Increasing interval sizes of zones 8 to 2

Characteristics of Widely Held Titles

Forty-five items in the sample (9.1 percent) are non-scarce, with more than 50 holdings each. However, the previous analysis shows that this category constitutes quite neatly the top two zones, with a residue of just 1 item in the third zone. At this point one can focus on the highest zone, which consists entirely of print books in English, dividing it even further into two equal halves to arrive at a core consisting of the most ubiquitous items. Dividing the top zone into two halves of 11 items each results in group A, with 381 to 1491 holdings, followed by group B, with 105 to 335 holdings. The size of the sample does not enable pinpointing with precision the lower boundary of this highest fraction, and one ends up with a gap of 46 between the two sub-zones. Given a margin of error of 4.4 percent, one should probably say only that highly ubiquitous holdings begin at about 360, plus or minus 40. With 11 items, group A makes up 2.2 percent of the sample. Remarkably, this is the same percentage that has either no matching record or a record with no attached holdings. Even more striking is that White, in a study of library collections, found the midpoint for library holdings to be about 400; furthermore, he asserts that this figure of 400 library holdings represents the dividing line between highly specialized books and books aimed at more general readership.[23] This suggests that fewer than 5 percent of items in WorldCat account for the great majority of library holdings.

A look at the items in groups A and B supports White's statement. All items in group A are published in the United States. Five are works of fiction. One is a translation. Publication dates range from 1954 to 2003. Publishers include industry giants—Warner Books, Houghton Mifflin, Viking, Basic Books, and St. Martin's (with two entries in this group)—along with a major scholarly publisher, Princeton University Press, as well as Orchard Books, a division of the Watts Publishing Group. The occurrence of two less-prominent publishers of group A books, the Center for Urban Policy Research at Rutgers University and John F. Blair (located in Winston-Salem, North Carolina), suggests that many relatively small publishers have at least some titles with high levels of holdings.

Most group B books seem to be located in a different sector of the book industry than group A books. Publication dates range from 1972 to 2002. Three of the 11 titles are published in England. This group includes 1 title each in the fiction, juvenile, and government publication

categories. In order of descending ubiquity, publishers of group B titles include Routledge, W. B. Saunders, Prentice-Hall, World Publishers, Kar-Ben Copies, Scholar Press/Ashgate, CRC Press, Overlook Press, Oneworld, Paulist Press, and the United States Government Printing Office. Most of these publishers are not in the trade or mass-market sectors of the publishing industry, but are dedicated to educational and scientific subjects.

Changes in Holdings over Time

By marking changes in the holdings of WorldCat records after a lapse of time, one can detect activity in the database. Ideally, one would like to know whether items circulated or were consulted, but WorldCat does not provide that information. Items whose holding levels have changed (upwards or downwards) over time are active; the others are static. As one might expect, items with the most holdings are the most active, while those with the fewest holdings are the least active. Surprisingly, however, items at all holding levels except nonexistent experience change over time.

Five months after first obtaining data on holding levels, each record in the sample was checked. Forty-six items (9.3 percent of the sample) had added holdings between July and December of 2004, while 24 (4.8 percent) had lost holdings during the same period of time. The largest number of holdings gained by any item was 91, while the largest number of holdings lost was 8.

Group A experienced the most marked change, with 5 items gaining and 5 losing. Only 1 item in this group did not change. In group B, 4 items gained and 3 lost. In the non-scarce category as a whole, 15 items gained in holdings, while 16 lost holdings. Therefore, 69 percent of non-scarce items changed in the absolute number of holdings in a five-month period. All but 3 of these items were print books.

Thirty items with scarce levels of holdings gained, while 8 lost. This means that 15.4 percent of scarce items changed in the number of libraries holding the item. Twenty-six of the scarce items that gained in holdings were print books. Only 2 unique items gained in holdings during the course of the study. While these items are exceptional, the study finds that activity is not exceptional at the level of 2 holdings. Indeed, 3 items that were not unique at the beginning of the study became unique by the end of the study by being deaccessioned from libraries that previously owned them. This study suggests that unique items are occasionally removed from holding libraries without deleting the records, resulting in nonexistent items.

Widely Shared Titles and Collecting Levels

This investigation finds that most items in WorldCat are rare and unusual items that do not have a place in most library collections and are unavailable for acquisition by libraries other than those that already own them. To use an oceanographic metaphor, most known content in WorldCat consists of the small fraction of materials at or near the surface, while a far greater amount of content is unknown, lying undisturbed deep below the surface and toward the ocean floor. The majority of items in WorldCat are extremely scarce, with two-thirds held by no more than 5 participating libraries. Even excluding manuscript and other archival materials (which make up a considerable portion of WorldCat), nearly 75 percent of items are held by a maximum of ten participating libraries. Although extremely scarce items are sometimes added to library collections, only a small fraction of WorldCat items are widely distributed among libraries and play a role in selection, acquisition, and collection management.

White has suggested that OCLC (i.e., WorldCat) holdings data can be used as a guide for collection development by checking a sample of items against the number of holdings in WorldCat to determine those items' collecting levels in terms of the RLG Conspectus.[24] The Conspectus provides guidelines for the acquisition of library materials, differentiating between minimal, basic information; study or instructional support; research; and comprehensive levels. These levels refer to libraries' objectives for specific call number ranges and subject collections, not whole libraries, and it should be noted that some levels have gradations and qualifications.[25] White developed what he called rules of thumb for determining the collecting level in these terms of given items. He reckoned that books with fewer than 150 holdings were at the research level; those with between 150 and 400 holdings were at the instructional level; those with between 400 and 750 holdings were at the basic information level; and those with more than 750 holdings were at the minimal level.[26]

From 2001 through 2002, Lesniaski reexamined the holding counts of items in White's sample lists, finding that the number of holdings for titles in each collecting level had risen significantly.[27] By adjusting each collecting level upward, he found that the approximate ratios between levels remained approximately constant. By Lesniaski's new rules of thumb, level 1 (minimal level) corresponds to WorldCat holdings above 1,000; level 2 (basic information level) corresponds to between 500 and 1,000 holdings; level 3 (study or instructional support level) corresponds to between 200 and 500 holdings; and level 4 (research level) corresponds to fewer than 200 holdings.[28] The average numbers of holdings of titles at these levels are 1,541, 751, 389, and 153 respectively, producing ratios of 10 : 4.9 : 2.5 : 1.[29]

The cut-off point for group A (ubiquitous items) is near the average for the study or instructional support level. Because the RLG Conspectus allows for the division of this level into two sublevels, 3a (study or instructional

level, introductory) and 3b (study or instructional level, advanced), one might connect group B with collecting levels 3b and 4, and group A with levels 3a and below. The category of ubiquitous holdings, therefore, encompasses the Conspectus introductory instructional or study level, the basic level, and the minimal level.

One would like to know the relative proportions of items in WorldCat at the various collecting levels. Only 1 item in the sample has more than 1,000 holdings and thus qualifies through Lesniaski's rule of thumb as minimal level. Six items are at the basic information level, having holdings between 500 and 1,000. Ten have holdings between 200 and 500, thus qualifying for the study or instructional support level. Although the maximum for the research level is set by Lesniaski at 200, and the mean is known to be 153, the line dividing the research level and the comprehensive level (level 5 of the RLG conspectus) is not established. White suggests that the research level goes down to 1 holding, and does not define a range for Level 5 titles, as he views the comprehensive level as "gap-filling over Levels 1 to 4. . . . For libraries already at Level 4 and seeking comprehensiveness, 'Level 5' titles are simply any remaining desiderata."[30] But the present study suggests that most unique and scarce items in WorldCat have either been superseded by more recent publications or pertain to such localized interests that they have little research significance beyond the communities in which they were produced. Based on an admittedly subjective and casual examination of bibliographic records from the sample, most print books in English with holdings fewer than 40 are *not* at the research level: 40 seems to be a reasonable cut-off point for such items. However, several other items with holding levels as low as 17—print books in languages other than English, along with items in other document and material types—also seem to be at the research level. Finally, academic theses (manuscripts) at the doctoral level with 1 holding should by definition have research value, and thus count as being at the research level, though whether libraries would seek to collect them is another question. White is correct if he means that items with 1 holding *could* be at the research level, and support for his view can be found in Perrault's discovery that 63.5 percent of monographic titles in research libraries are indeed unique.[31] But a qualitative examination of my sample suggests that, in practical terms, only a small fraction of the many items with fewer than 40 holdings actually *are* at the research level. If one accepts my suggestion of 40 as the lower cut-off point for the research level, and accepts that conspectus levels can in fact be equated with ranges of WorldCat holding levels, then the sample shows ratios for proportions of items at the various levels, from the minimal to the research levels, as 1 : 6 : 10 : 28. If, however, one accepts White's notion that the research level goes down to 1, the ratios are 1 : 6 : 10 : 478.

Summary and Conclusion

This study has analyzed a sample of bibliographic records in WorldCat to determine the proportion of items that have ubiquitous (widely shared) holdings versus lower levels of shared holdings. It left the term *ubiquitous* undefined, and worked with the heuristic categories of unique items, scarce items (items held by between 2 and 50 libraries), and non-scarce items (items held by more than 50 libraries). In these terms, the study finds that the large majority of items are either unique or scarce, with only 9.1 percent of items non-scarce.

Analysis of a random sample of 500 records in WorldCat shows that 2.2 percent of the sample is empty, matching no record or a record with no attached holdings. At the other end of the spectrum, the same percentage of the sample is matched by what may be called ubiquitous items. This uppermost fraction is obtained by taking the square root of the entire sample (22 for a sample of 500), using a formula devised by the information scientist Derek de Solla Price, to divide the sample into zones to account for items held by many libraries.[32] The highest zone, using this formula, consists of items with more than 100 holdings. Dividing this zone into halves results in a subzone of ubiquitous items consisting of items held by more than 380 libraries (out of more than 9,000 member libraries contributing to WorldCat). The items in this category consist of print-format books in English, all published in the United States in the last half century, mainly by giant mass-market publishing houses. Virtually all are cataloged and transcribed at the national level. Following a rule of thumb stating that the more ubiquitous an item is, the more basic it is as an educational or informational resource, the dividing line between ubiquitous and nonubiquitous items corresponds approximately to the line between introductory and advanced levels of instructional support, by the standards of the RLG Conspectus. Only .02 percent of items are at the minimal Conspectus level.

The distribution of items in the sample shows an inverse geometric relationship between the number of items at a given level of holdings and the number of libraries holding items, with the largest fraction (39.8 percent) held by just one library. Proportions of items with higher holdings can be described by Lotka's law, using the inverse square of the items with one holding.[33] While the fraction of unique items seems high, it is substantially lower than was found in a recent and much larger study.[34] Moreover, 20.9 percent of unique titles are actually archival materials (mainly manuscripts); bracketing this portion of the sample, only 34.6 percent of remaining items are unique.

This paper presents a categorical scheme accounting for the distribution of items with various levels of holding by member libraries—unique, scarce, and non-scarce—

along with a system of zones accounting for levels of nonscarce and ubiquitous items. Used with other quantitative and evaluative measures, these categories and levels are helpful in assessing individual library collections and online union catalogs.

References

1. Judith J. Senkevitch and James H. Sweetland, "Evaluating Public Library Adult Fiction: Can We Define a Core Collection?" *RQ* 36, no. 1 (Fall 1996): 103–17; Judith Serebnick, "Selection and Holdings of Small Publishers' Books in OCLC Libraries: A Study of the Influence of Reviews, Publishers, and Vendors," *Library Quarterly* 62, no. 3 (1992): 259–94; Debora Shaw, "An Analysis of the Relationship between Book Reviews and Fiction Holdings in OCLC," *Library & Information Science Research* 13, no. 2 (1991): 147–54; Danny P. Wallace and Bert R. Boyce, "Holdings As a Measure of Journal Value," *Library & Information Science Research* 11 (Jan. 1989): 59–71; Stephen E. Wiberley Jr., "The Humanities: Who Won the '90s in Scholarly Book Publishing," *portal: Libraries and the Academy* 2, no. 3 (July 2002): 357–74; Stephen E. Wiberley Jr., "The Social Sciences: Who Won the '90s in Scholarly Book Publishing," *College & Research Libraries* 65, no. 6 (Nov. 2004): 505–23.
2. Allan N. Mirwis, *Subject Encyclopedias: User Guide, Review Citations, and Keyword Index* (Phoenix, Ariz.: Oryx Pr., 1999).
3. Howard D. White, *Brief Tests of Collection Strength: A Methodology for All Types of Libraries,* Contributions in Librarianship and Information Science, no. 88 (Westport, Conn.: Greenwood, 1995); Anna Perrault, *Global Collective Resources: A Study of Monographic Bibliographic Records in WorldCat* (report of a study conducted under the Auspices of an OCLC/ALISE Research Grant, July 2002), www.oclc.org/research/grants/reports/perrault/intro.pdf (accessed Oct. 6, 2004).
4. White, *Brief Tests of Collection Strength,* 135–36.
5. Perrault, *Global Collective Resources.*
6. Ibid.
7. OCLC Online Computer Library Center, "WorldCat," www.oclc.org/worldcat (accessed Aug. 11, 2004).
8. OCLC Online Computer Library Center, "Watch WorldCat Grow," www.oclc.org/worldcat/grow.htm (accessed Aug. 11, 2004).
9. OCLC Online Computer Library Center, "WorldCat Gold Records," www.oclc.org/worldcat/goldrecords.htm (accessed Aug. 11, 2004).
10. Thompson M. Little, "OCLC's International Initiatives and the Online Union Catalog," *Cataloging & Classification Quarterly* 8, no. 3/4 (1988): 67–78.
11. OCLC Online Computer Library Center, "OCLC Annual Report 2002–2003: Year in Review," www.oclc.org/news/publications/annualreports/2003/yearinreview.htm (accessed Aug. 16, 2004).
12. Ibid.
13. Random.org—True Random Number Service, www.random.org (accessed Mar. 10, 2005).
14. OCLC Online Computer Library Center, "WorldCat."
15. *Anglo-American Cataloguing Rules*, 2nd ed., 2002 rev. (Ottawa: Canadian Library Association; London: Chartered Institute of Library and Information Professionals; Chicago: ALA, 2002), v.
16. See J. W. D. Dougherty, "Salience and Relativity in Classification," in *Language, Culture, and Cognition: Anthropological Perspectives*, ed. Ronald W. Casson (New York: Macmillan, 1981), 163–80; George Lakoff, *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind* (Chicago: Univ. of Chicago Pr., 1987); David B. Kronenfeld, *Plastic Glasses and Church Fathers: Semantic Extension from the Ethnoscience Tradition* (New York: Oxford Univ. Pr., 1996); Anna Wierzbicka, *Lexicography and Conceptual Analysis* (Ann Arbor, Mich.: Karoma Pub., 1985).
17. Perrault, *Global Collective Resources*.
18. Ibid., table 4-1a.
19. Jay Jordan, "Global Networking of Information—OCLC's Strategy for the Future" (paper presented at 7th International Bielefeld Conference, Feb. 3–5, 2004, Bielefeld, Germany), http://conference.ub.uni-bielefeld.de/proceedings/jordan.pdf (accessed Oct. 12, 2004).
20. OCLC Online Computer Library Center, "WorldCat Facts and Statistics," www.oclc.org/worldcat/statistics/default.htm (accessed Sept. 1, 2004).
21. Concepción S. Wilson, "Informetrics," *Annual Review of Information Science and Technology (ARIST)* 34 (1999): 173; Alfred J. Lotka, "The Frequency Distribution of Scientific Productivity," *Journal of the Washington Academy of Sciences* 16, no. 12 (1926): 317–23.
22. Dietmar Wolfram, *Applied Informetrics for Information Retrieval Research,* New Directions in Information Management, no. 36 (Westport, Conn.: Libraries Unlimited, 2003), 46; Derek J. de Solla Price, *Little Science, Big Science* (New York: Columbia Univ. Pr., 1963).
23. White, *Brief Tests of Collection Strength*, 135–36.
24. Ibid.
25. Anthony W. Ferguson, Joan Grant, and Joel S. Rutstein, "The RLG Conspectus: Its Uses and Benefits," *College & Research Libraries* 49, no. 3 (1988): 197–206.
26. White, *Brief Tests of Collection Strength,* 127.
27. David Lesniaski, "Evaluating Collections: A Discussion of Brief Tests of Collection Strength," *College & Undergraduate Libraries* 11, no. 1 (2004): 11–23.
28. Ibid, 17.
29. Ibid.
30. White, *Brief Tests of Collection Strength*, 129.
31. Perrault, *Global Collective Resources*.
32. Price, *Little Science, Big Science.*
33. Lotka, "The Frequency Distribution of Scientific Productivity."
34. Perrault, *Global Collective Resources*.