DOCUMENT RESUME

ED 047 743 LI 002 624

AUTHOR Resnikoff, Howard L.
TITLE On Information Systems, With Emphasis on the Mathematical Sciences.
INSTITUTION Conference Board of the Mathematical Sciences, Washington, D.C.
SPONS AGENCY National Science Foundation, Washington, D.C.
PUB DATE Jan 71
NOTE 31p.
AVAILABLE FROM Conference Board of the Mathematical Sciences, 834 Joseph Henry Building, 2100 Pennsylvania Avenue, N.W., Washington, D.C. 20037 ($1.00)

EDRS PRICE EDRS Price MF-$0.65 HC Not Available from EDRS.
DESCRIPTORS Automation, Conferences, Electronic Data Processing, *Information Dissemination, *Information Services, *Information Systems, *Mathematics Materials, *National Programs
IDENTIFIERS *Information System for the Mathematical Sciences, ISMS, Scientific and Technical Information

ABSTRACT

In the area of the mathematical sciences, a naticnal information system will principally be concerned with the dissemination (including publication) rather than the creation of primary information. Here "dissemination" refers to both general distribution of information which is not responsive to specific pre-existing inquiries as well as demand dissemination which does respond to specific requests. Part I discusses the need for an information system for the mathematical sciences (ISMS) and proposes some ISMS access systems. An access system is one which provides the user access to the comprehensive corpus of existing information that may be of interest to him. Part II discusses some proposed ISMS access systems from the standpoint of their "automatability" and their connection with the publication process. (NH)

CONFERENCE BOARD OF THE MATHEMATICAL SCIENCES

Committee on a National Information System in the Mathematical Sciences

ON INFORMATION SYSTEMS,

WITH EMPHASIS ON THE MATHEMATICAL SCIENCES

by

HOWARD L. RESNIKOFF

R & D Consultants Company
and
Rice University

Prepared with the support of the National Science Foundation

January 1971

Additional copies available at $1 from
Conference Board of the Mathematical Sciences
834 Joseph Henry Building
2100 Pennsylvania Avenue, N.W.
Washington, D.C. 20037

PREFACE

The Conference Board's Committee on a National Information System in the Mathematical Sciences was established in 1970 as an outgrowth of earlier activities of the mathematical community looking toward systematic improvement in the information services available concerning the full range of subject matter of the mathematical sciences.

The Committee is presently engaged in the preliminary planning of a program designed to supplement the present primary publication activities of its member societies, and other publishers, through improved and expanded secondary services which will provide better access to the primary literature. As a basic component of this preliminary planning, the Conference Board has commissioned Professor Howard L. Resnikoff to make an analysis of systematic approaches to information storage and retrieval problems, especially as they apply to mathematical knowledge. With his dual experience as a consultant in information problems and as an active research mathematician, we feel he is especially qualified to develop a theoretical framework for scientific information transfer applicable to the mathematical sciences and probably to other fields as well.

The information system planning activities of the Conference Board, including the preparation of this report, are supported by a grant from the Information Systems Program, Office of Science Information Service, National Science Foundation.

Robert M. Thrall
Chairman, Committee on a
National Information System
in the Mathematical Sciences

January 1971

ON INFORMATION SYSTEMS, WITH EMPHASIS ON THE MATHEMATICAL SCIENCES

PART I

1. An Information System should be the response to a pressing need as well as a laboratory for future developments. There is ample testimony to the need for an information system for the mathematical sciences (ISMS) as well as for other sciences and indeed for many other areas involving the national welfare.

In general an Information System involves the storage, transmission, and retrieval of information but insofar as the systems aspects are concerned it is the retrieval of information that specifies the urgency with which the other problems will receive attention and funds. In the area of the mathematical sciences a national information system will principally and normally be concerned with the dissemination (including publication) rather than the creation of primary information. Here we understand by "dissemination" both general distribution of information which is not responsive to specific pre-existing inquiries as well as demand dissemination which does respond ("on-line" or "off-line") to specific requests.

Systems of the first type are *passive* in that their users are typically not required to provide input to the system in advance of their use of it. The *active* systems of the second type do require such input and are consequently more precise in their response capabilities, more expensive to use and maintain, and slower to respond to the changing demands of users. *Mathematical Reviews* is an example of a passive system component whereas the *Mathematical Offprint Service* and all on-line systems are examples of active components.

Active systems appear to have a significantly higher mortality rate than passive ones in all fields of application, perhaps partly due to the greater degree of investment--of time and effort if not money--demanded of the user. For instance, the oft-heralded desktop computer console demand inquiry information system for use by executives of large corporations has made no headway because, principally, of the unanticipated difficulty of posing useful inquiries in addition to the problem of framing them within the rigid syntactic limitations imposed by the system. On the other hand, although most of the information it contains is "noise" for any single user, the telephone directory is one of the most successful and convenient information access systems currently in use.

From this viewpoint it appears that the principal problem that must be solved by an information system is that of providing the user access to the comprehensive corpus of existing information that may be of interest to him. The primary notion is that of *access*; the critical ISMS problem is to design an economically viable system that will provide access at all useful levels to the entire primary data base of the mathematical sciences, the latter conceived as the collection of all journal papers, monographs, textbooks, analyses, tables, working papers, etc., in all fields that use and teach mathematics.

Access is a problem because the archive is now large and grows rapidly. We reproduce as Figure 1 a well-known graph illustrating the growth of the number of scientific periodicals and abstract journals [1]. Each class grows exponentially with time, at the same rate (approximately 5% per year), doubling every 15 years. When the amount of information in the archive becomes larger than some "critical" quantity, it apparently becomes necessary to introduce a new method for accessing the archive; for scientific periodical information this point was reached with the existence of about 300 journals around 1830, and abstract journals were born.

The figure shows that the number of abstract journals reached about 300 in 1950, which suggests that it is past time for the introduction of an access system that will bear the same relationship (in the sense of size) to abstracts that abstracts bear to journal articles. Indeed, the permuted title index such as that produced by Chemical Abstracts Service functions in essentially this way, and the interest of the Federal Government in information retrieval systems since 1950 could be interpreted as a response to the need for a new level of access which will correspond to a third line on Figure 1, parallel to the other two but intersecting the time axis at about 1950, as shown labeled "second order access system".

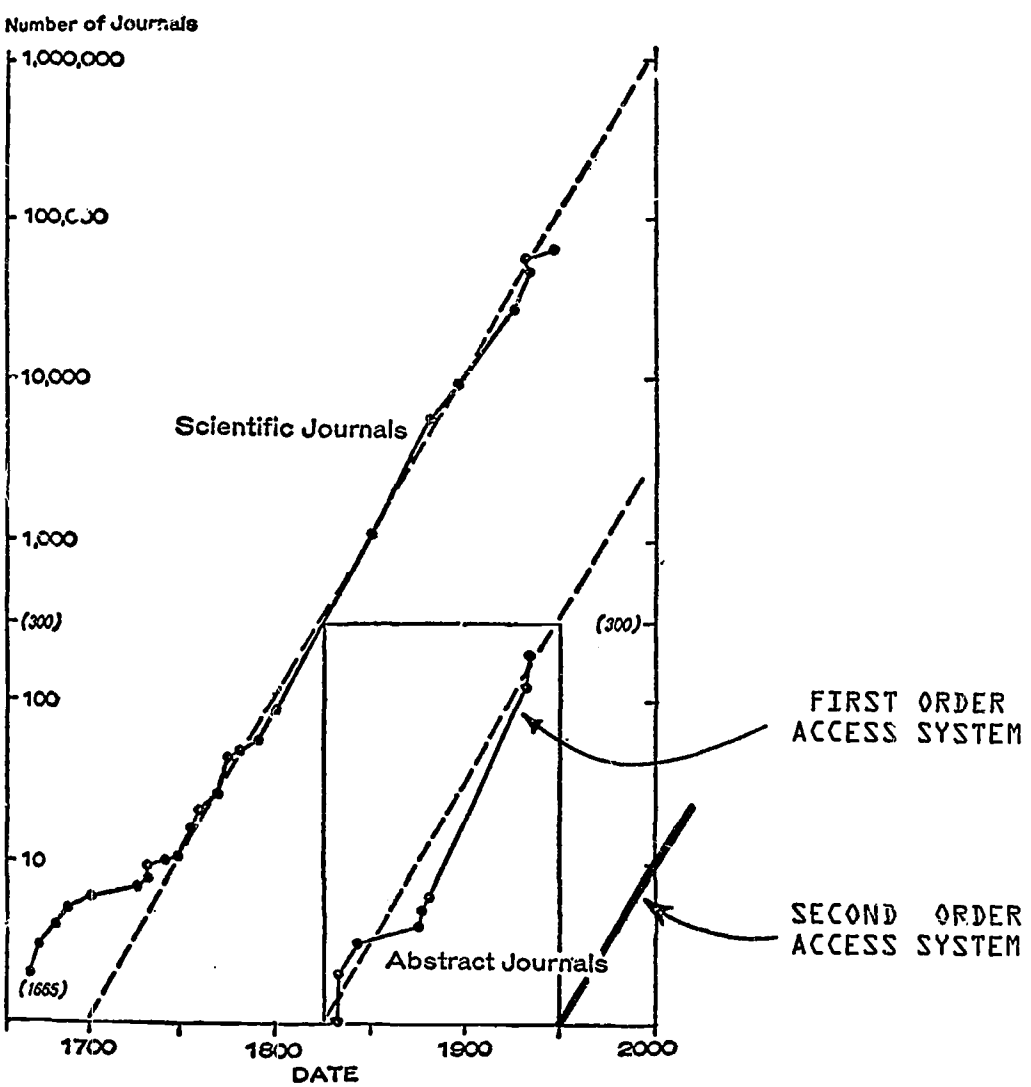
Data for this figure and a substantial body of additional evidence show that the need for access is primarily--perhaps entirely--a function of the *size* of the information base regardless of the nature of the specific contributions which constitute the data archive; hence, an organization of information bases and access systems according to size considerations is meaningful. The following sections explore this kind of organization in detail for some of the more important traditional access systems and primary information bases and discuss some of the more obvious implications for an ISMS.

The nature of the information in the archive, of its contemporary relevance, or of the probably specific utility of an archival document play only minor roles in our study, for once a document has reached the archive, having passed whatever admission tests may have been imposed, it is reasonable to argue that a comprehensive access system should provide means for finding it and for ascertaining the nature of its content. Moreover, this view affords an important simplification for statistical analyses of access systems and their interrelations.

In an analogous way the library catalog card is a traditional access system which does not pass judgments but simply records the presence and grossly specifies the subject matter of each monographic item in the archive. The unusually useful *Reviews of Papers in Algebraic and Differential Topology, etc.* [2] compiled by N. Steenrod do the same thing within certain formally defined temporal and spatial confines.

These observations are of more than passing interest because they hint at the possibility, which we shall later show to be highly probable, that the basic modular components of a large scale information system are automatable.

Regarding this point, we cannot today see how to automate all of the desirable features of existing access systems, but when these are restricted to special areas of the total archive which have a high degree of internal consistency and a relatively small size, such as the mathematical sciences archive, it becomes possible to consider the implementation of a highly automated system which can serve both as an effective Information System for that area as well as a pilot project for the extension and development of



Number of Scientific Periodicals (Data from D. J. de Solla Price, *Science since Babylon* [New Haven, 1961], p. 97).

FIGURE 1

national information systems of broader scope. In this sense we consider an ISMS to comprehend two essentially disjoint components: a *general component* concerned with the problems of access that are common to most subportions of the total information archive, and a *special component* which addresses itself to those access problems peculiar to the mathematical sciences including, for example, access to tabular material such as tables of integrals or series sums as well as the notational problems shared to some extent by the other sciences. It is the general component which currently does most of the work in an access system, as will be adequately confirmed by the statistics for the total monographic information archive which will be presented below. This will be obvious to anyone who has systematically used the existing abstract journals, current contents journals, library card catalogs or (where generality of application is perfectly clear) line-printer-produced permuted-title and related listings. Consequently it can be anticipated that the general modules developed for an ISMS will have a high degree of utility for other national information systems outside of, as well as within, the scientific fields; and, conversely, some of the modules developed for other information systems such as that currently in use in chemistry will be adaptable for use in the mathematical sciences.

2. If indeed size is the motivation behind the use of access systems, then it is important to determine the size relationships of the primary archival material to the various access systems currently in use and to arrange these systems in size-ordered levels so as to discover whether there are any potential access levels not now used but which could be used in an ISMS. We rely primarily on [3] and [4] in addition to data developed specifically for this study.

2.1. First, consider the size of mathematical abstracts. From a recent volume of *Mathematical Reviews* we calculate the mean page length of a mathematical journal article as approximately 13.8 (unnormalized) pages, whereas the mean length of an abstract in that volume is 0.45 pages normalized to the page and font size of the *Transactions of the American Mathematical Society* as a standard. Therefore the mean mathematics paper is 30.6 times as large as the mean mathematics abstract.

2.2. Consider the type of index normally found at the back of a book. For a random sample of 706 books with indexes drawn from the complete shelf list of the Fondren Library at Rice University it appears that the mean number of index entries per indexed book is 836. The average number of text pages per book is 276.6, which implies, after appropriate adjustments for variations in font and page size, that the average book is 31.8 times as large as the mean index; details appear in [3]. Hence a book index bears the same size relationship to the indexed text as a mathematics abstract does to its primary text. This observation assumes significance when the problems inherent in automating various existing access systems are considered, for indexes and abstracts belong to the same access level, but one is relatively easily automated whereas the other is probably unautomatable.

2.3. Next, consider the size of index entries. In a uniform sample from an alphabetized combined list of the index entries from 79 books in statistics, the mean number of characters per index entry, inclusive of pagination, punctuation, and internal space characters, but excluding the (two character) codes which specify the source book, is 33.6 characters per index entry. This is of no directly practical interest for the construction of an ISMS because index entries are themselves too small to require access modules in the usual sense, but this example helps to delineate the nature of the access model that could underlie an ISMS design, for

the one-letter Library of Congress class designation "Q" is indeed an existing access system which would be of utility in this context were the statistics index terms amalgamated with index terms from books occurring throughout the scope of the total library archive. The important point for us is that once again we observe that the access system provides a compression of information (with loss, of course) of a ratio of about 30 to 1.

2.4. The mean character length of titles of monographs represented in the Fondren sample [4] is 28.15 characters per title. We direct the reader's attention to the fact that this random sample includes items from all categories in the general library classification schedule.

2.5. Consider a more complex example which refers directly to the access problem. It is usual to find so-called "subject headings" at the foot of library catalog cards which are intended to provide cross-reference access to subject areas other than those associated with the class number of the item corresponding to the catalog card. There are nearly 93,000 subject headings in the Library of Congress standard list. A uniform 1/66 sample drawn from an alphabetized list of these headings shows that the mean number of characters per subject heading is 22.3, which cannot be considered as particularly close to 30. However, the distribution of subject headings per catalog card as determined from the Fondren sample has a mean of 1.2 per card; assuming that the distribution of subject headings per card is independent of the distribution of characters per subject heading, the mean number of subject heading characters per catalog card, including the associated ordinals and space characters, is 29.16. Hence, the collection of subject headings per card provides *in the mean* about the same level of discrimination above the one-letter Library of Congress class that is provided by the title (*cp.* section 2.4); sometimes the subject headings prove a better guide to content than the title.

2.6. The phenomenon that the means of access level sizes are in the ratio of about 30 to 1 is not confined to the access systems normally associated with library archives and journal articles. Consider ALTEXT, a recent text-processing higher level (macro expander) computer language [5]. Such a language consists of computer instructions which have two parts: a *generic instruction* such as the GOTO of FORTRAN which specifies the general function of the instruction, and certain other more particular components which contain the details of data location and transfers. The implementation of a higher level language instruction consists of a sequence of one or more "machine language" or "assembly language" instructions; the advantage of the higher level language is that it frees the programmer from the burden of keeping numerous housekeeping details in mind at the cost of lower (local) efficiencies in execution. This is another way of stating that the higher level instructions act as an access system for the sequences of assembly language instructions that are their implementation.

With this preamble in mind, we can examine the number of assembly language instructions required to implement each of the higher level instructions in a higher level language. For the generic instructions in ALTEXT, the average number of assembly language instructions per ALTEXT "macro" is 30.32 (including implementation of the "ALTEXT macro" which provides the interface with the operating system) for implementation on the IBM 360/30 computer.

3. The examples in the previous section suggest that many existing access systems occur in a level structure for which the mean sizes of successive levels are in the ratio of about 30 to 1, so that the absolute size

(in the mean) of elements belonging to level n is some fixed multiple of K^n where K is a constant nearly equal to 30. Before such an hypothesis can be seriously considered the distributions of elements of the various access levels according to size must be examined since the constant K was obtained solely from a consideration of sample means. If the distributions are multimodal, the mean might provide a misleading statistic; even if the distributions are unimodal the mean may be both misleading and insufficient as a descriptor of the distribution.

In this section we show that all of the distributions in question are *lognormal*. Moreover, for access levels spanning the range from 1 to 1.7×10^{13} characters, corresponding to 10 access levels, the distribution standard deviations (which, with the mean, completely determine these lognormal distributions) all lie within a narrow range. The distribution mean is therefore essentially the only significant distribution statistic and effectively determines the distribution.

Figures 2 through 6 illustrate the distributions corresponding to the access systems discussed in section 2 on logarithmic probability graph paper; straight lines thereon correspond to lognormal distributions.

It is pertinent to record that it is also known that the following distributions are lognormal:

dictionary entries measured in characters;
sentence length measured in words;
syllables measured in vowel strings;
words uttered in telephone conversations measured in characters for equivalent written forms.

Of greater immediate significance and at the upper end of the size scale, the distribution of university libraries measured in the number of books held is lognormal. Figure 7 illustrates this for the 64 largest university libraries under the hypothesis that there are 100 libraries properly associated with that level. The one university library which departs from the lognormal trend line to an appreciable extent is the largest --Widener at Harvard University--which the model shortly to be presented classifies as a member of the next larger access level (K^9) inhabited by the National Libraries.

4. These data and others reported in [3] support the following model for information access systems:

The distributions of items occurring in the various access levels is of the form

$$\frac{1}{xs\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\log x - m}{s}\right)^2}$$

The distribution means are powers of some constant K nearly equal to 30. There is not yet sufficient information to accurately fix the value of K ; it is however suggestive and consistent with the data to assume

$$K = (2e)^2 = 29.54,$$

which we do in what follows.

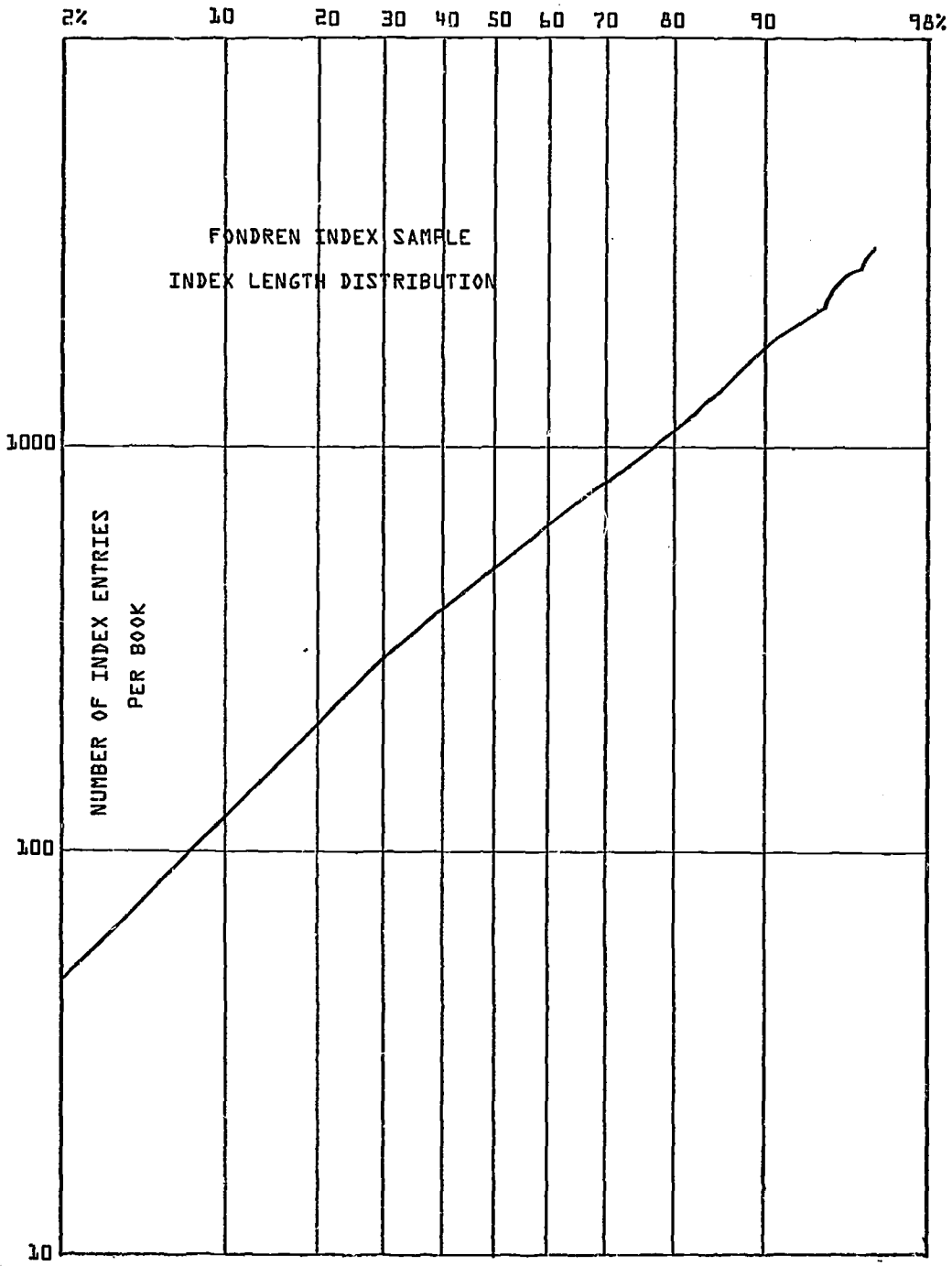

2%
10
20
30
40
50
60
70
80
90
98%
FONDREN INDEX SAMPLE
INDEX LENGTH DISTRIBUTION
1000
100
10
NUMBER OF INDEX ENTRIES
PER BOOK

FIGURE 2

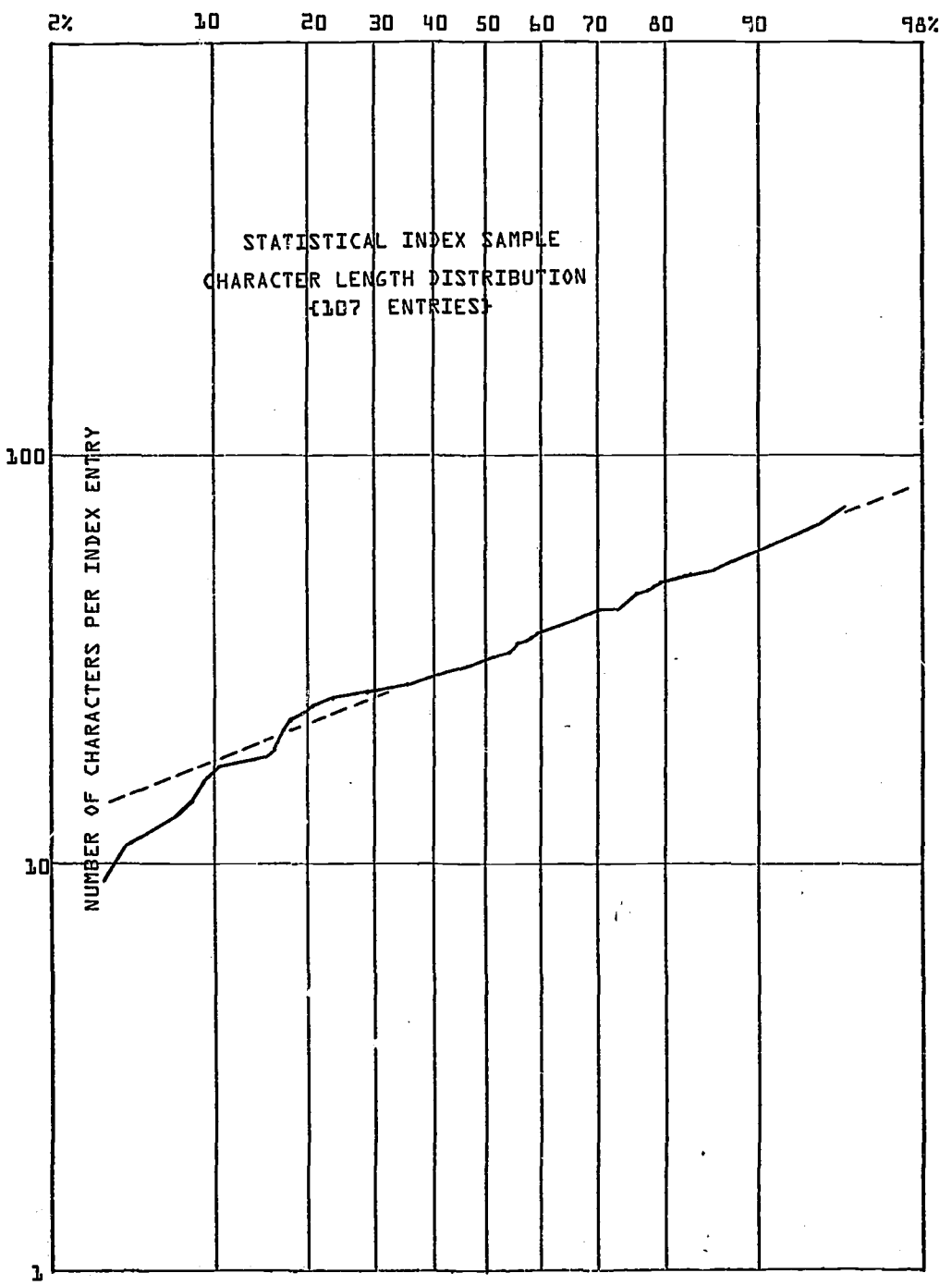

2%
10
20
30
40
50
60
70
80
90
98%
STATISTICAL INDEX SAMPLE
CHARACTER LENGTH DISTRIBUTION
(107 ENTRIES)
NUMBER OF CHARACTERS PER INDEX ENTRY
100
10
1

FIGURE 3

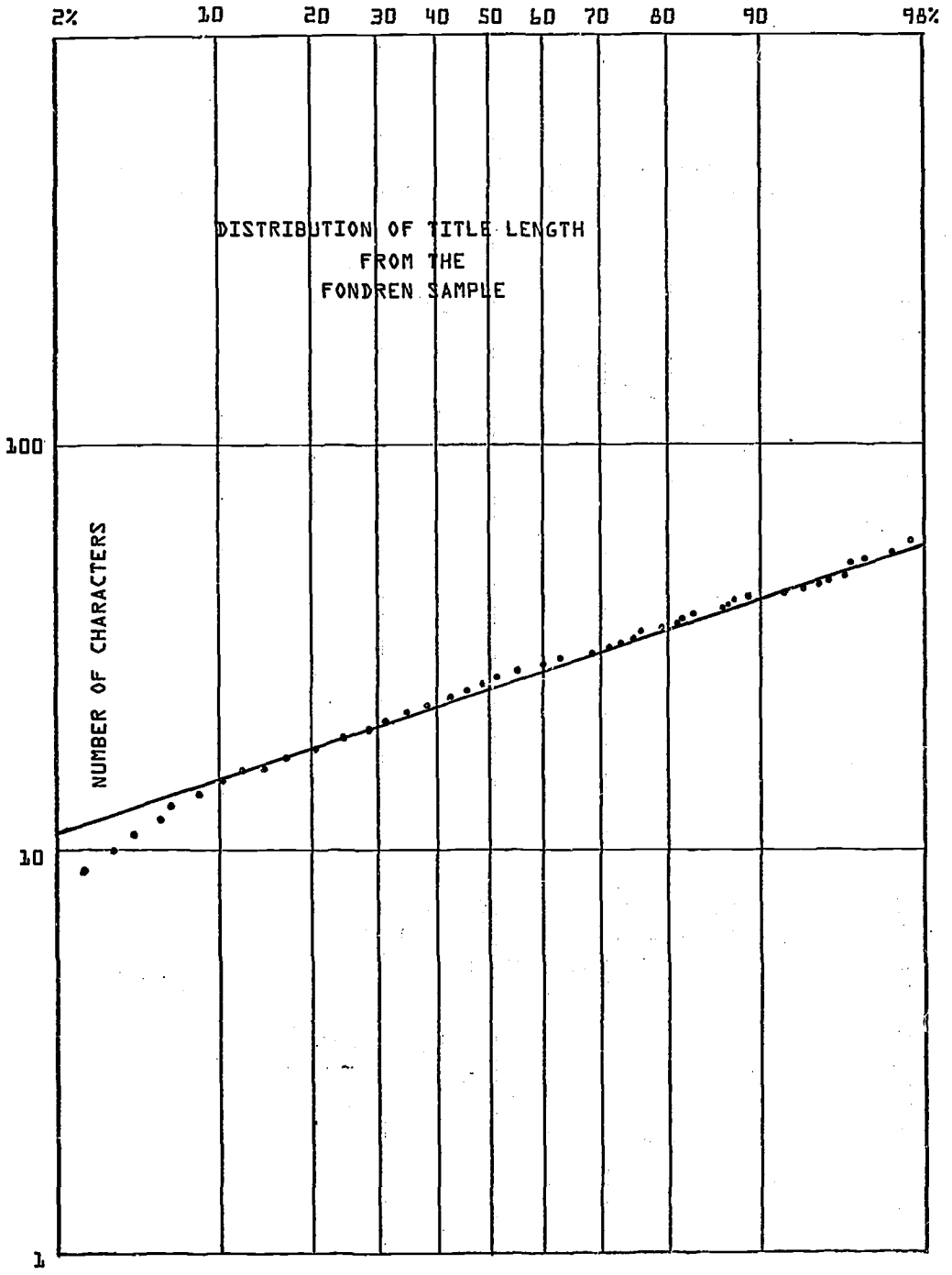

2%
10
20
30
40
50
60
70
80
90
98%
DISTRIBUTION OF TITLE LENGTH
FROM THE
FONDREN SAMPLE
100
10
1
NUMBER OF CHARACTERS

FIGURE 4

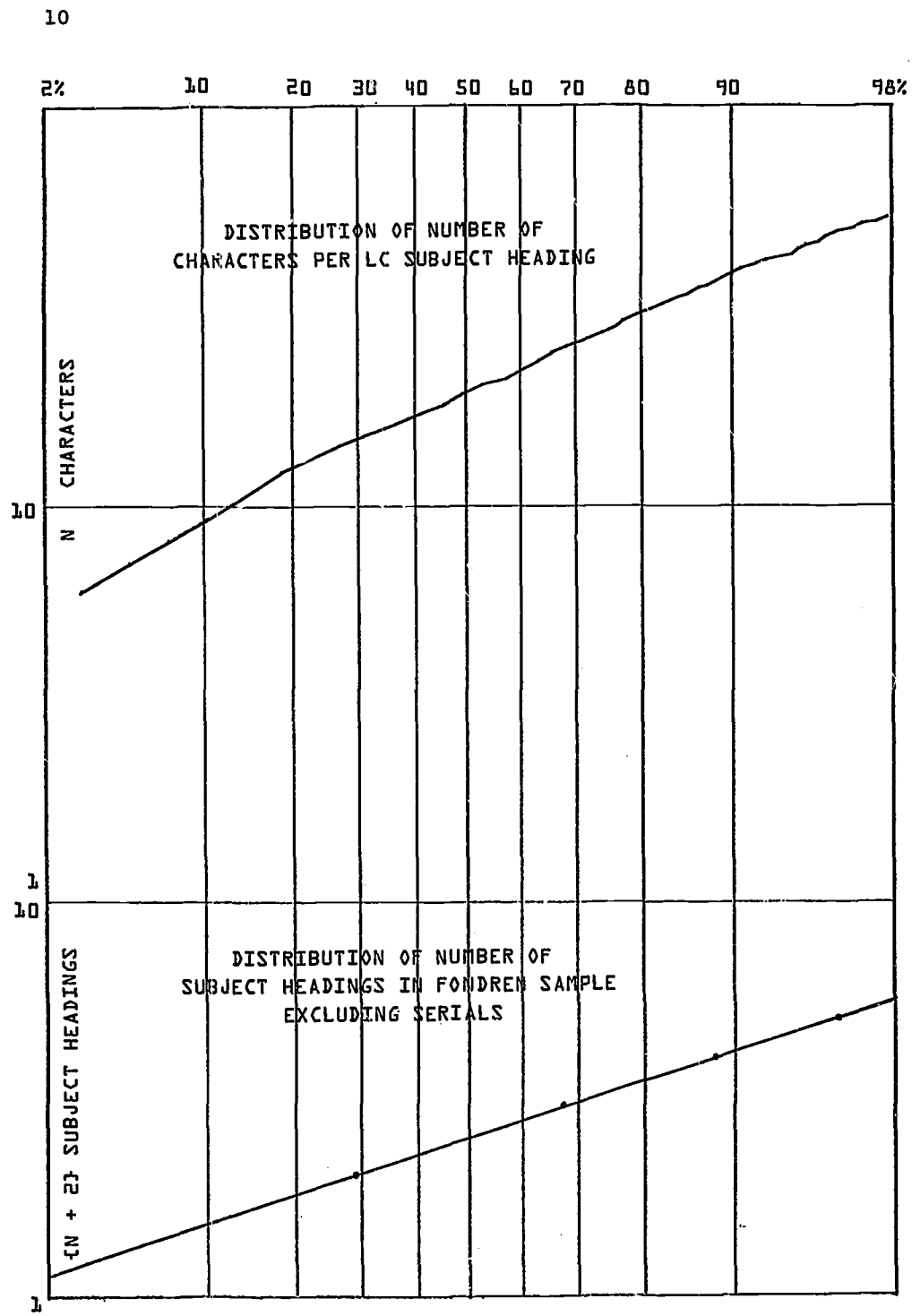

10
2%
10
20
30
40
50
60
70
80
90
98%
DISTRIBUTION OF NUMBER OF
CHARACTERS PER LC SUBJECT HEADING
N CHARACTERS
10
1
10
DISTRIBUTION OF NUMBER OF
SUBJECT HEADINGS IN FONDREN SAMPLE
EXCLUDING SERIALS
(N + 2) SUBJECT HEADINGS
1

FIGURE 5

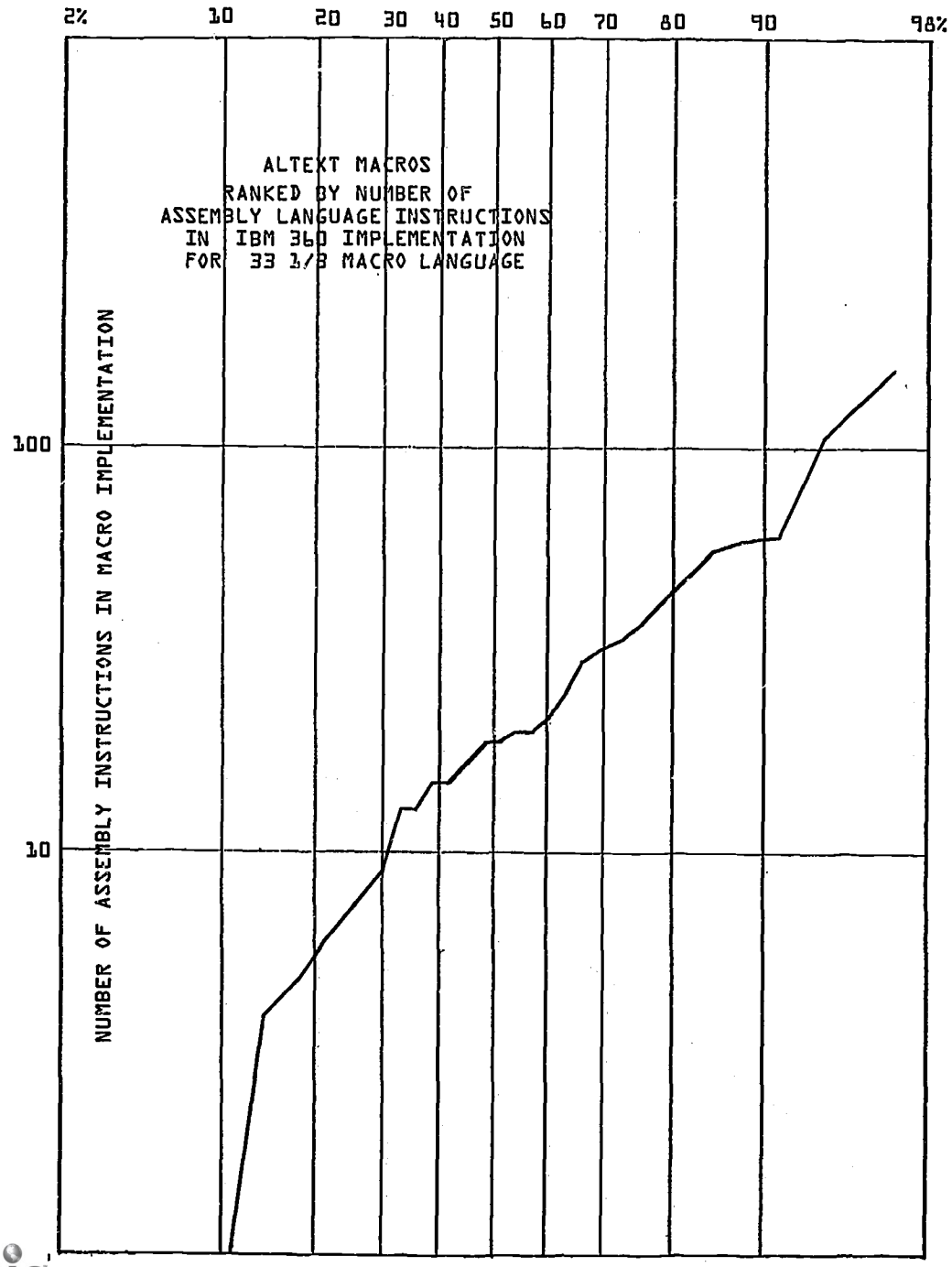

2%
10
20
30
40
50
60
70
80
90
98%
ALTEXT MACROS
RANKED BY NUMBER OF
ASSEMBLY LANGUAGE INSTRUCTIONS
IN IBM 360 IMPLEMENTATION
FOR 33 1/3 MACRO LANGUAGE
NUMBER OF ASSEMBLY INSTRUCTIONS IN MACRO IMPLEMENTATION
100
10
1

FIGURE 6

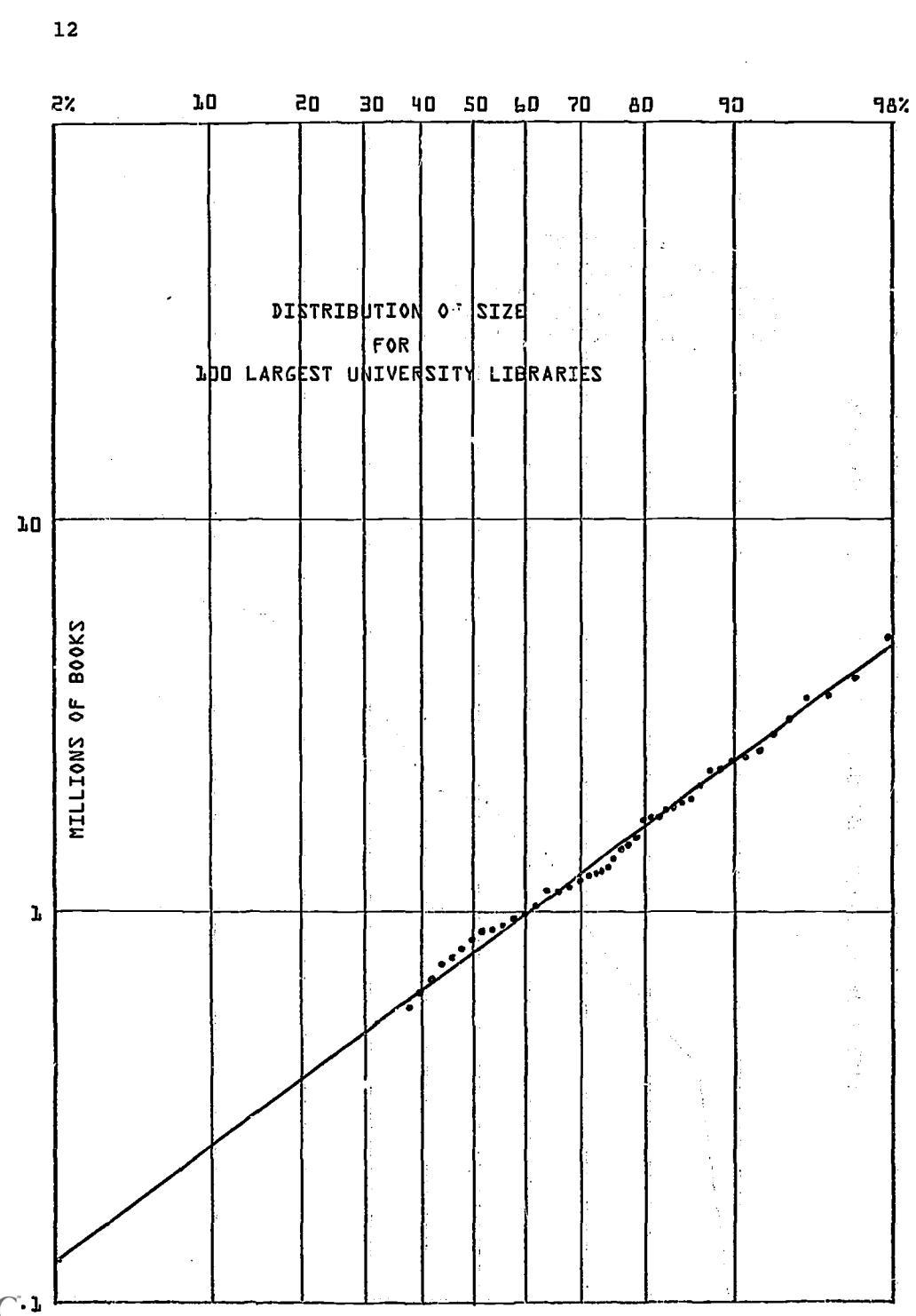

12
2%
10
20
30
40
50
60
70
80
90
98%
DISTRIBUTION OF SIZE
FOR
100 LARGEST UNIVERSITY LIBRARIES
10
1
.1
MILLIONS OF BOOKS

FIGURE 7

The standard deviations of the distributions corresponding to the access levels are all confined within a relatively narrow range when compared with the range of size under consideration. Indeed, the standard deviation, which in logarithmic coordinates is proportional to the slope of the line in Figures 2 to 7, varies between 0.18 and 0.43 as the size varies from about 30 characters to about 10^{13} characters.

It may be worthwhile to note that the above form of the lognormal distribution is not the most general, which latter is obtained by replacing x by $x-a$; for the access systems described in Figures 2-4 and 6-9 we may take $a=0$ without significant error.

Define the *level* of a given information base to be the integer closest to the logarithm of its size to the base K . Then the level of the average mathematics journal article is 3, while its abstract and title are of level 2 and 1 respectively. Typical access systems of various size from level 0 through level 9 are shown in Table I, measured in characters through level 4 and in books thereafter.

Table I

Access Level Structure

Level	Size	Conventional Name
0	1 char.	Library of Congress Class
1	30	Title; Index Entry; Subject Headings per Catalog Card
2	874	Table of Contents; Abstract of Mathematics Journal Paper
3	25,822	Book Index; Mathematics Journal Paper
4	763,203	Book
5	30 books	Encyclopedia
6	874	Catalog for a Level 7 System; Personal Library
7	25,822	Catalog for a Level 8 System; Seminar Library (Mathematics Library)
8	763,203	University Library
9	22,557,422	National Library (Library of Congress)

Note that the Library of Congress currently holds about 16 million books but 58 million "items", the latter including audio and visual records and other non-book materials.

If x is distributed lognormally, then $\log x$ is necessarily normally distributed. Our model implies that when the logarithm of size is used as the variable in place of the size itself, the various access system distributions are *normal* with equispaced means coinciding with the level when base K logarithms are used. Moreover, the standard deviations vary little from level to level. This situation is shown for access distributions for levels 1 to 3 obtained from the data described in [3] in Figure 8. The relatively small amount of overlap of successive distributions supports the notion of classification according to size and even according to average size of a class of access systems of equivalent type. We stress that this figure exhibits the normal distributions corresponding to actual data in the sense that they are constructed from the means and standard deviations of that data, some of which are further exhibited in Figures 2 and 4.

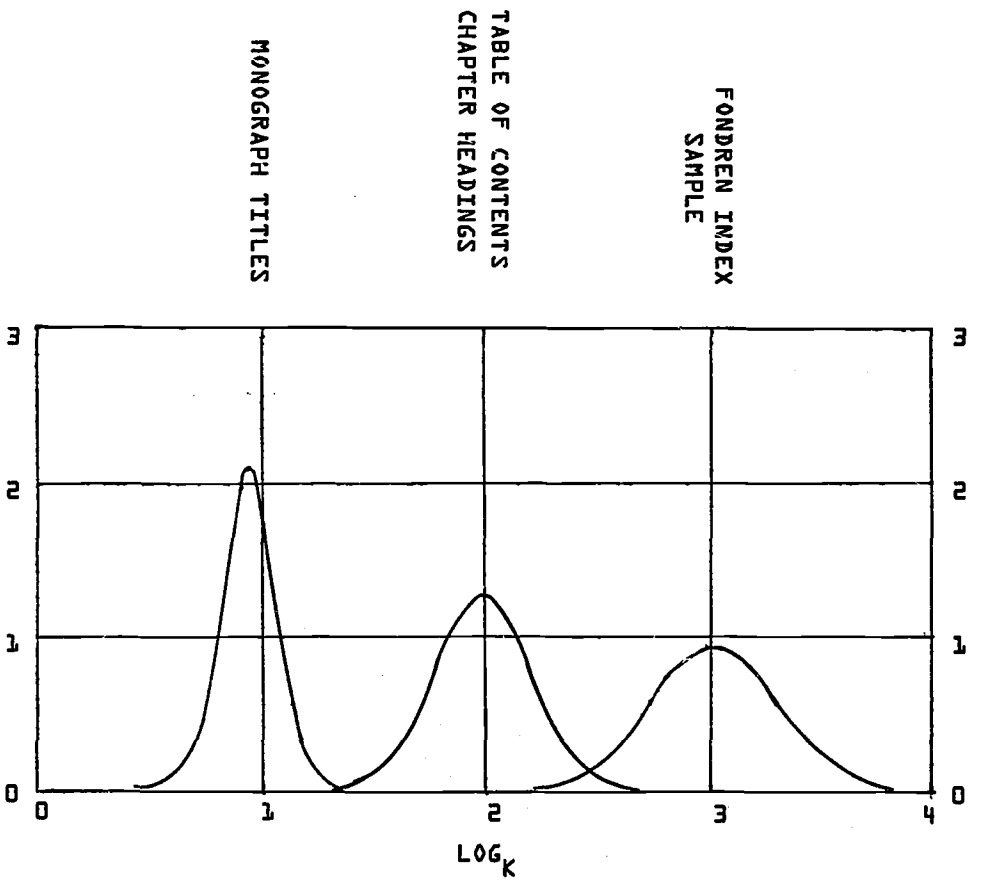


FIGURE 8. SOME ACCESS DISTRIBUTIONS IN LOGARITHMIC VARIABLES

The intrusion of the lognormal distribution is not really surprising, as the following argument shows. Consider information "states" S_x constituting some inventory, such as the words of a language as they occur in a large text corpus, or a large random sample of indexes or titles, ordered in some convenient fashion. Let $E(x)$ denote the "effort" required to utilize state S_x in an access system, and denote by $p(x)$ the probability of utilizing S_x from the inventory. Following Shannon, the expected amount of information per unit effort is proportional to

$$I = -\sum p(x) \log p(x) \; / \; \sum p(x)E(x) \; .$$

If the access system is such that the expected amount of information per unit effort is maximized, then maximization of I subject to the restraint $\Sigma\, p(x) = 1$ leads to the solution

$$\log p(x) = a_0 + a_1 E(x)$$

with constants a_i . This is a well known argument due to Mandelbrot [14].

It remains to specify the effort function $E(x)$. If the states S_x are words drawn from natural text and arranged in order of decreasing frequency of occurrence, then it is natural to argue, as Mandelbrot did, that $E(x)$ is proportional to $\log (x+a)$ with some small constant a . More generally, we may suppose that $E(x)$ has a Taylor series expansion about 0 in the variable $\log (x+a)$. If $E(x)$ be approximated by the terms through order two in this expansion, it immediately follows that $p(x)$ is a lognormal function of x .

The logarithmic measure of information impact (not to be confused with Shannon's "information") which is implied by this model is probably related to the underlying phenomena responsible for the well-known (and once again hotly debated) Weber-Fechner Law in psychophysics which asserts that (psychological) response is proportional to the logarithm of physical stimulus, as exemplified by the decibel scale of measurement of perceived acoustic energy and the octaval musical pitch scale. This connection is explored more fully in [3].

A closely related phenomenon is discussed in [7] where it is shown that the distribution of term usage from a manipulative index vocabulary is lognormal for nine large document indexing systems, including the Uniterm Index and a DDC (Defense Documentation Center) collection of 195,000 documents.

References [6] and [8] should also be noted; the former includes data which support the view that the lognormal distributions of access means and the logarithmic level structure of the model are consequences of psychophysical properties.

5. Examination of the variance of the lognormal distributions of the access systems studied shows that the various levels have only a small degree of overlap, so that the bulk of the access systems which belong to a given level in the structural sense actually lie within it in the following natural sense: the base K logarithm of their size lies between $(n - 1/2)$ and $(n + 1/2)$ if they have a logical structure belonging to level n (see Figure 8). Because of this it is possible to introduce the notion of *boundaries* between levels, which correspond to sizes that are $2e$ (= 5.4) or $1/2e$ times the mean size of the levels. For example, a book is a level 4 access system; its index belongs to level 3 in the structural sense. If a particular book index is larger than $1/2e$ the size of the book, then it is too large to be properly called an index, and in fact it does not function effectively as an index. Similarly, if it is smaller than the book by a factor greater than $(2e)^3 = 160$, then it will function more like a table of contents than an index.

The implication for the design of access systems is that the average size of the access structures should be approximately 1/30th that of the information base to be accessed for first level access, but in no case should any proposed access system provide less than a factor of 5.4 compression, for then it provides access at essentially the same level as the information base and one might just as well search the base directly.

Several different access systems may belong to the same level; for instance, a book title and the card catalog subject headings associated with it are two distinct access means that are the same size. Access is evidently increased by combining several systems belonging to one level, but if more than 5 such are combined, the resulting system belongs to the next higher level (at least) so it should be carefully considered whether the

the combination provides as effective and economical access as a structurally unified single access system belonging to the higher level. In this regard the problem of efficiently packaging diverse access systems to form one conglomerate system is critical. An example will help to clarify the importance of this observation. The traditional library card catalog system is a combinative system which includes class number, title, author, subject heading, and some additional information. The catalog card system occupies approximately 1/30th the volume of the library and therefore should function as a first level access system; unfortunately, the conglomerate catalog card only contains information corresponding to a second level access system. Inappropriate selection and packaging of access information has led to a situation where libraries traditionally and almost instinctively pay for a first level access system but obtain only a second level system. Such undesirable consequences of ignoring the level structure of access systems must be borne in mind in the design of an ISMS.

It is instructive to examine the recent collection of abstracts from *Mathematical Reviews* in fields related to topology compiled by Steenrod [2] in terms of the access model presented above. This collection contains approximately 6,400 abstract entries. First level access to the papers abstracted is provided by the abstracts themselves. Second level access could be provided by a list of the titles (and associated bibliographical information to make retrieval possible) of the abstracted papers. Third level access could be and is provided by distributing the titles (hence papers) among classificatory headings. The model implies that $6400/K = 217$ classes would be appropriate; Steenrod has 290. Fourth level access could be provided by grouping the classificatory headings into major classes; from the model, $217/K = 7.4$, whereas Steenrod has 14 major headings.

If Steenrod's 290 classificatory headings are accepted as the currently appropriate number for topology, then $290/K = 10$ major headings would be sufficient and $290K = 8561$ papers could be classified by the system operating at the mean of its level. All of these estimates refer to means; if distribution boundaries are considered, then one finds that 6,400 should correspond to between 217x5.4 and 217/5.4 classificatory headings and to between 7.4x5.4 and 7.4/5.4 major headings. Steenrod is neatly near the means.

Figure 9 displays the distribution of papers per class. Observe the generally lognormal nature of the distribution for that 80% of the collection consisting of the least populated classes, but note the significant underrepresentation of classes with large numbers of papers and the overrepresentation of classes containing between 17 and 25 papers. Steenrod's introductory remarks coupled with the access model provide a useful explanation for the variations which illustrates both the general applicability of the model and the importance of basing the design of an access system on a sensible model.

Steenrod states that he subdivided categories that initially contained more than 30 papers; no reason for the choice of 30 is stated. In addition to the overproduction of classificatory headings actually observed in the data, a consequence of this procedure is to replace classes containing from 30 to about 50 papers (very few classes with more than 50 papers remain) by pairs of classes of mean size from 15 to 25 papers, which results in the hump in the distribution shown in the figure as well as the underrepresentation of the more highly populated classes.

It is clear from the data that Steenrod did not really subdivide every class containing more than 30 papers; 25% of the final classes

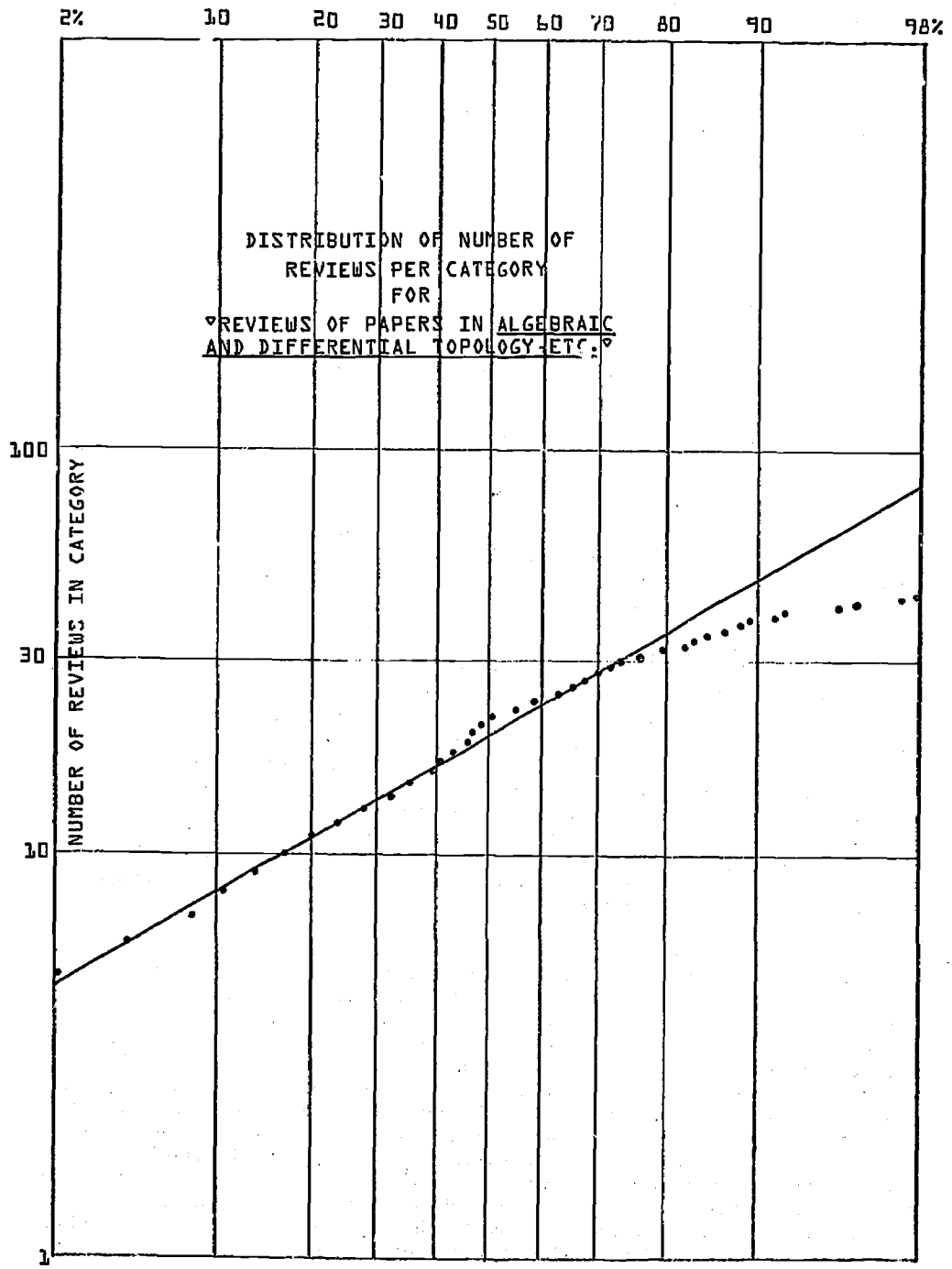

2%
10
20
30
40
50
60
70
80
90
98%
DISTRIBUTION OF NUMBER OF REVIEWS PER CATEGORY FOR REVIEWS OF PAPERS IN ALGEBRAIC AND DIFFERENTIAL TOPOLOGY-ETC.
NUMBER OF REVIEWS IN CATEGORY
100
30
10
1

FIGURE 9

contain more than 30, and one class contains as many as 55. One is therefore led to question the utilitarian purpose effected by this purely conventional and only partly implemented decision; it is clear that it cost the compiler some effort which possibly might have been better expended.

Steenrod's compilation is unusually useful; other similar efforts should be undertaken (some are now under way). The purpose of our discussion here is simply to provide a reasonable basis for asserting:

> It is not easy to deviate substantially from the structures defined by the access model;

and

> It is not worthwhile to deviate substantially from the structures defined by the access model; nothing is gained for the effort expended.

PART II

6. We open this part by stressing the inevitability of automation in the processing of information. This applies with particular emphasis to an ISMS. The following sections discuss some proposed ISMS access systems from the standpoint of their automatability and their connection with the publication process.

It is by now well known that the general archive and its special subsets such as the sciences and mathematics are growing exponentially with time ([1], [4], [8], Figure 1 above, and other sources too numerous to mention). This growth appears to be tied to the gross national product indicator (more accurately: gross world product) trend growth, with fluctuations and systematic variations which may be of great importance and size for certain special fields. It is essential to understand that this growth rate is, and for several hundred years has been, in general significantly larger than the rate of population growth. It follows that in an era of ever-increasing personnel costs it will not be possible to process the information produced by utilizing manual methods in a timely fashion and at reasonable--or at least stable--cost.

There is now no viable argument that suggests that the *trend* growth rate for the primary archive in the mathematical sciences will decrease within the next 50 years. From this remark it follows that automation of the automatable aspects of the archival publication, processing, and access systems is inevitable because the unit costs of computation are *decreasing* exponentially with time at such a rate as to offset the increase in the size of the archive itself and the associated personnel cost; Figure 10 displays how recent costs of computation have changed with time. Conversely, the non-automatable current methods of processing information will inevitably disappear as they one-by-one become too burdensome, costly, and inefficient as users of the limited amount of available human effort. Hence it is important to maintain an ongoing program of research and pilot-project experimentation to successively and successfully automate the important existing information processing functions and to devise replacements for those that cannot be successfully or economically automated.

In this regard it is worthwhile to examine the nature of abstracts of scientific papers more closely. We have already observed that an abstract

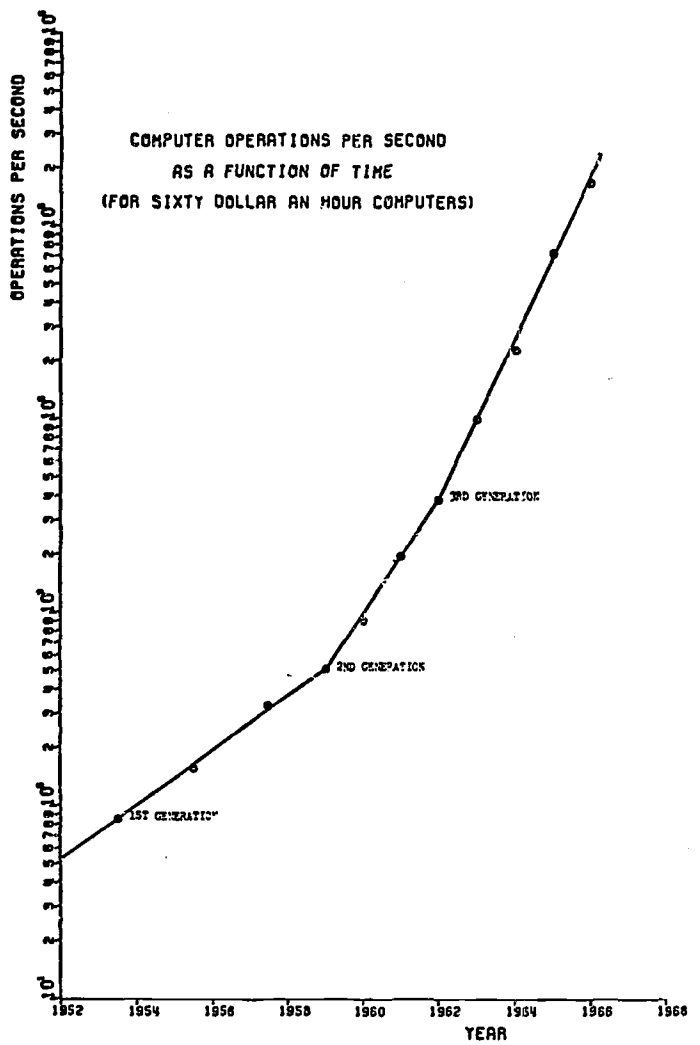


FIGURE 10

is a first-level access system. Abstracts are always written by trained personnel; however, the degree of competence of the personnel is neither uniform nor assessable by the abstract user. Abstract preparation is costly, and, as the ratio of partly machine-authored archival information to total information increases, abstractors will increasingly come to abstract machine production at the expense of their own creative production. The direct labor costs of abstract preparation are increasing. We cannot see a long future for the abstract as currently prepared by, e.g., *Mathematical Reviews*, although it is now the only available first-level access system covering any substantial portion of the archive in the mathematical sciences.

There does not now seem to be much hope for automating the production of abstracts. The only automated schemes which stand a chance of being

economically viable and of any use at all must extract information from the primary text. This assumes that there are word sequences in the text which could be extracted and then woven together to produce something which contains the essential information available in an abstract, and it is probably possible. The difficulty, which we think is insurmountable, is that the resultant product will in general belong to the same size level as the text itself although it will not contain more information than a first level access system. This, because the text redundancy which is usually eliminated by the abstractor and which accounts for as much as 75% of the non-symbolic text, cannot now and is unlikely to ever be excizable and reconstitutable in an economical way. This does not mean that an abstract journal would not be of significant temporary use in such fields as applied mathematics which currently lack any first-level access system; but, abstracts of the type discussed just cannot be viewed as a long range solution to the first-level access problem.

It has been proposed that authors supply abstracts with papers, and some journals currently require this. There are several defects associated with this process which may nevertheless have significant value as a first-level access system. One of the most easily remedied problems concerns the size limitations imposed on the abstract. In most cases *abstracts should be requested to occupy space approximately equal to 1/30th of the primary text of the article*. A substantially smaller abstract cannot in general adequately reflect the content of the paper and will therefore function as an access system belonging to some lower level, which will bury the paper. We think that if the paper is worth publishing, then it deserves representation in the access systems at the proper level.

A second defect is more subtle and ultimately of greater import. An author may be the last person to recognize what if anything is of real value in his paper; a competent reviewer probably does better although he cannot foresee what use future generations may wish to make of the content. It would be worthwhile to study the systematic variations, if any, that exist between author-composed and reviewer-composed abstracts.

A final remark in this regard: in order that author-composed abstracts provide an adequate substitute for services such as that currently provided by *Mathematical Reviews* it will be necessary to have a common international policy on author abstracts adhered to by all of the major journals.

7. There are not now any systems of first-level access to journal papers other than abstracts, but others are possible, and one of the goals of an ISMS development should be to define and evaluate them.

We will describe one possible alternative. It was noted above that the book index is a first-level access system to the text of the book. Back-of-the-book type indexes clearly can be constructed for papers. This is not normally done in the mathematical sciences or other fields, probably because of the great effort required and because there is no systematically published joint index compilation, but this situation can now be changed because papers can be indexed automatically. Salton [12] has recently described the state of the art of document retrieval using key-word type methods. Here we will describe an alternative procedure which does not rely on key word lists and therefore is in principle capable of identifying new significant information in a timely fashion. One elementary algorithm of this type, designed for application to general text (and whose effect on such input is presented as the index to [4] can be adapted for use on the non-symbolic portion of the text of mathematical papers. The result of applying this algorithm (by

hand simulation) to the text of a 12-page paper from the *Transactions of the American Mathematical Society* is shown in Figure 11; it is inadequate but promising. Two encouraging facts are that this algorithm could provide indexes as a byproduct operation on the data tapes used as input to a photo- or electronic composition system at low additional cost, and that the index produced by the algorithm is of just the proper size.

Some of the defects obvious in Figure 11 could be eliminated by adapting this completely general text algorithm to the peculiarities of the vocabulary structure of the mathematical sciences, but this is of secondary importance. Of major significance is the inclusion in the algorithmic index output of all of the pertinent terms which indicate the principal concerns and tools used. Diligent improvement of this primitive system or some alternative automatic indexing scheme should lead to the production of a first-level access system for inclusion in an ISMS which would operate as a byproduct function of the publication process. As was the case for abstracts, the impact of foreign publication and the possibilities for cooperative efforts must be explored.

Although the first-level access provided by index terms is inferior to that provided by abstracts in certain respects, index term access systems have some advantages not available to abstracts. It is possible, for instance, to amalgamate indexes to a number of papers or books and thereby provide specific access to all items containing a specified term or terms, some of which would not normally appear in the title or in an abstract of the paper.

This remark leads to the further observation that it may be of some utility to index abstracts and then amalgamate the index terms together with their bibliographical references. A primitive experiment of this type [10] has been performed on a sample of 50 abstracts taken from the *Annals of Mathematical Statistics*; Figure 12 exhibits one abstract and the index to it that results from an application of the primitive algorithm to the text of the abstract. With the exception of symbolic material, which was excised from the abstract and replaced by the symbol "*" , the index entries are representative of the subjects discussed in the abstract. Figure 13 exhibits part of the amalgamated index for all 50 abstracts. It would appear to be profitable to pursue this direction of increasing access at the second level.

8. Previous sections of this proposal have argued that an ISMS must be as automated as possible and that the data base upon which it operates must be compatible with other interfacing information systems. These requirements imply that the data base must be in machinable form.

There are liabilities as well as advantages to large scale machinable data bases; the costs of acquisition, storage, and processing are high and can normally only be justified if some form of generally distributable hard copy is a product of processing the data base. Interactive on-line information access systems can therefore only be practical if they operate on a data base that is several levels below the primary archive; the recent study of R. L. Van Horn [13] is implicitly of this type. Regarding data base acquisition, a significant fraction of the current primary publication in the United States of materials in the mathematical sciences could be obtained as a byproduct of the composition process; this problem is discussed in section 9. Because of the rapid growth of the primary archive, the problems of retrospective conversion which may now appear to be enormous will impose a decreasing burden on the system as time passes. Therefore a

PROTO-INDEX TERMS TO

"On differential operators and automorphic forms"
Trans. Amer. Math. Soc. 124 (1966), 334-346

	Term	Text Location Multiplicity
	automorphic form	12
	discrete groups	6
	algebraic differential equation	5
	horocyclic group	5
	upper half plane	5
*	algebraically dependent	4
	automorphic function	4
	classical modular group	4
	monomial terms	4
	fundamental domain	3
	meromorphic functions	3
	multiplier system	3
	rational function	3
	analytic continuation	2
	compact real Jordan algebra	2
	constant multiple	2
	differential operators	2
	explicit expression	2
	Fourier expansion	2
	full group of analytic automorphisms	2
	graded ring of modular forms	2
	linear combination	2
	local uniformizing parameter	2
	necessary condition	2
	negative integer	2
	nonnegative integers	2
	nonpositive integer	2
	nonzero constant	2
	normalized cusp form	2
**	order differential equation	2
	positive integer	2
	positive real number	2

****

*) "ly#" deletion suppressed.

**) One instance of "third order differential equation", and one of "fourth order differential equation".

FIGURE 11

34-0353

D. G. KABE, WAYNE STATE UNIVERSITY. (INTRODUCED BY A. T. BHARUCHA-REID)
SOME APPLICATIONS OF A RESULT OF HERZ'S TO NONCENTRAL MULTIVARIATE
DISTRIBUTION PROBLEMS IN STATISTICS.

IN PREVIOUS PAPERS (KABE, D. G., SOME RESULTS ON THE DISTRIBUTION OF TWO RANDOM MATRICES USED IN CLASSIFICATION PROCEDURES, ANN. MATH. STATIST. (TO APPEAR)), AND (KABE, D. G., ON THE DISTRIBUTION OF THE LATENT ROOTS OF THE WISHART MATRICES, (SUBMITTED FOR PUBLICATION)), THE AUTHOR SHOWS THAT SOME NONCENTRAL DISTRIBUTION PROBLEMS IN MULTIVARIATE STATISTICS DEPEND ON THE EVALUATION OF THE MOMENTS OF THE NONCENTRAL GENERALIZED VARIANCE. IN THIS PAPER THE AUTHOR USES A RESULT OF HERZ'S (HERZ, CARL S., BESSEL FUNCTIONS OF MATRIX ARGUMENT. ANN. OF MATH. 61 (1955) 474-523) TO GIVE FORMAL SOLUTIONS OF THE NONCENTRAL DISTRIBUTIONS OF THE ROOTS OF A SINGLE WISHART MATRIX, THE NONCENTRAL DISTRIBUTIONS OF THE ROOTS OF TWO WISHART MATRICES, AND THE NONCENTRAL MULTIVARIATE BETA DISTRIBUTION.

RANDOM MATRICES
CLASSIFICATION PROCEDURES
LATENT ROOTS
WISHART MATRICES
NONCENTRAL DISTRIBUTION PROBLEMS IN MULTIVARIATE STATISTICS
MULTIVARIATE STATISTICS, NONCENTRAL DISTRIBUTION PROBLEMS IN
AUTHOR USES
BESSEL FUNCTIONS OF MATRIX ARGUMENT
MATRIX ARGUMENT, BESSEL FUNCTIONS OF
FORMAL SOLUTIONS
NONCENTRAL DISTRIBUTIONS
SINGLE WISHART MATRIX
NONCENTRAL DISTRIBUTIONS
WISHART MATRICES
NONCENTRAL MULTIVARIATE BETA DISTRIBUTION

FIGURE 12. AUTOMATIC INDEX FROM TEXT OF ABSTRACT

project of conversion of selected portions of the retrospective primary information base to machinable form which systematically and uniformly works its way back through time, recording only that data necessary for specific lower access levels, should form part of an ISMS. Steps in this direction are already under way, and should be extended. For example, the Citation Index Project under John Tukey's direction has already compiled a cumulative list of titles of articles in statistics which includes more than 25,000 papers and most of the retrospective material. It is a remarkably powerful and efficient access system when presented in permuted form, highly superior to non-cumulative listings such as the bi-weekly *Chemical Titles*. (but note that the annual production of papers in chemistry is at least 100 times greater than that of the mathematics represented in *Mathematical Reviews*).

9. Material currently processed for publication is converted to machinable form at least twice in many publication systems. If portions of it (such as titles and other bibliographical information; abstracts provided

FIGURE 13. CUMULATED AUTOMATIC INDEX FROM TEXT OF ABSTRACT

by authors; extractive index terms, etc.) are required for information retrieval systems, they are normally re-keyboarded yet again.

It seems clear that significant cost reductions in the publication process and in all access system operations that utilize the primary text material in some way could be achieved if multiple keyboarding were avoided. Savings will be greatest if primary material is keyboarded but once. This implies that preparation of the original typescript should result in a machinable version as well. The machinable version--contained perhaps in a magnetic tape cassette--could be sent with the typescript to journal editorial offices; editorial modifications would be performed using a display screen and contact-pen terminal. Composition, utilizing either mask-generated characters (e.g., Photon or Linotron systems) or the preferable digitally-stored character generating techniques (Videocomp), would proceed without further human intervention. Since there would be no keyboarding of material after the author-stage keyboarding (which includes of course necessary options for author corrections and required manuscript revisions), there would be no introduction of typesetting errors with a consequent decrease in typesetting and proofreading costs and in the interval between acceptance of a manuscript and its appearance.

One of the goals of the ISMS development program herein discussed should be to study the economics of such an automated composition system. With the use of a common electronic composition machine, common computer software, and (possibly) a specially adapted digital character generator capable of producing the large symbol font typically required in mathematical publication, it should be possible to achieve a reduction in page composition costs.

REFERENCES

[1] de Solla Price, D. J.: Science Since Babylon, Yale University Press, New Haven, 1961.

[2] Steenrod, N.: Reviews of Papers in Algebraic and Differential Topology, Topological Groups and Homological Algebra, American Mathematical Society, Providence, 1968.

[3] Resnikoff, H. L. and J. L. Dolby: "Access", a report prepared for the U.S. Office of Education, R&D Consultants Company, January 1971.

[4] Dolby, J. L., V. Forsyth, and H. L. Resnikoff: Computerized Library Catalogs: Their Growth, Cost, and Utility, The M.I.T. Press, Cambridge, 1969.

[5] Dolby, J. L. and W. E. Houchin: A Preliminary Reference Manual for the ALTEXT Macro Language, R&D Consultants Company, 1970.

[6] Miller, George A.: "The magical number seven, plus or minus two: some limits on our capacity for processing information", Psychological Review 63 (1956), 81-97.

[7] Houston, Nona and Eugene Wall: "The distribution of term usage in manipulative indexes", American Documentation (April 1964), 105-114.

[8] Beaver, Donald DeB.: "A statistical study of scientific and technical journals", preprint, Yale University Department of History of Science and Medicine, 1964.

[9] Abramowitz, Milton and Irene A. Stegun: Handbook of Mathematical Tables, U.S. Department of Commerce, Government Printing Office, 1964.

[10] Dolby, J. L. and W. E. Houchin: "An experiment to measure the efficiency of a book-indexing algorithm when applied to abstracts", a report prepared for the Princeton University STRG Project, R&D Consultants Company, July 1970.

[11] American Mathematical Society: Final Report: Conference on Communication Problems in the Mathematical Sciences, December 5-7, 1967, Providence, 1969.

[12] Salton, G.: "Automatic text analysis", Science 168 (17 April 1970), 335-343.

[13] Van Horn, Richard L. and H. B. Back: "A system to improve the availability and usefulness of mathematical knowledge", a report prepared for the Conference Board of the Mathematical Sciences, Carnegie-Mellon University, January 1971.

[14] Mandelbrot, B.: "An information theory of the statistical structure of language", Proceedings of the Symposium on Applications of Communication Theory, Butterworth, 1953.

[15] Resnikoff, H. L.: "On information storage and retrieval systems", Proceedings of the 1970 TIMS Management Systems Conference, 124-138.