DOCUMENTATION
INCORPORATED

9521 CONNECTICUT AVENUE, N. W.
WASHINGTON 8, D. C.
COLUMBIA 5-4577

Errata
AD
43291

ERRATA

Technical Report No. 6. "The Mechanization of Coordinate Indexing" prepared under Contract Nonr 1305(00) for the Office of Naval Research, September 1954:

... ~~substitute attached~~ sheet for p. 15 and 16
(For File Copy 1st 3 only)

... page 5, paragraph 2 should read "dedicated positions 1/16" x 1/16"," rather than 1/16'

... eliminate page 19

descriptors. In fact a question may be asked under one term; and it is our experience that the great bulk of questions involve 2 or 3 terms and that additional terms are added only if the amount of information delivered by a 2 or 3 termed question proves to be too voluminous. In other words, actual reference experience and not mathematical considerations alone must influence the design of a system. A system of coordinate indexing which delivers too high a percentage of "noise" for one, two or three term questions is simply not adequate for reference use.

In the Zatocoding system random numbers correspond to the language elements in the indexing system we have described. Every term or subject is represented by a random group of four numbers chosen from a field of forty numbers, e. g.:

air	3	12	17	23
ducts	6	17	24	29
icing	11	12	23	27

We assume that for any item indexed, no more than 1/2 of the coding area is used. (This figure corresponds to our use of 1/3 of the dedicated positions in our machine development.)

The formula (approx.) given by Calvin Mooers for the dropping fraction in any search is $1/2^n$ where n equals the number of notches or holes used in a search. For example, if we search a collection for everything on Air, n = 4 (3, 12, 17, 23); for everything on Air Ducts, n = 7 (3, 12, 17, 23, 6, 24, 29); for everything on Air Ducts Icing, n = 9 (3, 12, 17, 23, 6, 24, 29, 11, 27).

n	4	5	6	7	8	9	[illegible]	11	[illegible]
$1/2^n$	1/16	1/32	1/64	1/128	1/256	1/512	[illegible]	1/2048	1/4096

It is only necessary to multiply the dropping fraction by the size of any collection to determine the approximate average number of false drops per search. This table shows the size of the dropping fraction for various sizes of n.

		SIZE OF COLLECTION			
n	Dropping Fraction	1,000	10,000	100,000	250,000
4	1/16	62.5	625.0	6,250	15,625
5	1/32	31.25	312.5	3,125	7,813
6	1/64	15.63	156.3	1,563	3,908
7	1/128	7.82	78.2	782	1,955
8	1/256	3.91	39.1	391	978
9	1/512	1.96	19.6	196	490
10	1/1024	0.[illegible]	9.8	[illegible]	[illegible]
11	1/2048	0.[illegible]	[illegible]	49	123
12	1/4096	[illegible]	2.5	25	63

If, as is the case, the average reference question involves the coordination of two to three terms, the average number of false drops in a Zatocoding system containing 100,000 items will range from 3,125 items (when n = 5, the minimum number of notches used for 2 terms) to 25 items (when n = 12, the maximum number of notches which can be used for 3 terms.)

In the mechanized system we are developing, for a search by 2 terms, we are planning a dropping fraction of 1/320. This means that for all practical purposes there

THE MECHANIZATION OF COORDINATE INDEXING

TECHNICAL REPORT NO. 6

PREPARED UNDER
Contract Nonr-1305(00)

FOR THE OFFICE OF NAVAL RESEARCH

September
1954

DOCUMENTATION
INCORPORATED
2521 CONNECTICUT AVENUE, N. W.
WASHINGTON 8, D. C.

TECHNICAL REPORT NO. 6

THE MECHANIZATION OF COORDINATE INDEXING

In the summer of 1952, the Armed Services Technical Information Agency awarded a contract to Documentation Incorporated for an investigation and experimental installation of a then completely novel system for filing and retrieval of information, the Uniterm System of Coordinate Indexing.

This manual system of storage and retrieval of indexing information as conceived by Documentation Incorporated has since been developed and refined in a series of test programs and has been adopted by numerous government agencies and industrial organizations.

One of the reasons this system has found such ready acceptance is that it combines high efficiency and specificity of information retrieval, compactness and low cost, and most important, the real possibility of mechanization. Although, for a system of limited size, the manual Uniterm System of Coordinate Indexing far exceeds any presently available machine system in efficiency,* it was realized that larger systems would still require a form of mechanization.

In the early part of 1953, certain pre-existing machine designs for data retrieval were combined with the logic of the Uniterm System of Coordinate Indexing, resulting in the conception of the indexing machine described herein.

---

*"Efficiency" defined as a combination of cost and effectiveness of information retrieval.

Shortly after its conception, this system came to the attention of the Armed Services Technical Information Agency, which requested Documentation Incorporated to make an investigation of the logic of this machine and of some of the proposed constructional features to determine its suitability to Armed Services Technical Information Agency's particular requirements. This investigation has now been completed, and we are pleased to report that this system of instantaneous information filing and retrieval exceeds existing and contemplated machine systems in efficiency to such an extent as to create an entirely new dimension in cost and retrieval time.

Until now, systems of sequential scanning of information such as the I. B. M. and Remington Rand systems, have completely monopolized all machine systems as well as all thinking in the information field. While such machines are excellent for tabulations, computations, and, in general, bookkeeping purposes in the widest sense of the word, they are not especially suited for information storage and retrieval.

Only a system of simultaneous scanning of superimposed data storage sheets, corresponding to natural language elements, as proposed, will provide low cost instantaneous filing as well as retrieval of information. It is our conviction that this machine will be a major and history-making step in the direction of adequate information control.

Coordinate indexing as a generic term covers all forms of indexing in which the retrieval of specific items of information involves the determination of the logical product of a number of classes. The Uniterm System as a species of coordinate indexing is a manual method of determining the logical product of two or more classes through the device of "arithmetical" coordination. The discovery of a common number on two or more Uniterm cards establishes that there is a class which is the logical product of the classes denoted by the Uniterms and that the class has members. The members are, of course, the documents or other items designated by the common numbers.

One advantage of the Uniterm System's device of "arithmetical" coordination is that numbers can be written on the card as the numbered items are received and indexed. It is not necessary to dedicate in advance any portion of the card to any number. There is, however, another manual method of Coordinate Indexing, known as the Batten system or the Peek-a-Boo system, which determines the logical product by the coincidence of identically numbered spaces rather than the mere identity of numbers. In the Uniterm System, even though some use is made of geometry (the arrangement of numbers in ten columns) a number at the head of a column on one card may be identical with a number at the foot of a column on another card. But, as we can note on Exhibit I in the Batten system, the position of each number must be preassigned and must be in the same place on each card. This preassignment of numbers to dedicated positions gives the Batten system one important and obvious advantage over the Uniterm System, in that coordination becomes a simple rapid process involving as many items as are desired; whereas in the Uniterm System numbers must be matched for any two terms before going on to a

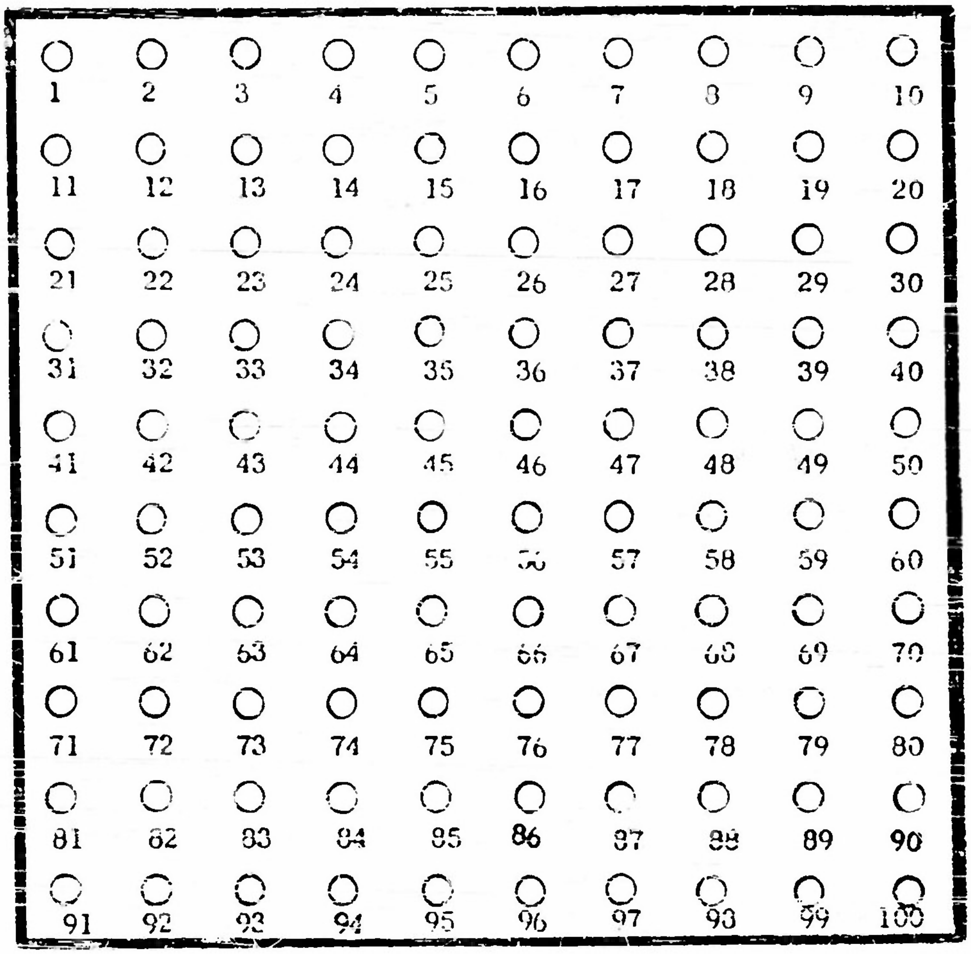

EXHIBIT I

Batten Card for 100 Items

third, or fourth term, etc. On the other hand, the use of dedicated positions limits drastically the capacity of the Batten system as compared with the Uniterm System. As will be seen from Exhibit I, by the time item 101 enters our system, we are beyond the capacity of our Batten cards and must set up an entire new set. And this may occur with only an average use of .01% or less of the dedicated positions on the first set.

The Office of Basic Instrumentation of the National Bureau of Standards is experimenting with Batten cards having a capacity of 18,000 positions. An interesting and important account [1] of this experiment has appeared recently, but this account does not reveal any information concerning the average or mean density of use of the dedicated positions. We, ourselves, lacking the evidence, which we hope will be forthcoming from additional reports on the Office of Basic Instrumentation project, felt that for large rapidly growing files of material the limitations imposed by the dedicated positions of a Batten system outweighed the apparent advantages of rapid coordination. These limitations can be expressed in terms of certain theoretical considerations together with certain evidence which is now available. We know, for example, that the indexing of 18,000 documents will involve a vocabulary in the range of 5,000 terms. We also know that the average document can be indexed using 8 to 10 terms.

---

[1] Documentation in Instrumentation, National Bureau of Standards Report 3276, by W. A. Wildhack, J. Stern, and J. Smith, April 1954.

Taking the higher figure, we obtain the total of 180, 000 postings for 18, 000 documents or 5, 000 Batten cards. This means an average utilization of $\frac{180,000}{5,000}$ or 36 holes on each card. Each Uniterm card (front and back) for a manual system has a capacity of 520 numbers if all columns are equally utilized. In actual practice, all columns will not be equal, but the percentage of difference between the longest column and the shortest column will decrease as cards fill up (Exhibit II makes this clear). It is, therefore, conservative to assume that we can get 400 numbers on a Uniterm card. With the same size vocabulary, 5, 000, and the same number of postings per document, our 5, 000 cards will provide for 2, 000, 000 postings or 200, 000 documents. In a Batten system having 18, 000 holes we would require $\frac{200,000}{18,000}$ or 10.2 sets of 5, 000 cards for 200, 000 documents; that is to say, 11 sets, since there can be no partial sets in a Batten system. Further, the advantage of rapid coordination in a single set, would be lost through the necessity of selecting cards out of 11 sets and of performing 11 separate coordinating operations.

It is obvious that one can cut down the number of sets in a Batten system by increasing the size of the card and the number of dedicated positions on each card; but such an increase in size does not increase the relative efficiency of the system since the percentage of dedicated positions used will remain constant.

In spite of the above considerations, we never lost

TERM ACCELEROMETERS

0	1	2	3	4	5	6	7	8	9
10	51	52	83	84	35	36	37	88	99
20	81	72	143	124	75	136	147	108	149
80	111	82			85			138	
120	141	132			145				
130	161								
200									

UNITERM card © DOCUMENTATION Incorporated

TERM

TERM ACCELEROMETERS

0	1	2	3	4	5	6	7	8	9
10	51	52	83	84	35	36	37	88	99
20	81	72	143	124	75	136	147	108	149
80	111	82	563	784	85	976	1037	138	849
120	141	132	973	974	145	1026	1047	1028	1029
130	161	972	1023	1024	975	1036	1057	1038	1039
200	971	1022	1033	1044	1025	1046	1067	1048	1049
210	981	1032	1043	1054	1035	1056	1077	1058	1059
240	1021	1042	1053	1064	1045	1066	1087	1068	1069
350	1031	1062	1063	1074	1055	1076	1097	1078	1079
	1041	1072	1073	1084	1065	1086	1107	1088	1089
	1051	1082	1083	1094	1075	1096	1147	1098	1099
	1061	1092	1093	1104	1085	1106	1267	1108	1109
	1071	1102	1103	1114	1095	1146	1287	2248	3829
	1081	1112	1113	1124	1105	2846	2297		

UNITERM card © DOCUMENTATION Incorporated

TERM

EXHIBIT 2

sight of the manifest advantages of geometric coordination as compared with arithmetic coordination, especially with respect to systems large enough to make mechanization feasible and desirable. For quite early in our program it seemed clear that the mechanization of arithmetic coordination (what is usually called collation[2]) is not a promising technique in information storage and retrieval. The machine comparison of two or more sets of numbers in order to find any numbers common to the sets, must in principle, be a sequential process, whereas the machine comparison of dedicated positions in two or more Batten cards can be relatively instantaneous.

As was said above, one apparent solution to the limitations of the Batten system involves increasing the size of the card; and if the card is to be positioned and scanned mechanically, only the size of available materials, the size of holes, and the problems of mechanical positioning affect the decision on this matter. In a series of tests we ascertained that a 3' by 3' metal or plastic sheet between 0.15" and 0.20" thick, which is not too big or too heavy for mechanical handling, will provide dedicated positions 1/16' x 1/16' for 250,000 items. Having made this determination, we were still faced with three problems:

1. No. of sheets:

   5,000 3" x 5" or 5" x 8" cards is a comparatively small file; but 5,000 3' x 3' metal sheets each

---

(2) "Machines and Classification in the Organization of Information", Technical Report No. 2, Documentation Incorporated, December 1953, p. 10 to 27.

of which is .020" thick, represent a cube of metal 3' x 3' x 8.3'. This seemed to carry us beyond the range of practicality.

2. Addition of New Sheets:
Even though the number of Uniterms in any system remains fairly constant and new Uniterms are added infrequently, we must provide for such additions. This means that in addition to our initial cube of metal 3' x 3' x 8.3', we must have a stock of extra 3' x 3' sheets for coding and adding to our basic set within the machine. This, too, we determined was impractical.

3. Use of Dedicated Area:
The recognition that our total coding space is a cube brings with it conclusive evidence that the Batten system, regardless of the number of positions on a card, is an inefficient mechanism for storing information. Of the metal cube, 0.2% would be used for storing information, the remainder would be potentially useful but actually unused space.

If we neglect, for a moment the problem of adding sheets, we can see that the problem of the nunber of sheets and the problem of the percentage of space used on each sheet are two sides of the same coin. For the same number of bits of information a higher density of use of each sheet would involve fewer sheets. The solution to this problem which is expressible in the same type of analysis as that which led to the concept of Uniterm, was first developed and patented by Dr. Frederick Jonker, who, besides being a member of the staff of

Documentation Incorporated, has organized his own business machine company. Dr. Jonker recognized that just as the ideas in any system of information could be expressed as logical products of terms, so any term can be expressed as a logical product of simpler elements. Further, just as the number of terms in any system is much less than the number of ideas, so the number of letters, as is obvious, is much less than the number of terms.

It will be said that a word is not a simple logical product of letters, but a product of letters in a certain order. Thus, tar is not the same word as rat, nor is dog the same word as God. But suppose, instead of one alphabet of 26 letters, we have an alphabet of three times this number or 78. In this artificial alphabet we will have three "A's"--A1, A2, A3; three "B's"--B1, B2, B3; three "C's"--C1, C2, C3, etc. With such an alphabet it is clear we can express any three letter word, without ambiguity, as a logical product of letters. For example: T1 A2 R3; R1 A2 T3; D1 O2 G3; G1 O2 D3. If we extend the number of letters in our alphabet in this fashion, or, what is the same thing, extend the number of alphabets in our system, we can express as logical products, longer and longer words. With 260 letters or 10 alphabets, for example, we can express every word up to 10 letters long not only in the English language, but in any language using the Roman alphabet.

Let us examine, now, the manner in which this development solves the three problems we have presented above:

1. No. of Sheets. Instead of 5,000 cards or sheets, that is, one for each Uniterm, we now need only 26 times the number of alphabets. For ease of multiplication in what follows, suppose we assume 10 alphabets and round off the size of system at 300 sheets (10 x 26 + 40 for symbols, Greek letters, and other devices).
2. Addition of New Sheets. It should be apparent, that like the letters in the alphabet or the digits in the number system, our system is constant and no new sheets will ever be required because of the addition of new ideas or terms to the system.
3. Use of Dedicated Area. On simple mathematical grounds it is clear that the same amount of information, stored in a smaller number of sheets, must operate to increase the density of use of each sheet. The exact manner in which this comes about can be seen from a comparison with the manner in which a manual Uniterm system stores information, which is dispersed in a standard library system.

Suppose that in a standard indexing system there are three headings:

air ducts
air blowers
air cleaners

In the Uniterm System all items on air ducts, all items on air blowers, and all items on air cleaners, are entered on the Air Uniterm card.

Similarly in the letter system, all items on accuracy, airplane, acid, accelerometers, etc., will be entered on the A1 card or sheet; all items on acid, accuracy, accelerometers, etc., will be entered on the C2 sheet; accuracy and accelerometer will be entered on the C3 sheet; airplane and accelerometer will be entered on the L5 sheet, etc.

It will be recalled that with the Batten system, with a vocabulary of 5,000 terms with each item indexed by 10 terms, the average use of dedicated positions was 0.2 of 1 percent. For a system of 18,000 items or 18,000 dedicated positions, 0.2 percent would be 36 per sheet. For our 3' x 3' sheets with dedicated positions for 250,000 items, 0.2 percent would be 500.

With the letter system, if we assume an equal or random use of all letters, and the same depth of indexing, we would use one-third of the dedicated positions on each sheet. There are three ways in which we arrive at this figure;

1. For the same number of items of information and the same depth of indexing, the density of postings varies inversely as the number of sheets, thus $\frac{5,000 \text{ sheets}}{300 \text{ sheets}} = \frac{x \text{ postings}}{500 \text{ postings}} = \frac{25,000}{3}$. Multiplying this figure by 10, to provide for 10 postings per term in the letter system, we have $\frac{250,000}{3} = 33\ 1/3$ of the possible positions posted on each sheet.

2. If there are 300 sheets in the system and each sheet contains 250,000 positions, we have a total number of 75,000,000 storage positions, in the system.

If each item is indexed by ten terms, each term being 10 letters long, this would involve using 100 holes per item. Thus, the indexing of 250,000 items will utilize 100 x 250,000 or 25,000,000 positions, which again is one-third of the possible positions in the total system.

3. Suppose our 300 cards represent 10 alphabets of 30 letters. Then any document indexed by 10 terms will utilize under random distribution, one-third of the letters in each alphabet for each document or dedicated position.

Hence, for 250,000 documents one-third of the positions will be used on all the sheets.

It should now be clear that the use of language elements rather than Uniterms solves the three problems of mechanization, namely, the size of the file, the permanence of the file and the efficient use of coding area or dedicated positions. We have, as we shall see below, introduced a new problem which is not present in the Uniterm System or the Batten system.

The logical product achieved in a Uniterm or Batten system may be ambiguous. Thus, if we combine fish and food, we will get information on food for fishes and fish as a food. This type of ambiguity can usually be resolved by adding another term as follows: fish-food-vitamin or fish-food-plankton. "Noise" or information which has no relevence to the terms coordinated can never appear in the logical product presented by a Uniterm or Batten system. But when we go from Uniterms to language elements, we, in effect, go from a system of direct coding to superimposed coding and this introduces the problem of "noise" or false drops in the information system.

We have effectively solved this problem in our machine development so that noise or irrelevant information delivered by the machine in answering any question, although possible in principle, is for all practical purposes eliminated. The exact nature of our solution involves certain mathematical and mechanical considerations which we will not present at this time. But we can illustrate the problem as it occurs in a simpler system, the well known Zatocoding system, developed by Calvin Mooers.

In any indexing system we must make a distinction between the number of headings or entries by which an item is indexed and the number of headings used either singly or in combination to retrieve any item of information. In a standard subject-cataloging system, or a system of indexing like the Index to Chemical Abstracts, we may enter items under a number of different headings. But we always search for information under one heading at a time. If the material we retrieve under any heading is not sufficient or adequate to our purposes, we may try another heading, which may be broader, narrower, or just different from the heading first used. But in no sense is such a search a search for material indexed under a combination of headings. Zatocoding is a species of coordinate indexing. This means that in Zatocoding items are indexed under a group of headings and searched for under any heading or any combination of headings. But the fact that an item is indexed under 10 terms or descriptors does not mean that questions put to the system should or do consist of 10 terms or descriptors. In fact a question may be asked under one term; and it is our experience that the great bulk of questions involve 2 or 3 terms and that additional terms are added only if the amount of information delivered by a 2 or 3 termed question proves to be too voluminous. In other words, actual reference experience and not mathematical considerations alone must influence the design of a system. A system of coordinate indexing which delivers too high a percentage of "noise" for one, two or three term questions is simply not adequate for reference use.

In the Zatocoding system random numbers correspond to the language elements in the indexing system we have described. Every term or subject is represented by a random group of four numbers chosen from a field of forty numbers, e. g.:

air	3	12	17	23
ducts	6	17	24	29
icing	11	12	23	27

We assume that for any item indexed, no more than 1/2 of the coding area is used. (This figure corresponds to our use of 1/3 of the dedicated positions in our machine development.)

The formula (approx.) given by Calvin Mooers for the dropping fraction in any search is $1/2^n$ where n equals the number of notches or holes used in a search. For example, if we search a collection for everything on <u>Air</u>, n = 4 (3, 12, 17, 23); for everything on <u>Air Ducts</u>, n = 7 (3, 12, 17, 23, 6, 24, 29); for everything on <u>Air Ducts Icing</u>, n = 9 (3, 12, 17, 23, 6, 24, 29, 11, 27).

n	4	5	6	7	8	9	10	11	12
$1/2^n$	1/16	1/32	1/64	1/128	1/256	1/512	1/1024	1/2048	1/4096

It is only necessary to multiply the dropping fraction by the size of any collection to determine the approximate average number of false drops per search. This table shows the size of the dropping fraction for various sizes of n.

		SIZE OF COLLECTION			
n	Dropping Fraction	1,000	10,000	100,000	250,000
4	1/16	62.5	625.0	6,250	15,625
5	1/32	31.25	312.5	3,125	7,813
6	1/64	15.63	156.3	1,563	3,908
7	1/128	7.82	78.2	782	1,955
8	1/256	3.91	39.1	391	978
9	1/512	1.96	19.6	196	490
10	1/1024	0.98	9.8	98	245
11	1/2048	0.49	4.9	49	123
12	1/4096	0.25	2.5	25	63

If, as is the case, the average reference question involves the coordination of two to three terms, the average number of false drops in a Zatocoding system containing 100,000 items will range from 3,125 items (when n = 5, the minimum number of notches used for 2 terms) to 25 items (when n = 12, the maximum number of notches which can be used for 3 terms).

In the mechanized system we are developing, for a search by 2 terms, we are planning a dropping fraction of $1/3^{20}$. This means that for all practical purposes there will be no false drops or noise in a system of 250,000 items.

We present in the following, a summary statement of the characteristics of the machine development which are presumed to be of general interest:

Nature of the Development: The proposed machine is based on proven and tested principles and does not contain any speculative elements.

Operation of the Machine: In entering information - merely typing out the indexing terms or phrases on a keyboard; in searching information - merely typing out the search words or phrases on this same keyboard. The document number of the relevent documents* will be printed on a paper tape.

Proposed Document* Storage Capacity Per Machine: 250,000 documents per unit. However, units of double this capacity are feasible.

Ultimate Total Document* Storage Capacity: Unlimited. A number of machines can be operated simultaneously from a single control console.

Information Capacity Per Document*: About 10 descriptive words or phrases per document is at present contemplated. This is equivalent to approximately 2×10^8 digital units of information. This number can be increased if necessary.

---

*"Document" should be understood as an article or item on which information is to be stored. "Documents" can be books, reports, medical case records, persons, criminals, patents, court decisions, articles of a supply catalog, etc.

Vocabul... C...acity of the System: The system ... poten... unlimited capacity of words ... etc., in any language and using any type of ... such as ... bic and Roman numerals, Greek ... etc.

Coding Requirements: None; actual language ... used as machine code. No code conversio...

Storage Spee... ... fying information on a keyboard. Followi... storage is virtually instantaneous.

Searching Speed: Following typing of search ... matter of a few seconds per document yie...

Nature of Storage Elements: Large metal she... most permanent, durable type of storage.

Machine Dimensions: Approximately 7' x 4' x ... is about 1/20 of volumetric space occupied ... valent conventional catalog card drawers.

Assembled Weight: 1,000 lbs. giving approxim... 30 lbs. per sq. ft. floorloading.

Nature of Techniques Used: Entirely mechanic... electromechanical. No electronic element... ... reliability.

Reproduction and ... of Stored Infor... Stored information can be reproduced and ... economically by means of photo-etching ...

Cost of the Machine When Manufactured in S... ... 1,000,000 documents the cost of the ... be as low as the cost of the filing cabine... index cards plus the index cards, requi... ventional card catalog system.

UNCLASSIFIED

UNCLASSIFIED