have been gathered together with great labor and expense, from one library to another.

From time to time a professor who has built up a great subject collection in one institution transfers to another and will then, if unable to bring his special collection with him, want to build up a second one even if most of the books concerned will very rarely be used. One of the questions that the future must face is whether collections that fall in this category can be transferred bodily from one institution to another as the demand for the books shifts. And, carrying the problem one step farther, can university authorities get together and say that only one will do the advanced work in certain limited fields and that others will refrain from duplicating that work? Would this be restraint in trade?

Another question which libraries have not yet faced squarely is whether a library which takes the responsibility for a limited field, spends the money to buy the books, catalogue them and serve them, can be expected to make them available without charge to scholars who have no connection with it, when its own students pay high tuition before they are allowed to use the library.

Whatever may come out of the situation described in the above paragraph, it must be realized that coöperation among libraries at a distance will always be confined largely to books that are little used or very expensive, and this expense must be considered to include not only the cost of purchasing the material, but the cost of cataloguing it, storing it, and making it readily available. Each of these four types of expenditure is important.

No attempt will be made here to consider the place of new methods of reproducing printed materials in coöperative development of research collections, but that place is sure to be of growing importance.

Finally, the point should be made that books that are used frequently should be in even small college libraries. The same is true for many reference works and for standard bibliographies which are of use in learning of the existence of material and, if possible, in determining the library in which it is located. And this leads us to say again that coöperative acquisition should in most cases deal with material that is used almost altogether by advanced scholars in their original research or for graduate students preparing their doctoral dissertations. These are the books and pamphlets that make the difference between the good college and the great university library collection, that make necessary the tremendous staffs for cataloguing and public service and the library buildings that at present prices may cost $5,000,000 to $10,000,000 or more to construct. It is the costs involved directly or indirectly in these books and pamphlets that have made coöperative acquisition desirable, will undoubtedly make it necessary, and should make it more satisfactory as the years roll by.

ZATOCODING APPLIED TO MECHANICAL ORGANIZATION OF KNOWLEDGE

CALVIN N. MOOERS*

ABSTRACT

The mechanical organization of knowledge for retrieval of stored information can no longer neglect the developments of point-to-point communication theory, since both deal with information and handle it by machines. The most versatile retrieval systems are those which delegate a separate tally to each information item, and which impress marks on the tally for the machine to read and to use for selective purposes. Coding is the relationship between these marks and the intellec-

*President, Zator Company, Boston, Mass.

tual content of the information items. Coding determines the complexity of the selective machine and the utility of the whole process. A set of invariant coding principles is stated which define maximum coding efficiency for any tally selecting machines, and parallels are drawn between these principles and the conclusions of modern point-to-point communication theory. Zatocoding is defined--the system which superimposes random subject code patterns on the tally--and it is found to obey each of the invariant principles of coding efficiency while still allowing the simplest possible selector machine structure.

I Introduction
II Information retrieval by machine
III Retrieval systems and communication theory
IV Principles and practice of Zatocoding

I - INTRODUCTION

Zatocoding is a system of coding for selecting recorded intelligence by means of a machine. It is of special importance to the documentalist because it gives him powerful new tools of idea specification for retrieving information from storage. Zatocoding has practical economic consequences because it can enormously enhance the information-handling capabilities of the most elaborate large-scale machines, and also because it can be practiced with extremely simple machines.

Machine handling of ideas is an old art in the field of point-to-point wire and radio communication. Yet, over many years, communication technology has had little influence upon the closely related problems of documentary information retrieval. Within the last decade there has been a burst of developments and fundamental advances in communications technology which can no longer be ignored if documentation techniques in information searching are not to be left far behind. Since the war, the author's research has been directed primarily toward this aspect of documentation.

The nature of the advance represented by Zatocoding can best be followed by drawing parallels to recent progress in communications and related technology. Typical advances are frequency-modulated radio-telephony, military radar, electronic digital computing machines, and pulse-code systems of speech modulation. These advances stem from two sources: improved components and improved systems. Typical new or improved components are high-powered magnetrons, cavity resonators, photoelectric tubes, and semi-conductor diodes. Equally important is the recent development of entirely new systems or schemes for using these improved components. The components needed to construct a frequency-modulated radiotelephony system were available in the earliest days of radio broadcasting. Yet frequency modulation as a possibility was actually discarded because at that time there was no realization of the tremendous advantages that could be secured through its use in a properly designed system.

Recently developed communications systems --including computing machines which have to talk to themselves and their operators--owe a great deal to the new discipline called "communication theory." Communication theory[1] deals with systems of transformation of information from one medium to another, as from voice to electric current; with modes of representation of information within circuits, storage elements, and transmission channels; with interaction between noise, operator and equipment imperfections and the modes of information representation; and with ways to overcome these imperfections. The relevance of communication theory to documentary problems is obvious when we realize that a document of paper, a voice, or a television picture can each carry intelli-

[1] C. E. Shannon, "A Mathematical Theory of Communication," Bell System Technical Journal, v. 27, pp. 379-423, 623-656 (July, October, 1948).

C. E. Shannon, "Communication in the Presence of Noise," Proc. I.R.E., v. 37, pp. 10-21 (January 1949).

William G. Tuller, "Theoretical Limitations on the Rate of Transmission of Information," Proc. I.R.E., v. 37, pp. 468-478 (May 1949).

Norbert Wiener, "Cybernetics," John Wiley & Sons, New York, 1948.

The bibliographies of the Shannon and Tuller papers are recommended.

gence, and that their transmission and utilization have many common problems.

Zatocoding[2] is a new system, related to these new communication theories. It is especially capable of handling the problems of documentary information retrieval by a machine. As measured by capability of dealing with ideas, and by economy in machine structure, Zatocoding is the most efficient coding system presently known. Its efficiency is related to the several close analogies that exist between it and certain of the system requirements for high-efficiency point-to-point signalling. In particular, I wish to mention that Zatocoding has parallels to the probability and choice concept in coding of messages,[3] and to the use of random codes[4] for transferring message ideas in communication theories.

Zatocoding is a considerable departure from the earlier methods of coding used for information retrieval, and therefore it has some unusual characteristics. Referring to Fig. 1, code patterns, representing the descriptor ideas, are superimposed in the single coding field, and thus they intermingle and mix. The coding patterns are initially generated by a mechanical random process, and they are then assigned to the subject ideas in any order. For selection, the marks and spaces of the codes are not matched, but instead pattern inclusion of the marks only is employed. Rejection of unwanted material is governed by statistical rules, allowing a slight percentage of "extra" unwanted material to come out with the desired material. In most cases, the coding and idea-manipulating capabilities of a given set of equipment can be enormously enhanced by Zatocoding, as a few examples will show.

The Rapid Selector[5] operates on a film memory in which the coding field for each frame has some 216 positions that can be marked by opacities. Assuming a collection of 5,000,000 documents--comparable to the Library of Congress--a Zatocoding pattern of 8 marks per subject idea can be used. The coding field can hold up to 18 such subject patterns, and selections can be made upon any combination of these patterns. The size of the Zatocoding descriptive vocabulary is unlimited. In comparison, the present Rapid Selector coding system (and associated optical-electronic adjuncts) records only 6 subjects in the field, it searches and selects upon only one subject, and combinations of subjects cannot be used to specify selection.

A Hollerith card has 960 positions that can be marked by punching out holes. If the collection of records is moderate-sized, that is, of less than 10,000 pieces, Zatocoding patterns of only four punches can be used. In this case, 165 different subject ideas or record statements can be put into the card by Zatocoding for use in selection and correlation. An entire medical or sociological case history might be recorded on only one card. Again, the size of the descriptive vocabulary is unrestricted. With conventional procedures, it is impossible to record 165 statements on a Hollerith card because there are only 80 columns which are available for numerical codes.

The very high efficiency of Zatocoding makes it the technique of choice for coding the simpler edge-notched cards. The typical Zatocard shown in Fig. 1 has only 40 positions that can be notched. Yet this card is suitable for a collection of 10,000 items, or larger, and it can be sorted with Zator equipment shown in Fig. 2 at speeds of around 800 cards per minute. (The electronic Rapid Selector is only 12 times faster.) Seven independent, cross-referencing subjects can be notched by Zatocoding into the 40-position Zatocard.

Another version of the Zatocard has 72 field positions and it can hold 13 subject ideas.

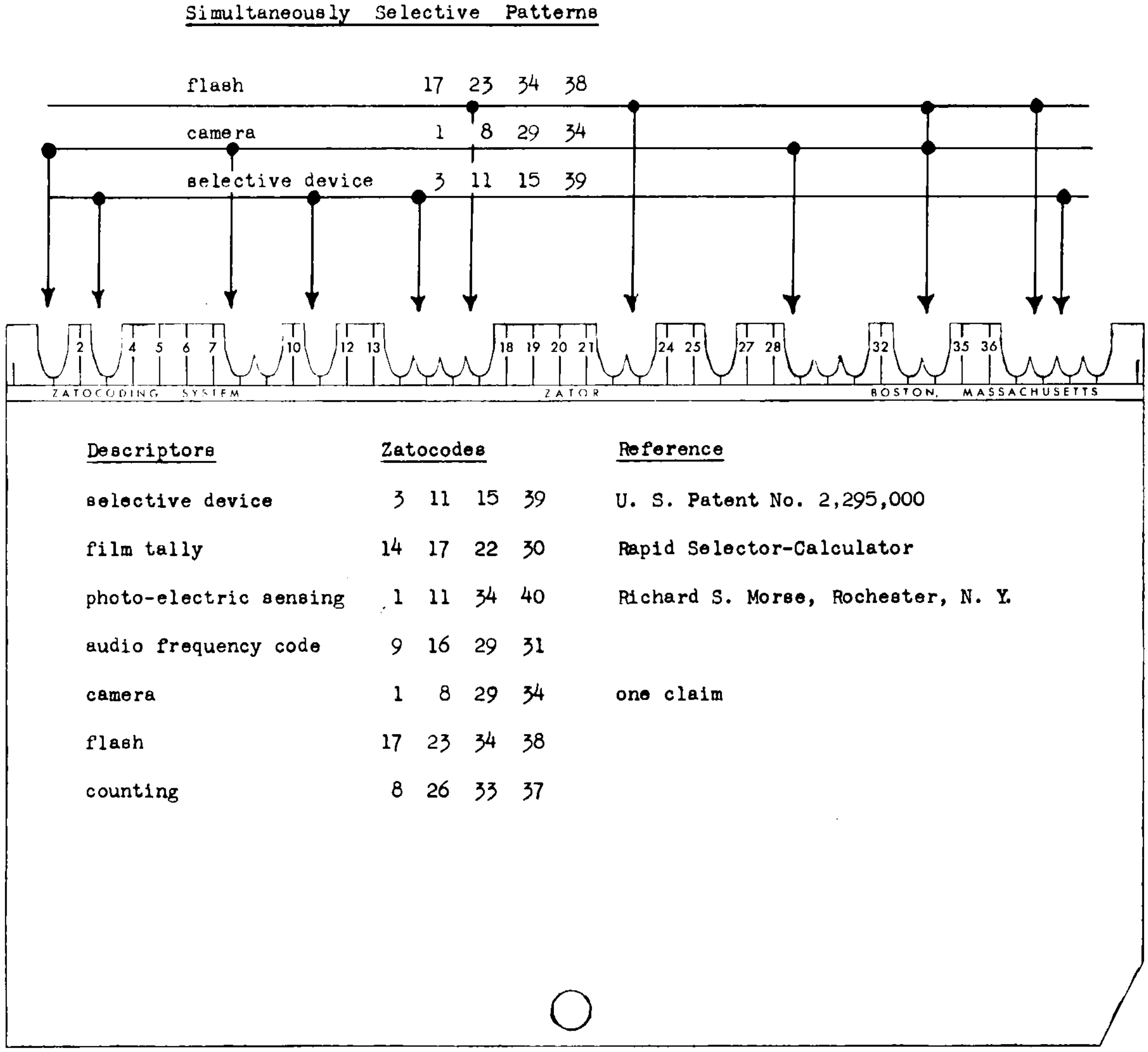
[2] C. N. Mooers, "Putting Probability to Work in Coding Punched Cards," paper, 112th Meeting American Chemical Society, New York City, September 1947. U.S. and foreign patents pending on Zatocoding.

C. N. Mooers, "Information Retrieval Viewed as Temporal Signalling," Proceedings of the International Congress of Mathematicians, Cambridge, Massachusetts, September 1950.

[3] Shannon (1948) ibid. p. 379.

[4] Shannon (1949) ibid. p. 17.

[5] Anon., "Report for the Microfilm Rapid Selector," Office of Technical Services, U.S. Dept. of Commerce, Washington, D.C. (1949).



Descriptors	Zatocodes	Reference
selective device	3 11 15 39	U. S. Patent No. 2,295,000
film tally	14 17 22 30	Rapid Selector-Calculator
photo-electric sensing	1 11 34 40	Richard S. Morse, Rochester, N. Y.
audio frequency code	9 16 29 31	
camera	1 8 29 34	one claim
flash	17 23 34 38	
counting	8 26 33 37	

Fig. 1. ZATOCODING, illustrated with a 5 by 8 inch edge notched Zatocard for the tally. Note that the random Zatocode patterns in the field overlap and intermingle. Selection on the combination of three descriptors, "flash," "camera," and "selective device," is according to the inclusion of the pattern of arrows into the pattern of notches in the coding field. Zatocards are sorted by the selector shown in Figure 2.

It is suitabie for more complicated indexing problems or larger collections.

To explain why Zatocoding is able to give such large advantages over older methods, it is necessary to examine concepts inherent in information retrieval by a machine and to compare these concepts with the more familiar methods of classification, indexing, and card-cataloging. Following this discussion, the section on retrieval systems and communication theory states some new principles of information retrieval systems and shows how they apply to machine operation irrespective of the intellectual schemes. The final section, on principles and practice of Zatocoding, outlines the characteristic features of Zatocoding and gives the step-by-step procedure for the design of a Zatocoding system.

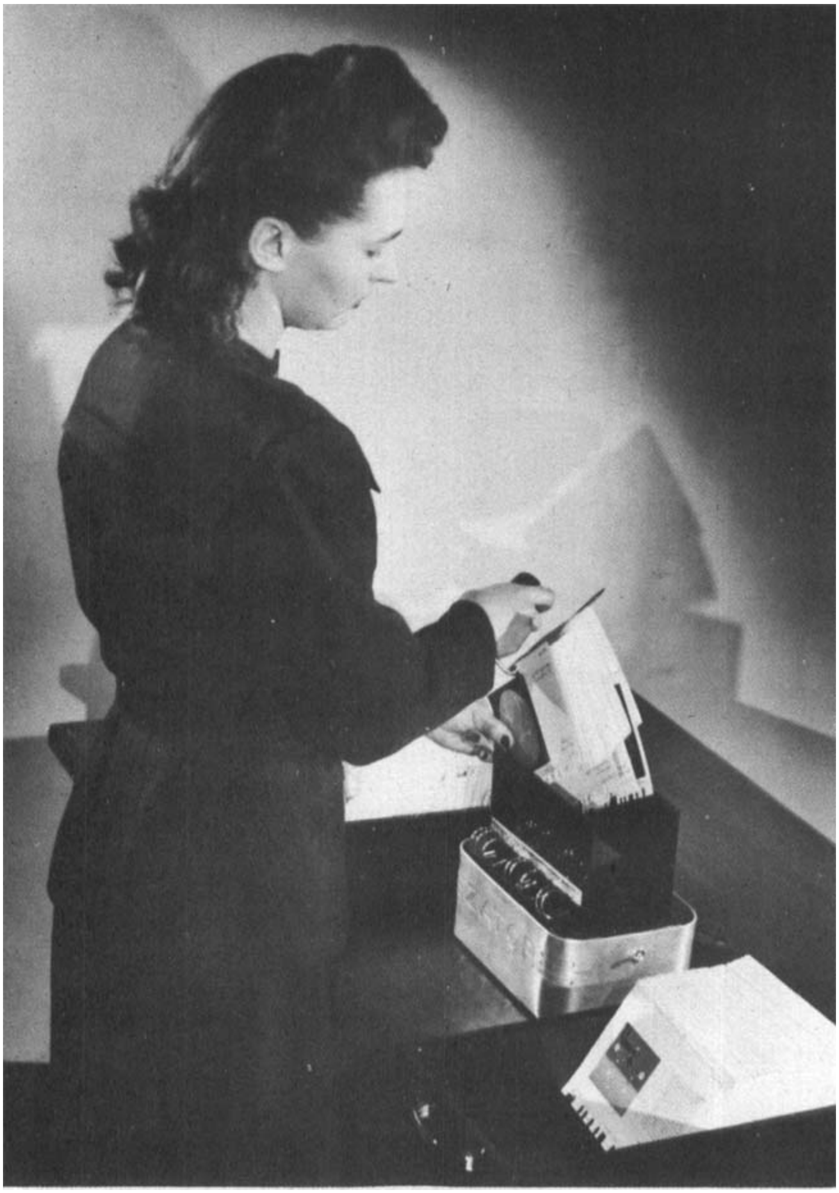


Fig. 2. THE ZATOR SELECTOR, capable of sorting edge notched Zatocards at the rate of 800 cards per minute. The black box-like portion of the Zator Selector is vibrated by a motor in the base. At the bottom of the box, a row of holes allows pins to be inserted to form the selective grid pattern. For selection, a pack of Zatocards is placed notched edge down on the vibrating grid. The rejected cards are those whose edge notches do not fit the pattern of the grid; hence they are supported on top of the grid. The edge notches of the accepted cards fit the grid and the cards drop down through the grid to a distance equal to the depth of the notches. In this way, the accepted cards are offset slightly below the rejected cards. To separate the rejected cards from the accepted cards the operator spears the pack of rejected cards through the small hole on the edge opposite the notches and lifts the pack out. The accepted cards, being offset below the rejected cards, are not engaged by the spearing tool; they drop out from the pack onto the table.

II - INFORMATION RETRIEVAL BY MACHINE[6]

Information retrieval is the name for the process or method whereby a prospective user of information is able to convert his need for information into an actual list of citations to documents in storage containing information useful to him. It is the finding or discovery process with respect to stored information. It is another, more general, name for the production of a demand bibliography. Information retrieval embraces the intellectual aspects of the description of information and its specification for search, and also whatever systems, techniques, or machines that are employed to carry out the operation. Information retrieval is crucial to documentation and organization of knowledge.

We will be concerned here with information recorded and stored in a collection. The unit of information is a single document, paper, or report. Storage facilities and reproduction processes are not considered.

The subject matter of each document or other unit of information is characterized or described by means of a set of "descriptors" taken from a formal vocabulary of such terms. A "subject heading list" will call to mind a rough approximation of what is meant here.

If we allow the word "machine" to extend to any physical or material structure that is useful for information retrieval, we find that classified documents, index lists, and card catalogues are all machines for the organization of knowledge. This being so, it will be useful to examine them here to give perspective to our discussion.

Classification is a system of mechanical organization of knowledge in which usually the documents themselves are placed in a physical arrangement defined by a "classification schedule." The classification schedule is a predetermined listing of combinations and patterns of descriptors, with the order of listing of the combinations and patterns usually supported by a set of logical, or arbitrary, arguments and conventions. A classification of documents is thus a passive machine. Classification systems are usually held to be systems for information retrieval, and not merely elaborate forms of storage. If so, then it should be possible for a prospective user of information to find his information by first consulting the classification schedule, and then by going directly to the desired documents. Moreover, he should have a fairly high expectation of success in efficiently finding what he wants to by this operation. As we all know, classification systems do not in fact meet this ideal. I believe there are some fundamental reasons why they can never do so.

An index is (usually) an alphabetical listing of single descriptors and their simpler combinations. The index format may be either a list on a sheet of paper or entries on a set of cards. To each such entry there is a page number or other citation to the location of the original document or unit of information. The list of index entries is the tangible machine of information retrieval. The documents themselves may be stored in any order, which is a great advantage. Information retrieval is accomplished by naming one or two or three descriptors, finding an entry with the desired combination in the list, and then following the citations to the original document. Indexes are powerful tools of information retrieval, though they run into difficulty when handling a large number of interacting descriptors. They must be subdivided finely enough so that a human searcher will not tire before finding his entry--a serious weakness of a practical index. We note, however, that an index has greater flexibility than a classification system in handling useful combinations and variations of descriptor patterns.

Human frailties can be avoided by turning the search job over to an active non-human machine. Each unit of information in the collection is then given a tally which can be manipulated by the machine. The tally may be a punch card, a frame of photographic film, or a section of a magnetic recording tape. The tally bears two kinds of information: a) a citation to the storage location of the unit of information, and b) the set of all descriptors applicable to that unit of information. These

[6] Compare: C. N. Mooers, "The Theory of Digital Handling of Non-Numerical Information and its Implications to Machine Economics," Zator Technical Bulletin No. 48, Zator Company, Boston, 1950.

descriptors are written on the tally in a fashion that can be "read" by the machine, such as by punches, dots, or magnetic spots.

When the tallies are cards and a human being instead of a machine reads the descriptors recorded on them, the system reduces to the ordinary card catalogue. However, humans are expensive to use--especially compared to the simplest machines, such as the Zator machine--and they have many frailties. They are slow, inaccurate, inattentive, and quite definitely limited in their ability to search through multitudes of cards for descriptor combinations of high multiplicity.

To return to the non-human machine: The descriptors used on the tallies must be broad for greatest utility. "Aircraft" and "helium" are excellent descriptors, but the typical subject heading "aircraft, lighter than air, helium filled" is too narrow and specialized. Such a narrower meaning can be generated with broad descriptors, for finding purposes, by specifying both "aircraft" <u>and</u> "helium" among the descriptors of the information.

Each document or unit of information is characterized by a set of descriptors taken from the vocabulary of descriptors. Each descriptor of the set applies to, or is true in some way of, the information content of the unit of information. The descriptors operate independently in this type of characterization. The fact that there are several descriptors in the set may mean that they formed some interacting combination in the original document, or it could just as well mean that they relate to independent ideas scattered through the document. Using descriptors in this fashion drops almost all relationships between the ideas represented by the descriptors.

Is this an undesirable degeneration of the information? Is it therefore abhorrent? Evidence shows that it is not, for several reasons. In the first place, it actually works very well in actual information retrieval systems. Also, relationships are subtle things, depending upon the point of view in most information situations. A person looking for information cannot be expected to have a cut-and-dried point of view, nor can he be sure of the relationship. All he is sure of is a very few of the objective or concrete facts of the situation, that is, the descriptors of the kind mentioned.

Here a distinction is very important. The distinction is between communication of information by language (which we don't need) and the appropriate language for finding information (the kind we must use in information retrieval). The point is worth a longer discussion than can possibly be given here.

The relationship between the machine language of punches, dots, or spots on the tallies and the intellectual content of the original documents given by descriptors is called the coding. The very nature and design of the machine used for information retrieval is determined primarily by the coding system adopted. The actual competence of the system to retrieve information also depends almost entirely upon the coding adopted.

Serious mistakes in coding schemes have been made by adopting or designing a mechanism first, and then using Procrustean techniques to force the coding scheme to fit the limitations of the machine. Certain experiments with tabulating machines provide examples of this approach. Equally serious mistakes can arise from taking the methodology of other systems--e.g., the decimal symbolism developed for classification by physical array--and forcing such a methodology upon machines which could otherwise be made highly capable and versatile.

Efficient coding systems for the tally method do exist. Therefore, it is pertinent for us to review the reasons why the machine-sorted tally method deserves special consideration as a technique of information retrieval:

1. The load of tally manipulation, descriptor sensing, and the problem of choice or selection is all turned over to a machine.

2. High-speed tally-searching machines already exist and speedier machines have been proposed. Such a machine can very rapidly scan all the tallies in a collection and compile an exhaustive demand bibliography to the specifications of any prospective user of information.

3. Pre-determined schedules of descriptor combinations are not necessary nor are they employed with tallies. The descriptors operate independently. Thus retrospective searches can be made for interrelations or correlations between descriptor ideas that were not foreseen when the descriptors were separately recorded on the tallies. In contrast, according to conventional library methods, the searcher is limited to interre-

lations actually foreseen and recorded by the cataloguer. Separate descriptor ideas are very objective, while the interrelations (dear to the cataloguer) depend upon the point of view. Unfortunately the points of view of the engineer and the cataloguer are often different.

4. The combination of many descriptors is used to specify a search. In coding the structure of an organic chemical compound, 15 or 20 descriptors might be used on one tally. To search for this compound and its relatives, as many as 6 or 8 descriptors might be used to specify its class. With a capable code, a machine can handle this easily. Human methods would be most inefficient and liable to error.

5. The documents themselves are stored in any convenient order or fashion. Storage is uncomplicated by classification.

6. Changes, or the inclusion of new knowledge or descriptors, do not upset the system or the storage arrangement. Because the descriptors are used independently, new ones are simply added to the vocabulary along with the rest. Unlike a classification schedule, the vocabulary is a mere listing of descriptors and there is no attempt at logical structure.

7. Finally, an important conceptual point: By the tally method the knowledge contained in the collection of documents is given no actual organization until a user requests information. Only at that time is "organization" brought into being. Then the request is framed by a set of descriptors, the machine scans the tallies, and prepares a demand bibliography. In this sense, every time a request is met, the collection is "organized" from a new point of view.

III - RETRIEVAL SYSTEMS AND COMMUNICATION THEORY

Tally methods hold outstanding possibilities for mechanical organization of knowledge--provided that a competent coding system is adopted. In this section we will consider those general principles of coding which hold irrespective of machine details, tally details, or philosophy of coding. These are invariant principles of coding systems. When these principles are obeyed, the coding system is efficient--and conversely. These coding principles have some remarkable parallels to principles otherwise developed in communication theory. Certain of these parallels will be indicated through references.

The speed of scanning the tallies is roughly proportional to the complexity of the searching machine, and is inversely proportional to the size of the tally field. Certainly, it will take longer, or require a better machine, to search a tally having a field with 1,000 positions than to search one with 100 positions. It is highly desirable to scan all the tallies.

> Principle 1: SMALL FIELD. The size of the tally field should be made as small as is compatible with the other requirements of the problem.

With a small tally field, made desirable by selector simplicity and overall speed, it is necessary to plan for maximum utilization of the field. To give the tallies the greatest possible powers of separation or selection, the patterns of marks and spaces on the tally fields should have the greatest variety possible. No blank areas should recur on tallies because of information not recorded. Neither should similar patterns recur frequently on the same portions of the field.

With these assumptions it can be shown algebraically that the greatest number of different patterns in the tally fields--and thus the greatest variety--will occur when approximately one-half of the positions in the fields are marked, and when the patterns are as random as possible.[7]

> Principle 2: FIFTY PER CENT. Whatever the coding system, maximum utilization occurs when the density of marks in the field is in the neighborhood of 50 per cent.

> Principle 3: RANDOM PATTERNS. Maximum utilization requires random patterns in the tally fields.

[7] Completely random patterns correspond to "white noise" of signal theory. Compare Shannon (1949) ibid. p. 17. Randomness of this kind gives an ultimate coding of intelligence, and Zatocoding utilizes this property.

The coding system should be specifically designed only for finding, searching, discovering, or retrieving information. The coding should not be influenced in any way by features of arrangement or storage of documents. Neither should the coding be concerned with communicating the information in the document. The coding should talk about--not communicate--information. If a man wants information on the effect of temperature on his experiment, he should not be forced to anticipate that the critical temperature is 10 degrees before he can cause a machine to produce his information from the files. He wants to ask the retrieval system a question about temperature in his experiment, not tell it. If he knew the answer was 10 degrees, he wouldn't need to go to an information source. The descriptors used on the tallies should only outline the information of the document. They should specify only what kinds of information can be found there.

> Principle 4: RETRIEVAL LANGUAGE. The coding of the tallies should carry only "retrieval language." The ordinary "communicative language" that conveys actual information remains in the document itself, or if it is on the tally at all, it is confined to the written abstract and it is not marked in the coding field.

The distinction between communicative and retrieval language is stressed here because it has long been an ideal in library science to "pin-point" the documents by a classification symbolism. Pin-pointing in this fashion is tantamount to communication. Theoretical studies in our organization, and practical experience gained from setting up a wide variety of successfully operating information retrieval systems, point to the definite conclusion that communicative language or pin-pointing is incompatible with successful retrieval.

Another invariant principle is the concept of choice[8] in an information retrieval system. One does not pick information or documents from an infinite universe of documents. Selection is always referred to a finite--though possibly very numerous--collection in storage somewhere. Therefore, all that information retrieval can accomplish is a choice of one or more documents from such a finite collection.

How much choice is involved? Certainly it is much easier to choose one item from a collection of ten than it is to choose one item from a collection of ten thousand. How much easier? Also, if it is easier to make the first kind of choice, a simpler coding should suffice for the smaller collection. Fortunately, these matters have a very simple and satisfactory quantitative formulation.

The choice of an item from a collection of two involves only a single two-valued decision: choose the first, or choose the second. We can designate the two values by 0 and 1, and we can call these digits. It requires two such elementary decisions in series to choose one item from a collection of four. We first split the collection in half, and decide which half we want, then in that half we decide which of the two objects we will finally choose. Thus for a collection of four items, the choice which represents any one item can be specified by two digits in series, and in fact the four objects can be respectively represented by the symbols 00, 01, 10, and 11. It is noted that this is precisely the binary enumeration system[9] that is much used in the internal operation of contemporary electronic digital computers. To enumerate eight objects will require three digits. Note that $2^3 = 8$.

> Principle 5: CHOICE. If there is a collection of no more than 2^S objects, we can specify a unique choice of one of them by means of a symbol having no more than S digits, each representing an elementary two-valued decision.

Values of S for collections of several sizes are as follows:

Size of Collection	Number of Two-valued Decisions for Choice
10	4
1,000	10
100,000	17
1,000,000	20

[8] Compare: Shannon (1948) ibid. p. 379.

[9] Also called "dyadic numbers." Compare G. Birkhoff and S. McLane, "A Survey of Modern Algebra," pp. 33-34. Macmillan, New York, 1944.

A machine for the selection of tallies scans or reads the marks and spaces in the tally field. The machine actually compares the patterns found on the tally field with some pattern of marks or spaces held in the machine to define the selection. Each mark or space in the pattern of the machine represents one value of a possible two-valued decision, and if there are $\underline{S'}$ such marks, then the pattern has the capability of making a unique choice among as many as $\underline{2^{S'}}$ objects. If the size of the searching pattern $\underline{S'}$ is larger than it needs to be, as compared to the magnitudes of the choice among the actual number of items in the collection, the coding is inefficient.

Most machines and systems discussed in the literature are extremely inefficient. Therefore they are able to record fewer descriptors in their coding fields, and they require more apparatus for searching than might otherwise be the case. Samain[10] employs a selection pattern of 36 marks and spaces to select upon a single descriptor. Most of his examples of selection are by two or three descriptors, having combined patterns of 72 or 108 places respectively. He mentions collections of the order of one million pieces. Such collections would allow adequate choice by a selection pattern of only 20 places. If he always used two or more descriptors for selection, he could represent any descriptor by a pattern of only 10 places instead of his present 36.

These considerations give rise to the next principle.

> Principle 6: DESCRIPTOR LENGTH. The number of effective marks or spaces in the pattern representing a descriptor should be set by the requirements of choice among the actual collection, and not by the size of the vocabulary as is the usual practice now.

Let us consider a collection of 4,000 documents which has a notched card information retrieval system with 1,000 descriptors in the vocabulary list. It is known in advance that all selections from this collection will be stated with three or more descriptors acting together. How many marks or spaces are required for the descriptor patterns?

The conventional answer is that each descriptor must be given a pattern 10 places long, otherwise the descriptor patterns will repeat. This pattern length is ridiculous when analyzed according to the choice required by selection. Since three such patterns will always be used in selection, the total selective pattern would always be 30 places long.

A choice among 4,000 pieces really requires a selective pattern of only 12 places. The correct answer, therefore, is that the individual descriptor patterns need have only four marks or spaces.

A paradox now appears. The patterns of some descriptors may be duplicates of other descriptors, yet we can still define the choice of the tally we want. The paradox is resolved when we notice that in addition to the desired choice, there is a statistical possibility that some unwanted descriptor patterns on unwanted tallies will by chance duplicate the patterns being searched. When this happens, extra tally selections will occur, though their number will usually be inconsiderably small and their occurrence can be approximately predicted from the details of the coding. The phenomenon of such "extra tallies" is inextricably bound up with the adjustment of the length of descriptor patterns in this fashion to increase coding efficiency.

As a final consideration in this section, we note that the manner in which the descriptor code patterns are recorded in the tally field has an enormous influence on the complexity of the selector machine and upon the competence of the whole system for information retrieval. While these factors can be evaluated quantitatively for different manners of recording, all that will be done here is to describe the several simplest possibilities. There are two design choices to be made. First, the tally field can be subdivided into mutually exclusive subfields with each subfield taking only one descriptor pattern, or the

[10] Jacques Samain, "A New Apparatus for Classification and Selection of Documents and References by Perforated Cards," pp. 680-685, also pp. 158, 230, and 265 in "Reports and Papers Submitted, The Royal Society Scientific Information Conference, 21 June to 2 July 1948," The Royal Society, London, 1948. See also Samain, pp. 22-26, Rept. 17th Conf. Int. Fed. Docmn. (1947).

field can be left undivided and the descriptor patterns can be superimposed one on top of the other in the same field.

Design Choice 1: "Mutually exclusive subfields" vs. "superimposed codes."

Second, the system can be operated so that the selector looks for a given descriptor pattern in some invariant position in the field (e.g., the descriptor for "red" is always found in the third subfield), or the selector mechanism can be made more complicated so that it can search out descriptor patterns in many alternative (subfield) and positions in the field.

Design Choice 2: "Invariant position" vs. "alternative position" scanning.

For example, Samain employs mutually exclusive subfields with alternative position scanning. Alternative position scanning is a very powerful technique. However, it requires an elaborate and expensive sensing mechanism because of the need to search many alternative locations in the field.

Conventional, small-scale, notched card systems employ mutually exclusive subfields, but avoid the expense and complications in sensing mechanisms by using fixed locations for descriptor codes, "invariant position scanning." This compromise severely limits the usefulness of such information retrieval systems and makes them incapable of handling many problems.[11]

IV - PRINCIPLES AND PRACTICE OF ZATOCODING

Zatocoding can now be defined as a system for using machine-sorted tallies for information retrieval in which the coding system has superimposed codes and the selector employes invariant position scanning. The great advantage of this combination of design choices is that the powerful capabilities of alternative position scanning can effectively be attained while using a very simple invariant position scanning machine. Moreover, Zatocoding follows each of the coding principles set out in the preceding section. Therefore, the Zatocoding system allows simple machine structure, gives the maximum efficiency of codes, gives minimum size of coding field, and allows maximum scanning speed for a given situation.

By means of an example, let us follow through the steps in setting up a Zatocoding information retrieval system.[12] To make the example quite definite, at each step we will give the appropriate numerical quantities along with the basic design formula when possible.

Example: We wish to design a retrieval system for a collection of 4,000 documents. There are 1,000 descriptors in the vocabulary. We wish to use as a tally a simple notched-edge card having only 40 positions in its coding field (Small field, Principle 1). From an analysis of our problem, we anticipate that, in almost all cases in which we wish to make a search for information in this collection, we can define the selection by at least three descriptors acting in combination.

The design proceeds as follows: The collection of 4,000 pieces is just less than 2^{12}, so that the choice involved in selection will require a pattern of 12 operating positions (Choice, Principle 5). Almost always, three or more descriptors will be used in selection, so a single descriptor pattern need have only 12/3 □ 4 marks in its code pattern (Descriptor length, Principle 6).

Zatocoding uses no subfields. The whole field receives the codes, and so the marks of the individual code patterns are superimposed one after the other on the field. The spaces of one code pattern may be filled up by the marks of another code pattern. Therefore, only the marks in the pattern have an invariant meaning. A Zatocode is then really a pattern of marks only, not of marks and spaces. In our example each individual pattern must contain four marks. Since the field has 40 positions in it, the possible number of different patterns of four marks is given by the number of combinations of 40 things taken four at a

[11] C. N. Mooers, "Zatocoding for Punched Cards," Zator Technical Bulletin No. 30, pp. 6-9, Zator Company, Boston, 1950. This contains a complete discussion of the effects of the compromise.

[12] Ibid. pp. 9-19.

time, which is precisely 91,390. However, only 1,000 patterns out of this set are needed, and the question is how to choose them.

It is imperative, for the successful operation of Zatocoding, that these patterns (assigned to the descriptors) be chosen by some mechanically random process from the set of all patterns (Random patterns, Principle 3). The reason for the randomness is that the marks from the successive patterns overlap and intermingle to some extent, as they are put into the field, with the marks that are already there. This intermingling is at a minimum only when the patterns are random patterns. There is no possible systematic way of assigning patterns to descriptors which is better than by completely random assignment. In fact, several systematic assignments that have been tried gave impossibly bad performance.

In order to assign Zatocoding patterns to the descriptors, we proceed as follows. The patterns are first generated. Balls numbered from one to 40 are mixed and then drawn four at a time from a container, to be replaced for the next random draw. These four numbers are an appropriate random pattern. A list of such random patterns is made. We already have a vocabulary list of descriptors. It doesn't matter in what order they are. The random patterns are then assigned to descriptors by giving the first pattern to the first descriptor, the second to the second, and so on for all the descriptors. That is all, for code assignment.

I am often asked whether the patterns of marks and spaces cannot be assigned according to some analogy to the successive digits of a decimal classification symbolism. Such a method seems attractive only until it is analyzed with respect to its consequences with superimposed codes. Because similar marks in the patterns would overlap with undue frequency, the method is worthless. Random patterns must be used with Zatocoding.

Since the Zatocoding patterns are placed in the same coordinate position in the single coding field of a tally, how do we know when the coding field is full and can hold no more patterns? With mutually exclusive subfields, the situation is very simple. Only as many codes can be recorded as there are subfields. A mathematical analysis of the Zatocoding situation shows that the optimum amount of information is carried in the tally field when around 50 per cent of the field positions are marked by descriptor codes. This result confirms, by another method, our conclusion of the previous section (Fifty per cent, Principle 2) concerning maximum utilization.

To return to our example, the codes have four marks apiece, and are recorded or notched into a field of 40 positions. Therefore, it would seem that we could record $\frac{(\frac{1}{2})(40)}{4} = 5$ different codes. Actually because the codes overlap, we can place more codes than this in the field before the density of marks reaches 50 per cent. The rule is that with random codes the sum of the marks of all the codes is equal to 69 per cent of the number of field positions when there is an average density of marks of 50 per cent. In our example, the number of Zatocode patterns that can be notched into the card is thus $\frac{(0.69)(40)}{4} = 6.9$ or effectively 7 patterns.

Machine Zatocoding selection of the patterns in the tally fields occurs when each mark in the selection-defining pattern is matched by a corresponding mark at the same position in the field of the selected tally. This is illustrated in Fig. 1. The tally field may have more marks than the selective pattern, but it cannot have fewer and be selected. This type of selection is called "selection by pattern inclusion," and the selective pattern is "included" in the field pattern. Zatocoding selections are made usually according to several descriptor patterns operating simultaneously. For instance, the selection might be according to "flash," "camera," and "selective device." Only those tallies which bear the marks of all three of these descriptors are specified for selection. In the terminology of symbolic logic, this selection is according to the "logical product" of the three descriptors.

It is expected that there may be other unwanted tallies which bear none of the desired descriptors, but whose patterns by chance combine in such a way as to imitate the selective pattern. These produce "extra" selected tallies. By proper design of the Zatocoding system according to the rules above, these extra tallies are made inconsequential. Desired tallies are never excluded by such chance operations; instead, a few other tallies may be included in the selection. This is all right. By application of probability theory, the average number of extra tallies to be expected can be predicted. If the design rules given here are exactly followed, only one extra tally per selection will occur on the average whatever the size of the collection.

Control of the number of extra tallies is accomplished through the length of the pattern which defines the selection of the wanted and the rejection of the unwanted tallies. The logic is as follows: The selective pattern always matches the field pattern of the desired tallies, and they are always selected. We can eliminate them from our consideration of how the extras are rejected. All the tally fields are coded with random codes having an average density of marks of 50 per cent. A single mark in the selective pattern will reject by chance only half of the unwanted tallies in the collection. It will accept the others. The second mark in the selective pattern will in turn reject half the unwanted tallies that were accepted by the first mark. This cuts the number of unwanted tallies to one quarter, and so on. Each successive mark added to the selective pattern improves the accuracy of the rejection by a factor of one half. In the example, the selective pattern of 12 marks can be expected to exclude all of the unwanted tallies in the collection of 4,000 except one, on the average. We note also that, if we were forced to make a search defined by only two, instead of three, descriptors, the selective pattern of only 8 marks would allow possibly 16 extra tallies to appear. These can easily be recognized among the desired tallies, and can be discarded.

The same design principles apply to any other size of collection, design of tally, or machine. A collection of a hundred million (10^8) items would take a Zatocoding pattern of 27 marks for a descriptor to give almost perfect performance (with an average of one extra tally) on a single-descriptor selection. A Hollerith card could take 25 such descriptors. Selection could be made on any combination of descriptors, or on single descriptors, taken from a vocabulary of unlimited size. By comparison, the more usual methods of coding, with numerical codes and alternative subfield position scanning, results in a very complex machine and a vocabulary limited to 1,000 descriptors at most.

The extreme simplicity of Zatocoding selection, employing pattern inclusion selection and invariant position scanning, reflects in a corresponding simplicity and economy of the selective machine. At ordinary speeds of tally scanning (in the order of 1000 per minute) a simple mechanical grid, sensing notches in the edge of cards in a pack, is completely adequate. By going to more elaborate technologies, of the kind now used for television broadcasting from movie films, more than 1,000 tallies can be sensed per second. By several further modifications in the manner of use of these electron-optical scanning devices, selection speeds in the order of 1,000,000 tallies per second are apparently possible. In another paper[13] the author has described such a super-speed machine, calling it a DOKEN or "documentary engine," and has discussed the associated organs appropriate to such scanning speeds.

In conclusion, where the simplest and highest speed selector is desired for sorting upon the logical product of descriptors, Zatocoding is preferred because it realizes the maximum efficiency in coding. It achieves this efficiency by matching the intellectual choice represented by the selective descriptors to the choice represented by the number of items in the collection. The innovations of Zatocoding are the use of superimposed random codes, and a statistical prediction and control of the associated phenomena of extra selected tallies.

[13] C. N. Mooers, "Making Information Retrieval Pay," paper, 118th Meeting American Chemical Society, Chicago, September 1950. Published as Zator Technical Bulletin No. 55, Zator Company, Boston, 1951.