INFORMATION THEORY AND OTHER QUANTITATIVE FACTORS IN CODE DESIGN FOR DOCUMENT CARD SYSTEMS*

By EUGENE GARFIELD, Director
Institute for Scientific Information, 1122 Spring Garden Street, Philadelphia 23, Pa.

In the past ten years, the field of information retrieval has witnessed the development of many new systems, devices, and theories. In particular, two opposing "schools" of thought on card indexing systems have developed. One claims that the term card (unit term) or "collating" system is the most desirable. The other advocates the document card (unit record) or "scanning" system. Dr. Whaley has noted many of the advantages and disadvantages of collating and scanning systems, and I am glad to adopt his terminology and agree with most of his comments. For the record, however, I wish to remind the proponents of term card systems that theirs was no new finding. Costello says Batten anticipated Taube by 15 years. Batten was anticipated by at least another 35 years.

One term card system began at the turn of the century at Johns Hopkins Hospital. Subsequently, it went through all the evolutionary stages which clearly demonstrate the inherent similarities between term card and document card systems. This does not mean that the rediscovery of the term card system was an insignificant development. After all, many useful ideas and inventions are rediscovered and we are grateful for these discoveries. However, when appropriate, our precursors ought to be given credit. Even the ten column posting card was anticipated by Paul Otlet, founder of the modern documentation movement. Indeed, long ago, the term card system was used in several medical institutions, including Johns Hopkins Hospital and the Mayo Clinic.

Texts on medical records management demonstrate such systems. These consist of one 3 x 5 card for each disease (term). Each card then lists the case history document numbers for all patients so diagnosed. Ultimately, the number of case history numbers grew larger and the time required to make any correlations between two diagnostic term cards increased to ridiculous, exponential proportions. Somewhere along the line it was decided that the document card system should be employed. At Johns Hopkins and Mayo, Hollerith cards were in use as early as the 1920's. The School of Public Health at Johns Hopkins was one of the earliest users of punched-card machines. Their equipment is still of early vintage. At Johns Hopkins, even the IBM card finally became a problem as the volume of patients grew into the hundreds of thousands. The "vicious circle" was continued when it was decided to use duplicate sets of cards -- i.e., rotated files, not unlike the system used at the Chemical-Biological Coordination Center (CBCC) several years ago. Finally, this semi-collating, semi-scanning system was abandoned because of the high cost of storing millions of cards. The entire file was tabulated on printed sheets and the punched-cards thrown out. This printed index arrangement is very similar to the original term card arrangement. However, in a separate section, the equivalent of the document card is also printed. Thus, one is able to do a search by both methods. Depending upon the individual search either one or both may be used. Pre-coordinations were made where appropriate before printing the index.

The Mayo Clinic long ago attacked the space problem in another fashion. The storage density of the IBM card was increased by a system of binary coding. These IBM methods, I believe, are still used there. The binary coding utilizes all of the 4024 combinations possible in a 12 position punched-card column. It is understandable that a group of statisticians would discover this method. After all, statisticians work with probability data constantly. However, it is interesting that many people, including the statisticians, have been clever in finding ways of increasing the number of codes that can be crammed on a card (Wise, Mooers, et al.). However, the problem of how many times each was used was not considered as important.

This aspect first troubled me while working with the IBM 101 at the Welch Medical Library Indexing Project. Some readers may recall the experimental 101 system we demonstrated in 1953 using five digit decimal codes, randomly strung along the first sixty columns of an IBM card. For each subject heading or descriptor there was one five digit decimal number. Each card contained 12 such numbers. The details are described in the final report of the project. To use the same code length for all descriptors regardless of their frequency was rather inefficient in terms of space utilization, input time and searching cost. Obviously, others have arrived at similar conclusions because their coding systems intuitively employ a statistical approach. It is surprising, however, how many extant systems still do not make provisions for "normal distribution." A good example is the CBCC system, and the same is true of Uniterm, Zatocoding and others. To reiterate: they all use the same amount of coding space for each descriptor, regardless of its frequency of use.

Working with the CBCC system, and utilizing Heumann's statistical data on about 25,000 chemical compounds coded with this system, it was possible to design a code which reduced

*Presented at the American Documentation Institute Annual Meeting, October 22, 1959, Lehigh University, Bethlehem, Pa.

significantly card space and the time and cost of searching. For the moment it is sufficient to state briefly that the statistical information available on the CBCC file was used to construct a normal distribution curve giving the frequency of use of each alpha-numerical code. One then arbitrarily breaks into the frequency curves in various sections to determine the space allocations for the descriptors. If a descriptor, such as benzene, occurs in half the chemicals and the code for uranium occurs rarely, why devote the same amount of space to both. Obviously, as Wiswesser, Steidle and many others have found, it is quite sufficient to assign permanent card locations to frequently occurring codes. On the other hand, descriptors which occur infrequently can be assigned some coding configuration which requires, relatively, a great deal of card space. This will be of little consequence since it will crop up so rarely. These "rare" birds are treated as a class and codes are used that permit many combinations in a larger space. The Mayo system is one example; another is the Zator system, as applied by Schultz. Indeed, one of the primary shortcomings of Mooer's Zator system is the indiscriminate, *i.e.*, random assignment of an equal number of code symbols regardless of actual occurrence in the file. This results in excess noise, *i.e.*, false drops. Incidentally, I wish to point out that I am well aware of Mooer's early attempt in *American Documentation* to set Wise straight on the folly of a superimposed coding scheme for the now defunct Rapid Selector. However, to use probability theory is one thing — to use information theory is something else. We all readily can visualize methods of utilizing card space that will *grossly* take advantage of the facts revealed by a statistical analysis of the use made of a particular descriptor dictionary or subject heading list. The theoretician, however, wants precise quantitative criteria for allocating code space to individual descriptors or groups of descriptors. Here is where Information Theory comes to the rescue. The design of the most efficient coding system does not depend upon the meaning of terms. The terms, by themselves, have no informational value. Rather, it is the frequency of use of a particular descriptor which determines its informational content. One can only measure the amount of information in the word benzene when transmitting it in English text. As a code or term in a document collection dictionary, the word has no value. It is only significant in so far as it occurs with a particular frequency. If half of the chemicals coded contain benzene then the knowledge that a particular chemical contains benzene reduces the remaining choices to one half.

Having cleared the cobwebs on what the real "coding" problem is in documentation systems it is then relatively simple to apply Shannon's basic formula for measuring informational content. I might mention that it is difficult, at first, to think of the card searching problem as a transmission problem. However, if you think in terms of magnetic tape systems (Univac) or paper tape systems such as the Western Reserve Scanner, it is easier to see an analogy between "transmission" and searching.

The information content of a document file is neither the number of descriptors used, nor the number of documents which the various combinations of descriptors constitute. The information content of a document collection is a function of the *probabilities* of the descriptors in the dictionary. *H*, the familiar thermodynamic entropy function, and Shannon's measure of information, is equal to the sum of the individual probabilities multiplied by the logarithm of the individual probabilities, *i.e.*, $H = -(P_1 \log P_1 + P_2 \log P_2 + . . . + P_n \log P_n)$.

From this we are able to draw many interesting conclusions. For example, a document collection of 1,000 documents may contain no more information than a document collection of one million documents. This fact accounts for the intuitive decision of the Patent Office to use a "composited" card, which in certain cases is quite justifiable. It also can be shown that the informational equality in two such files can be changed readily if the depth of indexing is altered. Indeed, if the informational content remains constant during such a growth one must either conclude that unnecessary cards remain in the file, new sub-dividing terms are required, or noise is present during a search. This situation is illustrated perfectly by our experience in coding steroid chemicals using the Patent Office code. In many instances a dozen different steroids were coded exactly alike. If the code dictionary is not changed, it is properly concluded that it is more economical to "composite" the 12 cards into one. However, one could increase the specifity of the coding. From the point of view of the Patent Office, with emphasis on the generic approach, the former conclusion, compositing, may appear simplest. From the point of view of the research chemist the latter approach, more specificity in coding, is more desirable. Taube's paper at the ICSI Conference implies that a term card system for the same steroid file could be used as readily as the Patent Office document card system. This has a theoretical validity in view of the fact that in *both* systems no attention whatsoever is devoted to the frequency of occurrence of the various codes. (The Patent Office uses one punched hole position for each descriptor and the Uniterm system uses a 4 digit document number for each descriptor.) Indeed, from a tabulation of the coding done by the Patent Office of over 2500 U. S. patents, involving about 35,000 codes, it is no coincidence to find that seven descriptors account for over 9,200 codes, 16 additional account for another 9,100, the next 52 another 9,400 and all the remaining

359 descriptors 6,800. Deciding the relative merits of working with a term card involving 1,500 document numbers (the highest frequency code) or the time to run 2,500 cards through a machine with a speed varying (according to price) from 500 to 2,000 cards per minute is meaningless. This becomes particularly ludicrous if one then considers the time required to find those chemicals containing both a 3-Hydroxy Steroid code and a 17-Hydroxy steroid which occurs with almost equal frequency (1,200 occurrences). Instead of matching numbers on Uniterm cards by eye, one can speed this up by "collating" on an IBM machine at speeds comparable to the sorting operation. Using a Ramac system or a high speed computer this can be speeded further. The point is that each system, according to the circumstances, has advantages and for this reason, in certain cases, I have used a combination of both — even going so far as to maintain two independent systems. This is commonly done, but not admitted, in many installations.

Returning to the discussion of the now measurable quantity H of an information file, to explain how this measure of information is determined and used, I must resort to basic Information theory. For that I have paraphrased Shannon's own words, to which I refer those who are not yet familiar with Information Theory.

Information theory is concerned with the discovery of mathematical laws governing systems designed to communicate or manipulate information. It sets up quantitiative measures of information and the capacity to transmit, store and process information. Information is interpreted to include the messages occurring in standard communication media, computers, and even the nerve networks of animals. The signals or messages need not be meaningful in any ordinary sense. Information Theory is quite different from classical communication engineering theory, which deals with the devices used -- not with that which is communicated.

I submit that most of the polemics concerning devices, *i.e.*, term card *vs.* document card systems have kept us in the dark ages of conventional engineering theory. Relatively speaking, we have paid little attention to the nature of the information itself. This led to the failure to design really efficient searching devices; anyone who rents an IBM machine knows this. The measure of information, H, is important because it determines the saving in transmission time that is possible, by proper encoding, due to the statistics of the message source. Consider a model language in which there are only four letters — A, B, C, and D. These letters have the probabilities 1/2, 1/4, 1/8 and 1/8. In a long text, A will occur 1/2 the time, B one quarter, and C and D each 1/8. Suppose this language is to be encoded into binary digits, 0 or 1 as in a pulse system with two types of pulse. The most direct code is: A equal 00, B equal 01, C equal 10, and D equal 11. This code requires 2 binary digits per letter. However, a better code can be constructed, with A equal 0, B equal 10, C equal 110 and D equal 111. The number of binary digits used in this code is smaller *on the average*. It will equal 1/2 (1) + 1/4 (2) + 1/8 (3) + 1/8 (3) = 1 3/4, where the first term derives from letter A, second B, *etc.* This is just the value of H found if the probability functions are calculated.

The result verified for this special case holds generally — if the information rate of the message is H bits per letter, it is possible to encode it into binary digits using, on the average, only H binary digits per letter of text. There is no method of encoding which uses less than this amount if the original message is to be recovered without noise. An average of 1 1/4 bits is possible if the message is allowed to be noisy, *i.e.*, not a completely faithful rendition of the original message.

Before we can consider how information is to be measured it is necessary to clarify the precise meaning of "Information" to the communication engineer. In general, messages to be transmitted have "meaning," but have no bearing on the problem of transmitting the information. It is as difficult to transmit nonsense words or syllables as meaningful text (more so in fact). The significant point is that one particular message is chosen from a set of possible messages. What must be transmitted is a specification of the particular message chosen by the information source. The original message can be reconstructed at the receiving point only if such an unambiguous specification is transmitted. Thus "information" is associated with the notion of a choice of a set of possibilities. Furthermore, these choices occur with certain probabilities; some messages are more frequent than others.

The simplest type of choice is from two possibilities, each with probability 1/2, as when a coin is tossed. It is convenient, but not necessary, to use as the basic unit the binary digit or bit. If there are N possibilities, all equally likely, the amount of information is given by $\log_2 N$. If the probabilities are not equal, the formula is more complicated. When the choices have probabilities $P_1, P_2, \ldots, P_n$, the amount of information H is given by the equation above. An information source produces a message which consists not of a single choice but of a sequence of choices, for example, the letters of a printed text or the elementary words or sounds of speech. In these cases, by an application of a generalized formula for H, the rate of production of information can be calculated. This "information" rate for English text is roughly one bit per letter, when statistical structure out to sentence length is considered (see *Bell System Tech. J.*, October 1949) or ("Encyclopedia Britannica" article on Information Theory).

The problem of applying information theory to documentation, I believe, is to be solved in properly defining the information source, which is the totality of descriptors assigned in any file. The next problem is defining the language units, *i.e.*, the descriptors and/or their components. A classification number, *e.g.*, has built into it much more information than a Uniterm. Each facet of the class number must be taken into consideration when measuring the information content of a classification system. It is then necessary to determine the probabilities of the units involved.

I will further hazard the statement that in the design of a document card of the IBM type *the most efficient space utilization will be obtained when the informational content of all card fields approach equality*. For example, in the case of the steroid file mentioned above, a card of four basic fields could be designed in which about 25% of the information was contained in each. The first "field" would consist of one column of 12 punches. The twelve most frequently occurring codes would be assigned to each of the twelve locations. The next eighteen codes would be accommodated in another column divided into six sections, each of which could accommodate three different mutually exclusive codes. You cannot have a steroid which is both an 11-keto and an 11-hydroxy compound. In actual punched-card application I suspect that one would continue to use the first five columns, at least, for direct codes covering the first 60 most frequently occurring descriptors. If not, another field could be used to accommodate the next 28 codes dividing one or more columns into 4 sections, each containing 3 punches. To accommodate the remaining 359 codes in one field would be quite simple by using all the 495 combinations (binary) of four hole punching patterns possible. The number of columns in the field would depend upon the average number of such codes possible in a single compound. Specific characteristics of existing equipment may modify this decision.

The preceding example of applying measures of information content to the design of an IBM card has been very brief and may not be entirely clear to those not familiar with IBM machines. It is important, at this point, to make clear the similarity between this simple code for an IBM card and a similar code that can be used for a variety of document card or scanning card systems. Let us take up a brief discussion of the qualitative aspects of document cards systems, particularly as they relate to coding.

By document card systems, as contrasted to term card systems, we mean systems wherein all descriptors, or codes for descriptors, are retained together in the particular storage medium involved. Thus, in a punched-card document card system, *i.e.*, McBee, E-Z Sort, IBM, Remington Rand, Underwood-Samas, *etc.*, the holes or perforations are used to encode descriptors assigned to individual documents. In a limited sense, the card *is* the document. Indeed, if the coding were sufficiently elaborate and detailed the card could be the document. The original Luhn Scanner employed an IBM card in which semantically factored words were stretched across the card to form an encoded telegraphic style message. The IBM card employed was the standard 80 column card with a total of 960 punching positions.

Punched-card document card systems have their counterparts in film (Filmorex and Minicard) where again all the descriptor codes are assembled together on a single piece of unitized film. The coding patterns may or may not be exactly of the type found on punched-cards. However, black or white spots correspond to perforations or the lack of perforations. The film-card (microfiche) may also contain a micro image of the original document. Similarly, an IBM card could contain the same micro image in a microfilm insert (Filmsort). Similarly, the Magnacard is the magnetic analog of a punched card. In this case information is coded as magnetized spots on magnetic tape.

The unit-card characteristic common to punched-cards, film cards, and magnetic cards is not only found in document-card systems. The same information found on Magnacards can be stored on continuous magnetic tape. This is done on Univac and the IBM 700 series computers. The mechanisms employed to scan the "card" (sections of tape) are naturally somewhat different. Similarly, the defunct Rapid Selector was a continuous series of Filmorex cards strung out on one reel of film. In the Benson-Lehner Flip system, the Rapid Selector system is partially revived. A compromise between Filmorex and the Rapid Selector was suggested in the AMFIS system by Avakian. The serial counterpart of perforated cards can be found in the Flexowriter tape used at Western Reserve where each document is represented by a series of codes exactly as in the fashion of the Luhn scanner. This is no different from teletype tape except for the number of channels involved and the selector circuitry.

The Zator card is another version of the punched card. The coding method employed has no basic dependence upon the card. It can be used with any type of document card system. Superimposition of codes is employed to make more efficient use of space. I mentioned earlier some of the limitations of Zator coding theory.

There are, obviously, many factors to consider in evaluating document card systems. Cost is one factor, but I believe its relative importance has been overly stressed by Taube and others. Document card systems are not inherently expensive, nor small collections of manual punched-cards. Dr. Whaley has covered more than adequately many other factors which may favor the document-card or scanning card system. He particularly stressed the need, sometimes, to retain relationships between various descriptors. He did not stress adequately the advantages in terms of input convenience and cost, where it is equally advantageous to keep codes together. Preparing a single IBM card is simpler than posting a dozen or more document numbers to individual term cards. It is also simpler than duplicating the same card a dozen times, each to be filed in twelve different file locations.

At the present time, punching a really efficient IBM card is difficult because the IBM machines are not designed for retrieval purposes exclusively. However, in my own experience, preparing elaborately punched cards is not an insurmountable obstacle. Key-punching costs are not considered major problems when a file is used repeatedly. Another factor to consider is searching time for large files. This can be cut down by converting to speedier machines--if time is a problem.

The major criticism of existing document-card systems is the need to operate in a "scanning" sense, i.e., each card or each unit of tape or file must physically pass by a scanning unit. When there are large volumes of records involved very high speeds may be required. This is not only costly, but it will be obvious that there is a limit to the speeds we can reach in mechanically transporting cards, film, etc. It is phenomenal how fast some sorting and scanning devices do work, and possibly these speeds will satisfy most requirements for a long time. However, these speeds are generally available only at a relatively high price. IBM machine rentals are higher in proportion to the speed at which they work, presumably because of greater maintenance and engineering cost. IBM tabulator rentals also vary according to the speed at which they are operated.

An ideal document card system would be one in which the basic advantages are retained--unit record input and storage, logical capabilities, etc. However, one would like to eliminate the need to scan the entire document file, in a physical sense, i.e., by passing cards through a sorter, or magnetic tape past a reading lead, running film by a photoelectric cell. I believe such a system is possible and required particularly if we are to achieve the ultimate in access time. Such a system would be a truly random-access system and not a term card system using so-called random access. Systems such as RAMAC or AMFIS do not appear to be as energy consuming as high speed tape readers or sorters on punched cards, but their mechanical characteristics would seem to be limiting. It is comparable to solving the problem of sorting at high speeds by using a dozen sorters all at once. Similarly to use the equivalent of a dozen magnetic tape readers is no fundamental solution. In the ideal, the file will remain completely stationary and the scanning mechanism will be able to identify the existence of desired codes by scanning in a non-mechanical fashion. An approach in this direction is seen in the Bell Telephone system of routing long distance calls by use of special punched cards. Verner W. Clapp once asked me why you couldn't wave a flashlight at a file and have it throw out the answers. This is not impossible. I have been exploring a similar principle utilizing electromagnetic phenomena which I have called Radio Retrieval.

In conclusion, I have tried to show the fundamental similarities between so-called term card and document card systems by tracing the cyclical evolution of a term card system into a document card system, then into a semi-document card system employing collating methods, and finally back to a term card printed index arrangement. I maintain that the differences between term and document card systems are basically illusory. You will find vigorous proponents for each system depending upon the circumstances. If one had no indexing system at all in the first place, any system is an improvement. Once a system is adopted, thereby improving access to documents, a proposal to merely change the mechanics will not usually excite people.

An area of research which requires more fundamental work is in coding. No matter what system is used, the same amount of information is produced if one uses the same code dictionary and code frequencies.

The Patent Office Steroid Code would be, theoretically, equally efficient with a term card system as in its present document card system. From a practical point of view, it would not. Using Information Theory the coding space required in a document card system can be reduced considerably. It is possible that similar efficiencies are possible in designing term card systems, but these are not yet apparent and may be difficult to find. In other words, term card systems are inherently inefficient because they seemingly cannot take advantage of the variations in code frequencies which are inherent to all information systems. According to Keckley, "there is a central tendency for 90% of the activity to be concentrated within 25% of the classifications." This appears to be well substantiated in the coding of 2,500 steroid patents and independently the coding of 8,500 steroid compounds from the literature. Furthermore, term card space requirements may increase exponentially as the size of the collection grows. A collection of 1,000 documents requires less than 7 bits per descriptor assignment. A collection of 10,000 about 12 bits per descriptor assignment, 100,000 16 bits, and 1,000,000 20 bits.

Mooers deserves credit for recognizing the value of Information Theory for retrieval theory. However, it is just as inefficient to use five punched holes for every descriptor on a document card as it is to use a five digit document number on a term card. By proper application of descriptor probabilities Information Theory can make Zato coding even more powerful.

It has been shown that one can quantitatively measure the amount of information in a document collection by the Shannon formula
$H = -(P_1 \log P_1 + P_2 \log P_2 + \dots P_n \log P_n)$

As a result of this expression, it is concluded that the size of a document collection is no realistic measure of its "information content." Indeed, two collections of entirely different size contain the "same" information if they use exactly the same code or dictionary with the same percentage distribution of descriptors. Thus, in this sense the Library of Congress Subject Catalog contains no more information than the local Public Library Catalog. This may sound startling or ridiculous to librarians. However, as long as the local Library uses the LC Subject Heading Authority List, it may even contain more information because it may add further refinements to the existing LC dictionary or use it with varying frequency assignments. A special library is of more use to its clientele than is the Library of Congress. To alter the information content of a collection one must index in greater depth — not index more documents. This point is most important in industry.

Analysis of the Patent Office steroid code frequencies illustrates in a simple case how Information Theory may be put to use. A brief summary and review of Shannon's Information Theory has been presented to show that the past preoccupation of documentalists with devices is comparable to the earlier preoccupation of communication engineers with machines rather than the information they were transmitting. The main problem in applying information theory in documentation is in defining the "information source" and the "channel." A completely successful retrieval system must combine the advantages of both term and document card systems in such a way that all inertial characteristics of existing systems are removed.