By ROGER PATTRELL BRISTOL

Can Analysis of Information Be Mechanized?

Mr. Bristol is head of the Catalog department, Peabody Institute Library, Baltimore.

THE Communicational Revolution, child of the Industrial Revolution, has created an unwieldy mass of information difficult to record, to analyze and to release in needed amounts and at high speed—difficult, that is to say, because recording, analysis and release have up to now been done by handicraft methods. In most libraries the most complicated mechanical tool used for these purposes is the typewriter, which permits information to be more swiftly recorded. In larger libraries there may be a duplicating machine of some sort also, directed toward the same purpose. Even in the most efficient library systems the whole complex of information-dispensing, surely the most important function of a library, usually employs no more complicated tools than these, and these only for recording information. For analyzing and releasing it libraries everywhere are almost wholly dependent upon hand-brain operation. It is a somewhat ironic reflection on and of our civilization that our swiftest, most analytical tools are devoted chiefly to secondary ends such as statistics and business operations.[1]

Accordingly it is worthy of note that over the past few years considerable attention has been paid to punched cards, both edge-notched and over-all punched, with regard to their use for controlling subject information. The peculiar advantage inherent in punched cards is their capacity for rapidly releasing information by mechanical means, in addition to recording it. The development of the Rapid Selector,[2] which even in its present stage permits by far the fastest release of information yet achieved, intensifies the need of studying the problem of analysis, the only step remaining completely in the handicraft stage.

It needs to be repeated, however, that although immensely fast mechanical tools now exist, in actual fact libraries are still using slow, simple tools; mechanization, even of the processes of recording and releasing information, is rudimentary. It is conceivable that *if* subject analysis could be made amenable to mechanization and *if* the necessary tools could be devised rapidly enough, the whole nexus of information-dispensing in libraries might undergo simultaneous revolution. It is the purpose of the present paper to review current thinking on the double conditions just mentioned. Is subject analysis amenable to mechanization? Can the immensely complex and swift machines that would be required be devised?

Most of those who have considered the need for a fresh look at subject analysis have depended on punched cards rather that the Rapid Selector for illustrations. This in no way invalidates their arguments; on the contrary, the development of the

[1] Hardkopf, J. C. "Cybernetics and the Library." *Library Journal*, 76:999-1001, June 15, 1951.

[2] Shaw, R. R. "Machines and the Bibliographical Problems of the Twentieth Century." In *Bibliography in an Age of Science*, p. 58-71. Urbana, University of Illinois Press, 1951.

selector lends weight and urgency to them.

It is generally agreed that "coding," the term most used by writers on punched cards, is not the same as classification. By some writers classification is tacitly omitted; by others it is specifically condemned, chiefly on the grounds of its being a "rigid network of pathways leading to rigidly grouped collections of item";[3] what is needed is not more pigeon holes but flexible groupings. For example, Shaw believes that today the only reason for classification is a quick, rough look at the holdings of a library.

Holmstrom comments that an obvious coding base would be the Universal Decimal Classification or other established system. But he quickly makes the point that punched-card subject coding should not be dependent on any existing system; coding and classification are alike in that the symbols must be mutually exclusive, unlike in that the coding symbols need not denote position.[4] Current indexes, such as *Chemical Abstracts*, could easily have code numbers assigned to the various headings and thus be mechanized so far as recording and release go. But Perry, for instance, believes that mechanization of conventional indexes and classification schemes "would fail . . . to extract from mechanization the full benefits which it promises."[5] Most efficient use of the new tools will be impossible without study of the relations between machine operations and the basic concepts of indexing and classifying. According to Holmstrom,

> To realize the possibilities of these mechanical methods of literature searching what we need essentially is a philosophy of subject coding. The philosophy of classification has been studied since the time of Aristotle but coding, I submit, is not the same thing, and its theory needs to be thought out.[6]

Perry distinguishes two aspects of coding: (1) The intellectual problem of "discerning which criteria most effectively characterize the subject matter under consideration,"[7] and (2) the mechanical problems of assigning notches or positions. Although the latter problem involves "mathematical analysis based on the theory of permutation and probability,"[8] it is the former which is the more important and the more difficult intellectually.

Similarly, Macdonald refers to the need of a coding scheme

> to determine those ideas in a specific item which might be coded to aid in selection of the item. Not only is such a process extremely long and difficult, but it is dependent upon the background of the person making the choice and so extremely subject to individual variation.[9]

Librarians will recognize this as a good lay description of the problems of subject cataloging. Without using library terminology Perry, Macdonald and Holmstrom see and deplore the pigeon-hole characteristics of classification schemes. With varying degrees of generalization they are expressing the need for fresh thinking on the theoretical aspects of subject analysis because of the impact of mechanized recording and release. To librarians, too often immersed in peripheral routine activities, fresh thinking on what is actually the central problem of

[3] Perry, J. W. "The ACS Punched Card Committee: An Interim Report." *Chemical and Engineering News*, 27:755, Mar. 14, 1949.

[4] Holmstrom, J. E. "Discussion of Punched-card Systems as to Their Application to Library and Technical Work." In Association of Special Libraries and Information Bureaux. *Report of Proceedings of the 22d Conference, 1947*, p.52. London, 1947.

[5] Perry, J. W. "Indexing, Classifying, and Coding the Chemical Literature." *Industrial and Engineering Chemistry*, 40:477, March 1948.

[6] Holmstrom, J. E. "Indexing and other Library Services; Opening Address for Section III, Royal Scientific Information Conference," p.13. Mimeographed.

[7] Perry, J. W., Ferris, Lorna, and Stanford, S. C. "The Use of Punched Cards in American Libraries." In Association of Special Libraries and Information Bureaux, *op. cit.*, p.41.

[8] *Ibid.*, p.42.

[9] Macdonald, J. R. "The Storage of Information—Its Evolution and Future," p.135. Cambridge, Massachusetts Institute of Technology, 1947. Unpublished seminar report in electrical engineering.

information-dispensing will always be welcome.

Most current discussion of the theoretical aspects of subject coding centers about two questions: Minuteness of analysis and the showing of relationships. As Holmstrom[10] points out, logical completeness requires that every significant word in the article or abstract be coded. This atomization of a paper into its component parts, though still impractical, is in line with the demands of researchers, especially in scientific fields. At the University of Montreal, for example, Hans Selye, director of the Institute of Experimental Medicine and Surgery, has indexed many of the papers in his special field under as many as 25 to 50 headings; his private library contains 250,000 entries.[11]

Samain[12] has made several rather extraordinary proposals which involve extremely minute subject analysis of information. He proposes first of all to break down documents into their component concepts. The sum of these concepts will build up "a vocabulary derived from French in which every word will possess 6 letters and which will comprise absolutely every idea, even the most scientific."[13] These six-letter concepts, which appear to be similar to cable-addresses, are then recorded on punched cards by a punch specially designed by Samain "which will produce multiple copies in a single operation and which has an alphabetical keyboard like that of a typewriter."[14] The result is a telegraphically worded abstract of the document on a punched card. Samain claims that use of his suggested vocabulary would allow 60,000,000 concepts to be expressed.

Furthermore, these concepts (*i.e.*, keywords) are punched without the necessity of locating a given word in a predetermined field; any coded item can appear anywhere on the card. Says Samain:

> By judicious organization of important documentation and in particular by the predetermined alphabetical classification of automatically reproduced cards, several million cards can be explored in only a few minutes.[15]

Moreover, by assigning "coefficients" to the keywords, it would be possible to tie concepts together in order to select successively a "word," an "elementary idea." The selector could pick out single words, combinations of words, ideas, or even roots of words!

Samain does not say that this *has* been done. In fact, in his paper it is hard to distinguish past achievement from hypothetical potentiality; since his field is pharmacology, he perhaps somehow evades the ineluctable vagaries of ordinary language. One's first reaction to his extravagance of statement and grandiosity of concept is to write his work off as fantasy. But to do this is to fail to realize the importance of his central idea: The logical need for a mechanized language to permit mechanized analysis which could be fed into mechanized recorders and releasers. In such a language each concept would be as fixed and unvarying as a chemical symbol and would be susceptible of incorporation into "something akin to a grammar of a highly simplified, concise nature."[16]

It is the organic chemists, who deal more with the building blocks of the universe than with vague, transitory and logically unsatisfying verbal concepts, who have gone farthest toward mechanization. G. M. Dyson, an English chemist, has recently de-

[10] Holmstrom, *op. cit.*, p.14.
[11] Letter, July 11, 1949.
[12] Samain, J. "Progrès du classement et de la sélection mécanique des documents: vers une mémoire artificielle." In Fédération Internationale de Documentation, 17th conference, Berne, 1947. *Rapports 1*, p.22-26. La Haye, 1947.
[13] *Ibid.*, p.24.
[14] Holmstrom, J. E. "A Classification of Classifications." In Fédération Internationale de Documentation, *op. cit.*, p.35.
[15] Samain, *op. cit.*, p.25.
[16] Perry, J. W. "New Horizons in Scientific Information Techniques," p.5. Mimeographed, 1949.

veloped a method

for translating molecular structural formulas into a linear set of symbols consisting of letters, digits and punctuation marks. The Dyson symbolism represents the full details of molecular structure and is also amenable to handling on punched cards.[17]

Yet there are rival languages even in chemistry; Dyson's is not the only symbolism capable of translation onto punched cards. More important, even chemical literature, though it consists very largely of these precise structural formulas and of "concepts, such as those of thermodynamics, capable of mathematical definitions," contains also "non-mathematical concepts, whose definition inevitably must involve problems in semantics."[18]

Whatever the field of knowledge, however, whether chemistry, pharmacology or patents, minute analysis, even to the point of coding every significant word, would not be enough, for

you would merely have enumerated the map co-ordinates, so to speak, of the number of points in the country which the author of the article or patent in question had been exploring. You would, so to speak, have pinpointed the most striking features of the landscape he describes. . . . But that is not the same thing as describing the scenery between those points and the reason why the author has proceeded from one point in the landscape to another. . . . It is precisely these *relations between the concepts,* not the concepts themselves, which you want to be able to select from the total mass of literature. . . . With punched cards or the Bush selector or the Univac it is going to be possible, when we have fully worked out the principles of coding, to dispose information in such a way as to enable automatic selection of those items which mention a number of required concepts *connected in a particular way.* We should be able to code not only map co-ordinates of particular points in the field of knowledge, but descriptions of particular types of scenery intervening between those points.[19]

Holmstrom believes that concepts and the semantic territory between them can be recorded and released; he does not say either that the raw material of language can be molded into forms sufficiently rigid to be amenable to mechanical analysis, nor that such a marvelous mechanical analyzer exists or is imminent. Swept away by the tremendous possibilities of punched cards and the Rapid Selector for recording and releasing information, one too easily concludes that it is only a question of time before human intervention will be eliminated from analysis itself.

Suppose we discount for the moment the elusive quality of language. There still remains, alas, the problem of the mechanical analyzer. Obviously, if subject analysis is to be successfully mechanized, some faster tool is needed, some tool comparable in range and speed to the Rapid Selector. One thinks at once of the increasing variety of electronic computers. They are extremely expensive, and so for single libraries now wholly outside consideration. Yet their capacities in numerical computation are so great and their cybernetic possibilities still so untapped that it is impossible not to wonder whether large-scale analysis of subject information will not some day be among their conquests. Ridenour is sure that it can be if the goal is desirable.

One should never ask whether a particular technical goal is possible of achievement; for it always is. The only sensible question is whether the achievement of a given technical goal is justified by economic considerations. . . . It is only a step from [the electronic pencil] to the electronic cataloger, which will read text for itself, recognize key symbols and phrases with which it has been provided, and construct appropriate catalog entries for

[17] Perry, J. W. "The Utilization of Scientific Knowledge." *Scientific Monthly,* 66:416, May 1948.
[18] *Ibid.,* p.414.
[19] Holmstrom, "Indexing and other Library Services," *op. cit.,* p.14-15.

the text it reads. . . . It is probably the steps involved in providing an analytical bibliography which would first engage the attention of an open-minded engineer determined to reduce library costs and raise library efficiency.[20]

Machines are better than brains when they are handling operations which have to be repeated many times rapidly. Conversely, brains are better than machines in nonrepetitive operations which involve a large number of variables. Mechanical recording of information is superior to nonmechanical because it is faster and because it permits rapid duplication. Mechanical searching, particularly by means of the Rapid Selector, is likewise superior in speed and accuracy to handicraft methods. If brainpower is still required for searching, it is because the coding, the analysis, was defective or incomplete; the searching itself does not require brainpower. But analysis does. The brain, as Holmstrom aptly comments, has one property not in any mechanical system: it can be "polarised" when a new subject is presented, so that the relevant data in the brain converge automatically upon this fresh focus. The brain has "what may be termed an automatic adjustment of its cross references."[21]

Electronic computers are capable of dealing with a number of variables, usually numerical. Yet the phases of analysis which they deal with are the repetitive phases; they solve equations at high speed and accurately, but they do not set up the equations. Mathematicians are not put out of their jobs by mechanical marvels. Instead they are given freedom to exercise more fully their analytical abilities. Ridenour, in the passage quoted above, implies the same distinction. Someone must still devise the "key symbols and phrases" which determine the success or failure of the electronic cataloger. The word "electronic" is not magical enough to turn a machine into a human being.

Even if we assume that some day it may be possible for librarians to feed documents into a machine which will analyze them into their intellectual components, it is highly unlikely that such a machine will handle other than the repetitive phases of subject analysis. It is conceivable that a machine can be constructed capable of recognizing and coding the reiterated indivisible concepts, the "map coordinates" (to borrow Holmstrom's phrase) of a document; it is much less conceivable that such a machine could survey and describe the "scenery" between those points, that is, the relations between concepts. The human brain will for a long time yet excel in handling a long series of involved variables.

In short, librarians need not fear mechanization. They need not fear it, in the first place, because mechanical recording and release of information will free them for the more vital, the more intellectual side of librarianship. Just as mathematicians welcome relief from the drudgery of equation-solving, so librarians should welcome the elimination of all possible repetitive operations. Librarians have nothing to lose but their drudgery.

They need not fear mechanization, in the second place, because even the most advanced mechanical brains show little sign of equaling the analytical capacities of their masters. At the very least, librarians may be sure that the more intricate and intriguing the subject information to be analyzed, the longer will that analysis remain unmechanized, a challenge to their best abilities.

[20] Ridenour, L. N. "Bibliography in an Age of Science." In *Bibliography in an Age of Science*, p.26-27. Urbana, University of Illinois Press, 1951.

[21] Holmstrom. "A Classification of Classifications." *Op. cit.*, p.29. Quoted.