Some Characteristics of an Efficient Information Retrieval System*

CLAIRE K. SCHULTZ**
Applied Mathematics Department,
Remington Rand Univac, Blue Bell, Pa.
Received October 13, 1961

This paper discusses some general characteristics of information retrieval systems.

At the onset of designing an information retrieval system there are many unknowns. Some things have to be "played-by-ear," that is, designed in a stepwise fashion. Frequently the assumptions on which a system is designed are proved wrong as systematic analysis of the system's operation brings new insights, or because, once the system begins to operate, its efficiency, glamor, or both, attract many new users. In the beginning, then, one must often plunge in; but it is possible to make some general statements about the manner in which the different elements of an information system should be combined.

A. THE ELEMENTS OF AN INFORMATION RETRIEVAL SYSTEM

1. The Thesaurus.—The thesaurus is the key to the construction and contents of the storage file, and embodies policy decisions about how the file will be built and how it will be used for search. It sets forth not only the "labels" to be applied to the material in the storage file but also the categorical inter-relationships among these. When properly designed and used it ensures that decisions are made consistently. Constructing the thesaurus is the most difficult intellectual task in designing a total system.

Though perhaps not recognized as such, probabilistic considerations enter into thesaurus construction. If there is very little chance of the file being searched for tertiary alcohols as a class, but a good chance that the file will be searched for alcohols of a particular chain length, with particular substitutions on the chain, the builder of the thesaurus should know and act on such statistics if he is to organize the file to meet user demands efficiently. Of course, he may not have the necessary statistics for deciding some of the issues that arise. This fact should be recognized, and the thesaurus should be kept flexible enough so that it may be modified when such statistics are obtained after the system starts operating.

2. The Storage File.—The storage file is only a means to an end—retrieval. Methods of file organization are dictated by the combination of input methods and output strategies. Therefore considerations of the storage file usually are reduced to considerations of physical media for storage and suitable formats for manipulating the data. The particular equipment used for searching plays a large part in determining what the storage media and data formats will be.

3. The Search Strategy.—In manual systems the search strategy is a product of the mind of the human making the search. In machine systems the search strategy is set by the group of instructions, or program, given to the machine.

We know that a human being approaching a card catalog to perform a search involving, for example, the concept *engineering* uses many alternative search plans almost subconsciously. By contrast, if a machine search of the file is begun for the concept *engineering*, the machine will not carry out alternative actions unless specifically instructed to do so by its program. As a result, it may retrieve far too many items; the search may be too time-consuming, or the machine may find no references pertinent to the question, even though, if the question had been phrased a little differently, references would have been forthcoming.

Whether a retrieval system employs humans or machines, a search strategy is needed that delievers only a "reasonable" number of references, should the file be very rich in items related to the desired information. On the other hand, the strategy should include a provision to inform the inquirer of the "next-most-pertinent" content of the file in the event of a null answer to the inquiry.

4. The Implementation of the Search.—In card catalogs, the brain of the human searcher translates the search strategy into action. The analogous capacities of machines are used. Evaluating the comparative efficiencies of machines for retrieval is almost as complex as comparing the skill of humans at these tasks. There is, as yet, too little experience with most machines to judge their capabilities by records of the machine's past performance. One must largely entrust such evaluation to persons who know how to translate into machine specifications the requirements set for a system with respect to such things as: (a) size of the file, (b) search strategy, (c) storage medium (media) and format(s), (d) frequency of use, (e) speed with which search results are required, and (f) physical format(s) and medium (media) of the output. It can be seen that the choice of equipment for implementing a system ideally post-dates basic decisions about most of the other elements of the system. If equipment is selected early, the freedom of choice with respect to variables in other elements of the system becomes more limited.

5. The Results of the Search (Output).—The output of retrieval systems can be in various formats—lists of identification numbers, titles of documents, abstracts and summaries of data, or printed indexes. Output media can also vary, *e.g.*, punched cards, magnetic tapes, microfilms and photocopies. The requirements for output formats and media are important to consider in the designing of the system, since possible formats and media depend not only on what equipment is chosen, but also on the input formats and the storage media. Some systems require more than one output format and more than one output medium.

*Presented before the Division of Chemical Literature, ACS National Meeting, St. Louis, Mo., March, 1961.

**Presently associated with the Institute for the Advancement of Medical Communication, 33 East 68th St., New York 21, N. Y.

6. The Measurement of System Efficiency.—For many decades reference librarians have wondered exactly how well they understood what their patrons really wanted, whether they were finding every shred of information available on the subjects for which they searched, and whether their catalogs and references books could be better organized. This kind of introspection reflects a concern for efficiency which necessarily precedes any improvement in system design.

All well-designed information retrieval systems need to provide a capacity for analyzing the system's functioning. The more sophisticated systems, that is, those using electronic computers, can perform many kinds of self analysis while they are searching. This is an advantage of automation that should not be overlooked. It is possible, for example, to keep a record of what terms from the thesaurus were actually used in searches, the frequency with which they were used, the other terms with which they were combined, the number of items in the file described by particular combinations of terms, and so on.

B. GENERALIZED FLOW-CHART FOR INFORMATION RETRIEVAL SYSTEMS

A diagram developed for another paper[4] illustrates general aspects of the process of making a search—whether it is documents that are to be retrieved or else information from a file of raw laboratory data, a file of structural formulas, a file of structural fragments, or even a personnel file about the employees of a laboratory (Fig. 1). Both the thesaurus and storage file for the system are assumed to have been built already. The steps postulated in this diagram, and enumerated below, hold for every type of implementation, from a manual system, through punched card systems and electronic computer systems, and for a wide range in the size of the collection and the number of inquiries to be handled.

Explanation of Figure 1.—First the search must be formulated. This requires choosing the "labels" of the file that will combine in a specified logical relationship to provide the information sought. Because some information retrieval systems require the batching of inquiries so that more than one question can be answered while scanning the storage file, a block on the diagram is provided for making up the search file. A generalized scheme must accommodate the flexibilities provided in some systems and not others, so the next block symbolized the step where some systems specify choices among the available variables. In the language of the systems analyst, which you see on the diagram, a choice is an option; a variable is a parameter.

1. The section from 1 to 2 in the top line of the diagram is a special courtesy to some machine systems. An electronic computer has to check routinely to find whether it has finished its work.

The section in the second line from 2 to 3 represents the operations in one general search strategy. These operations can be combined into a number of different patterns. Notice that some of the blocks on the diagram are drawn with broken lines. This means they are optional. The dotted oval in the center of this section provides the option of setting an upper limit on the number of references to be supplied in answer to an inquiry. If the machine finds more than this number, it can be instructed to drop the search until the question can be restated (next run of the computer).

3. The third line of the diagram is an operation in which a human or a computer keeps track of what is found on a scratch pad, or its electronic counterpart.

4. The section in the fourth line from 4 to 5 is where the answers to the batched inquiries are separated according to the individual inquiries in the batch. Also the option of a lower limit on the number of references desired for the search can be exercised here if alternate searches were set up to meet this possibility.

5. Either of the two operations of the section immediately after 5 in the fourth line may be adopted to obtain the search results from the memory of a computer in readable form. The last block provides for putting out statistical information about the work that the machine has just performed.

C. ANALYSIS OF INDIVIDUAL SYSTEMS IN TERMS OF THE GENERALIZED SCHEME

The two retrieval systems with which I am most familiar are at Merck Sharp and Dohme and at ASTIA. They are good examples because they differ considerably. Since both systems are described in the literature,[1,2,3] I will consider only a few of their more general characteristics.

The general characteristics of the Merck Sharp and Dohme system are: one search is performed per pass of the file; the file to be passed is usually 50,000 punched cards or less (one card per document). The machine used is the IBM 101 Electronic Statistical Machine. Upper and lower limits on the number of references desired in answer to an inquiry are set by human intervention. The only possible machine output is a set of punched cards with the serial number of a document printed on it. Supplementary human effort provides fuller information. Any statistics about the functioning of the system must be kept by hand. The corresponding general characteristics of the ASTIA system are: approximately ten searches are performed per pass of the file. The file to be processed is approximately 300,000 document descriptions recorded on magnetic tape according to the "inverted" system. The machine used is the Remington Rand Univac Solid State 90 Computer with tape components. It is possible to set automatic upper and lower limits on the number of references desired in answer to an inquiry. The output is usually a set of abstracts, printed on card stock, but other formats and media can be provided. Statistics about the functioning of the system can be compiled by the computer.

Figure 1 indicated in terms of the general scheme for information retrieval which operations are manual and which are automated for each of the systems, and also which operations are either non-pertinent or only indirectly pertinent to each of the systems. Though this figure gives only skeletal information about each system, it reveals much more than one would learn from spending an equivalent period of time (to that of reading the figure) trying to get a total picture of either or both of the systems by any other method. The figure also provides a framework for filling in additional information about either or both systems.

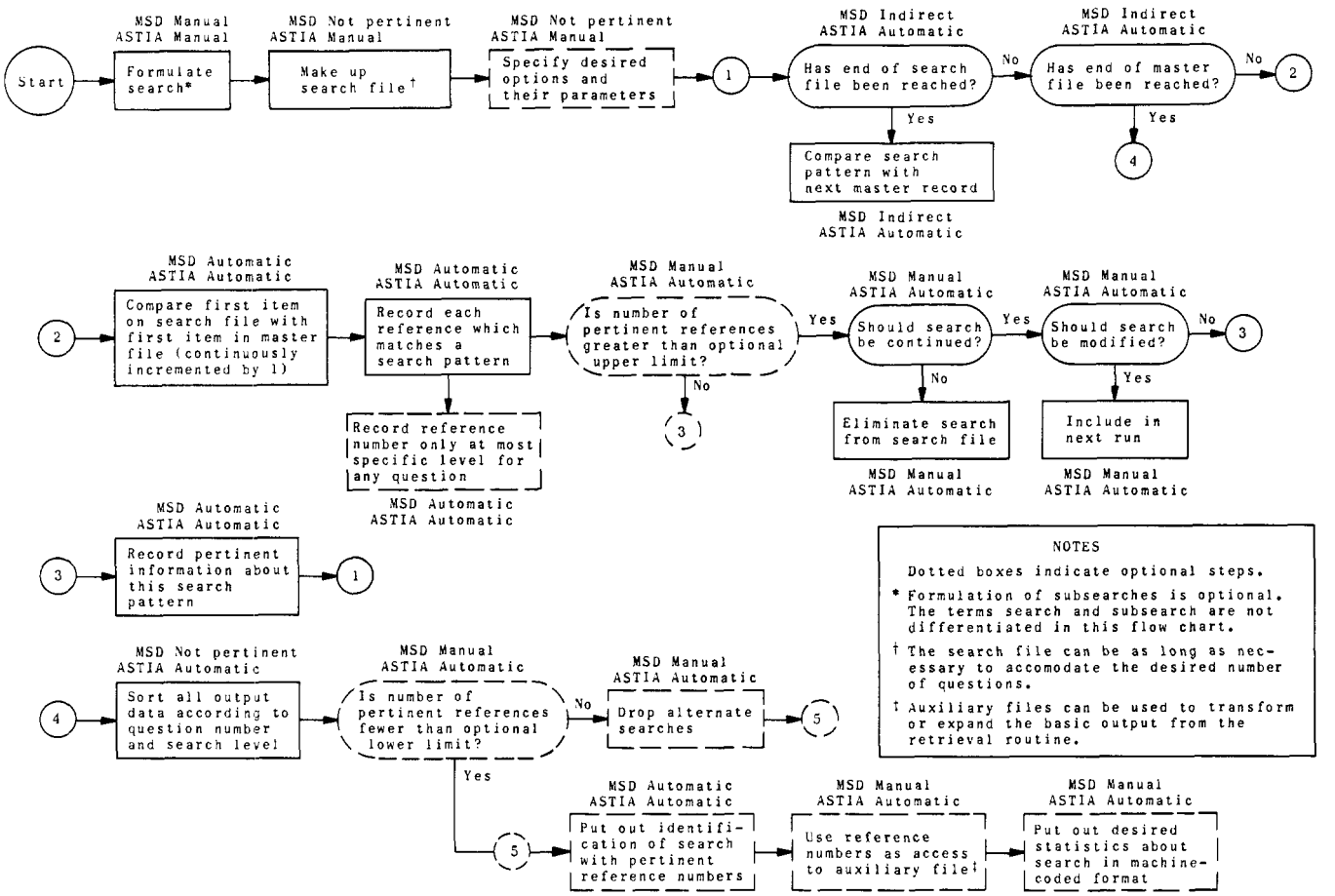

Fig. 1.—Generalized flow chart of a computer program for information retrieval. The figure also compares the relative automation of two retrieval systems.

D. VALUE OF THE GENERALIZED SCHEME FOR PLANNERS AND ADMINISTRATORS OF INFORMATION RETRIEVAL SYSTEMS

Just as the generalized flow-chart was used to follow and compare the operations in the two above systems, it can be used by persons concerned with other systems, extant or not. The diagram may also be useful for comparing, in a gross manner, the automatic *versus* non-automatic features of two different machines for implementing the same system. For example, one machine may have a large enough memory to keep statistics about the functioning of the system while it is operating and a second may not, or one machine may use only one medium of output, while a second may provide several media. By attaching quantitative significance to the various blocks in the diagram one might compare additional capacities of machines or systems: for example, the number of searches that can be performed during one pass of the file, or the time necessary for passing the file to perform searches.

The utility of such a diagram may stimulate further attempts to characterize objectively the processes of information storage and retrieval. The more objective we can become about our activities, the easier it will be to compare our work and our systems, to design new and more efficient systems, and train the new people needed for the field.

BIBLIOGRAPHY

(1) National Science Foundation, "Non-Conventional Technical Information Systems in Current Use, No. 2, September, 1959, pp. 24-26.

(2) Schultz, Claire, K., "An Application of Random Codes for Literature Searching," in Casey, Robert, and Perry, James, "Punched Cards," 2nd ed., Reinhold Publishing Corp., New York, N. Y., 1958.

(3) ASTIA, "Controlling Literature by Automation," Fourth Annual Military Librarians' Workshop, October 5-7, 1960, Washington, D. C.

(4) Schultz, Claire K., A Generalized Computer Method for Information Retrieval, to be published in *Journal of the Association for Computing Machinery*.