tors, introduced in this paper, should be utilized where appropriate. Care should be taken that the resulting contraction is neither identical nor almost identical with a regular contraction and will not be taken as a special contraction for some other group.

(5) No regular contractions for the heterocyclic rings are presented in this paper. Where their contraction is imperative, it appears best to omit as few letters as possible. Thus, QUINOLINE might be progressively reduced to QUINOLIN, QUINOLN, and QUINLN, and even to QINLN; PYRAZOLE might be reduced to PYRAZOL, and PYRAZL, and even to PYRZL.

REMARKS

In this paper a set of practices has been described for regularizing the transcription of chemical names into capital letters, arabic numerals, and four special characters, namely, the ampersand, the asterisk, the hyphen, and the slash, and also for contracting the resulting names so that they can be accommodated within a fixed field. The authors invite criticism of the proposals and will welcome any suggestions as to possible improvements or extensions. (A single alphabetically arranged list of the names and contractions presented in this paper is available from the authors upon request.)

LITERATURE CITED

(1) Barnard, A. J., Jr., Kleppinger, C. T., Wiswesser, W. J., *J. Chem. Doc.*, 6, 41 (1966).
(2) "J. T. Baker BATCH Directory," J. T. Baker Chemical Co., Phillipsburg, N. J., Sept. 1965, 39 pp.
(3) Wiswesser, W. J., Abstracts, 124th National Meeting of the American Chemical Society, Chicago, Ill., Sept. 1953, p 18G.
(4) Wiswesser, W. J., "Literature Sources of Mammalian Toxicity Data, with Special Emphasis on Tabulating Machinery Applications," Advances in Chemistry Series No. 16, American Chemical Society, Washington, D. C., 1956, p 64.
(5) Wiswesser, W. J., "A Line-Formula Chemical Notation," Thomas Y. Crowell, Co., New York, N. Y., 1954, 149 pp; Smith, E. G., Wisswesser, W. J., *ibid.*, 2nd ed, in preparation.
(6) Benson, F. R., Abstracts, 124th National Meeting of the American Chemical Society, Chicago, Ill., Sept. 1953, p 5G.
(7) Wiswesser, W. J., unpublished data.

Some Information Indexing Techniques in a Real-Time Hospital Computer System

RICHARD A. BOLT
Bolt Beranek and Newman Inc., Cambridge, Massachusetts 02138

SCOTT I. ALLEN
Medical Care Administration Branch, U. S. Public Health Service, Washington, D. C.

and JOHN W. WEBB
Massachusetts General Hospital, Boston, Massachusetts

Received September 15, 1965

A phonetic indexing technique has proved useful for computer retrieval of filed information, especially in those cases where misspellings in the retrieval request can readily occur. Further, the creation of special indexes to filed information can facilitate rapid selection on very large files.

INTRODUCTION

A cooperative research effort to develop a real-time Hospital Computer System is being undertaken by the Massachusetts General Hospital and Bolt Beranek and Newman Inc. The project is supported by the National Institute of General Medical Sciences and the American Hospital Association. The system uses a Time-Shared computer (the computer is a modified PDP-1, by Digital Equipment Corp., Maynard, Mass.), with remote input-output devices allowing rapid real-time collection, storage, retrieval, and dissemination of hospital information. It is hoped that the developing system will prove both a powerful adjunct to the medical and nursing staff in carrying out their tasks, and a significant aid in clinical research.

In this paper we describe two techniques that facilitate reference to information in files stored in the bulk memory of a computer system. The first is an adaptation of the Russell "Soundex" encoding scheme to provide an index through which a user of the system may reach a filed item despite misspellings. The second is an extension of the use of indexes to provide rapid selection of a population of interest from records of a large file. We need first, however, to give a brief description of the computer-based communication system on which these techniques have been applied.

SYSTEM DESCRIPTION

The hardware consists of a central Time-Shared computer which is connected to Teletypes ("Teletype" is a registered trademark of the Teletype Corp., Skokie, Ill.) at remote locations. The term "Time-Shared" means that the computer's time is allocated (by a program called the Executive) among the independent programs (User programs) associated with the various Teletypes. The central computer has a bulk storage in the form of a UNIVAC Fastrand magnetic drum on which some 60 million characters of information may be filed. It is through this common store of information that the User programs communicate.

The hardware is located on the premises of Bolt Beranek and Newman Inc., in Cambridge, Mass., and the terminal devices, Model 33 Teletypes in noiseproofing enclosures, are located at the Massachusetts General Hospital in Boston, a distance of about 4.5 miles. The hospital terminals communicate over direct current commercial telegraph lines. Members of the programming staff in Cambridge have office terminals for individual access to the computer.

A growing library of User programs designed to provide specific information-handling functions to the hospital are stored on the large drum, to be loaded upon request from a Teletype terminal. A number of User programs dealing with the functions of patient admissions, medication ordering, and test ordering are being run at the Massachusetts General Hospital on a limited experimental basis.

CREATING THE DRUG FORMULARY

One of the User programs, the Drug Formulary Update Program, enables the hospital pharmacist to create and maintain a file of drug information. Through a question–response sequence on the pharmacy Teletype terminal, the pharmacist enters information concerning a particular drug, such as permissible routes of administration and comments concerning preparation of the drug. The body of information is collected in the computer memory and organized compactly. When all of the information for the drug has been entered, the pharmacist commands the program to write out the information on the bulk storage drum. Thus the pharmacist may build up a file of drug information covering the entire stock of drugs in current active hospital use.

This file of drug information can be automatically referred to by an operator in the hospital who enters a medication order for a patient by using the Teletype terminal assigned to a particular Patient Care Unit. A drug entry in the formulary may be found by typing in the name of the drug, or one of its names if it has not only a generic name but also synonyms or legitimate abbreviations. For example, the operator may refer to aspirin by typing the name "ACETYLSALICYLIC ACID," "ASPIRIN," or the letters "ASA". Now, many drugs bear names that may be misspelled even by experienced personnel. This fact coupled with very high volume of referral to the drug formulary in the course of the hospital day presents an interesting retrieval problem: How can we find an entry for a drug even though the name we have been given is misspelled?

Let us examine one approach to that problem, namely the use of a phonetic indexing technique.

The Phonetic Index. We remarked that when the pharmacist had typed in all the information concerning a drug, he then typed a command to the program to cause the drug to be filed. This filing action consists of writing the item out on the large bulk storage drum and of filing the location at which the item was written in a special index file. The indexing procedure is carried out by a program that performs a phonetic transform of the drug name to a numeric combination. That is, each phonetic element of the drug name is classified according to its sound, and a number corresponding to the sound class is substituted for the letters. (A detailed description of the Russell Soundex alphabet is contained in Soundex Brochure, Form No. LBV-809, obtainable from Remington Office Systems Division, Sperry Rand Corp., 122 East 42nd St., New York 17, N. Y.) For example, the words "CAT" and "KAT" would transform phonetically into the same index code. Both letters "C" and "K" are encoded as the number 2. The encoding procedure ignores vowels and would bypass the letter "A," and would then assign the code number 3 for the letter "T." Thus, both "CAT" and "KAT" would be transformed to the code "23." The word "RAT," however, would transform differently as the letter "R" has a different sound. In this case, the "R" would transform to the code 6, producing the configuration "63." In passing we might remark that the letter "C" is assigned the same encoding value as the letters "S" and "K," contemplating both the hard and the soft "C."

The success of this transform is a consequence of the fact that a person will usually try to spell a word or name on the basis of how it sounds when he is unsure of the correct spelling.

Retrieval *via* the Index. To look up a drug then, the computer performs a phonetic transform of the character string that is the name of the drug we are trying to get. Then the index file is inspected to see if there are any drug item addresses filed under that particular code. If there are, the addressed items are read one by one, and a character-by-character comparison is carried out to detect a direct match between the name of the desired drug item and the names in the file. If no direct match is found, the fact that one or more drug items were indexed under that particular phonetic code combination suggests that we are trying to retrieve a drug item by a misspelled name. The names of the drugs phonetically similar to the name we entered can be typed out by the computer as alternative suggestions to what might have been meant.

By choosing a phonetic encoding of the names and alternate names of drugs in a file we provide two benefits: first, rapid localization of the field of search; second, a technique which contributes to effective cooperation between a near-perfect machine and a sometimes erratic human being.

The Drug Formulary File is a particular example within the Hospital Computer System of a file that is created on-line, and whose elements are the object of subsequent selective reference. Let us now turn our attention to some aspects of a system for general file generation and retrieval.

ON-LINE FILE GENERATION

A basic function of the Hospital Computer System is to facilitate research based on patient-care records. Some of the information making up these records flows into the computer as a by-product of the day-by-day entry into the computer of observations and tests on hospital patients; other data may be collected by special laboratory or observational techniques as part of individual research inquiries.

Common hospital files or individual private research files may be established by hospital investigators directly from Teletype terminals without the assistance of special programming personnel. Such research data files may be designed to handle both existing punched-card information of a retrospective nature and additional prospective data coming into the system—during future time periods. In the latter instance, data are often typed directly into the computer without the intermediate generation of punched cards.

On-line entry of research observations is advantageous in two important ways. First, the accuracy and consistency of information may be checked according to the definitions and criteria established when the file was initially created. Secondly, the investigator has the facility to examine and analyze the new measurements immediately, and to correlate the new data with previously collected data. The researcher deals with his data by means of retrieval programs. Let us see what this involves.

Ease of Language. A major objective in the design of a hospital-oriented retrieval system was the creation of an interrogation language as close to natural English as possible. In particular, it was considered highly desirable to avoid the repetitive use of punched card terminology such as column locations and numeric codes. To achieve this goal, the researcher initially designates English titles or field names for each class of information; for example, PATIENT NAME, AGE, SEX. These titles or names are organized into a dictionary.

Further, in the case of data entered on punched cards in coded numeric form, it is possible to transform such information to more easily referable alphabetic abbreviations. For example, the month "01" on the card becomes "JAN," the month "02" becomes "FEB," and so forth.

Stating the Retrieval Request. The following example illustrates the manner by which an investigator may set a task for the retrieval program. A typical request for hospital case data might call for the selection of all male patients with the diagnosis of diabetes treated with insulin. Such a search request begins with establishment by the researcher of a list of descriptors such as those shown below.

Descriptor	Tag
SEX = MALE	MALE
AGE > 40	OLD
AGE NOT > 40	YOUNG
DISEASE = DIABETES	DIABETIC
DRUG NAME = INSULIN	INSULIN

A tag is set up for each descriptor for ease of reference. As any descriptor most likely reflects an individual researcher's immediate interest, no permanent dictionary of these descriptors and tags is maintained.

Once such descriptors are defined and given a "tag" name, one indicates the subsets of the file that are of special interest. In our example, the researcher is interested in the male patients who are diabetic and are receiving insulin treatment. Further, he desires to divide this subset of the entire patient population into two further subsets on the basis of whether the patient is or is not over forty years old. Accordingly, his two study groups would be described by the following logical combinations of the previously entered descriptors, referring to the descriptors by their "tag" names:

Sample 1: MALES AND OLD AND
DIABETIC AND INSULIN

Sample 2: MALES AND YOUNG AND
DIABETIC AND INSULIN

The specifications of samples to retrieve are compiled into programs which begin the designated searches.

FIELD INDEXING FOR RAPID SEARCH

To speed up searches on large files with thousands of records, the researcher has the ability to index certain fields so that the computer may locate groups of records without resorting to inspection of every item of the file. In the above illustration it is reasonable to suppose that the diagnosis field would be indexed for rapid access to certain disease groupings. If drug preparations were also indexed, the search time would be reduced even more, since only records indexed on both fields would have to be scanned to look for patients of a specified age and sex.

The files of selected items themselves take the form not of items but rather of an index to the selected items which requires far less room to store than would the selected items themselves. Since the existence of selected items is frequently the only fact of interest for statistical purposes, it is often possible to operate on this index rather than on the items of the file.

Once relevant records have been isolated by the Search Program, a library of easy-to-use statistical output programs is available for call-up from a Teletype. Hence, the results of retrieval may be easily tested for statistical significance and the routine measures of variability, such as standard deviations, computed.

This research retrieval system, currently in experimental usage, permits the flexible entry of data, allows use of data editing routines, facilitates high-speed access through field indexes, and offers the user a simple English-like language with which to interrogate and manipulate computer files.

SUMMARY

We have discussed in this paper an approach to the rather difficult task of real-time retrieval of information filed away against complex character strings, in this instance rapid access to drug information by way of drug name. The approach has been a phonetic indexing procedure, which not only detects spelling errors, but actively seeks to be helpful and suggestive to the user.

An experimental research-oriented retrieval system of programs has been described wherein the researcher has been afforded the ability to converse with the computer in a reasonably natural language, in terms which the researcher himself has stipulated. In this system, the possibility of indexing items by selected information fields permits the much more rapid search technique of matching lists of item addresses, as opposed to linear item-by-item scans.

The Time-Shared system, with rapid program intercommunication, further permits the researcher to send the tabular results of his searches to an on-line, real-time mathematical program, with which the researcher may then analyze his data.

Rapid Structure Searches *via* Permuted Chemical Line Notations. IV. A Reactant Index

ALAN GELBERG
Diamond Alkali Company, T. R. Evans Research Center, Painesville, Ohio 44077

Received December 22, 1965

Previous publications in this series have discussed the concept of permuted Wiswesser chemical line notations (*1*) and methods of preparing a permuted index (*2*, *3*). The applicability of this technique has been extended to develop a reactant index. The line notations of the chemical reactants that formed the chemical products were added to the punched cards which contained the notations of the reaction products. These notations easily fit within a 60-column field of the 80-column punched card. Also, there is available space, in this field, for indicating the catalyst and the preparation conditions. There is no need to permute these last two items, but they can appear in the index as part of each notation entry.

It has been found that an input of 1050 punched cards containing the notations of the products and reactants generated a permuted record count of 10,068, or 9.6 lines of print per punched card entry. This is in contrast to 5.5 lines for the products alone, as previously reported (*3*), which had included the R symbol (phenyl ring) as well as an experimental use of maximum contraction (*4*). For this study, the "Revised Rules" (*5*) were followed. The R symbol was not made a separate line entry. Also, after the entry of the first G symbol (chlorine atom) in a notation card, additional G symbols were not entered as separate line entries. However, both the R and multiple G symbols were included in the Quick-Scan area (*3*). The R and G symbols are the most frequently occurring groups in this file and have been found to be useless index search terms. Wiswesser analyzed the frequency of occurrence of symbols in 66,660 notations of chemical structures and reported the R as being the most frequently occurring symbol after the space, while the G symbol placed 19th out of 41 characters (which had included 26 letters, 10 numbers, 4 symbols, and the space). However, it appears that he included locants in this table as well as multipliers (*6*). All of the G and R symbols appearing in the notations are of value in the Quick-Scan area and serve a useful purpose for browsing before reading the total notation. The Quick-Scan symbols are now alphabetized, as suggested by Sorter (*7*), rather than being listed in their order of appearance in the notations.

For the forementioned deck of cards, the card-to-tape input time was 78 minutes for the IBM 1401. The computer sorting time was 35 minutes to alphanumerically organize the 10,068 records. Listing, to create the index, took 22 minutes. At an estimated cost of $52.00/hour operation time, each line entry cost about $0.013, or approximately $0.125 for each punched card.

The new index has enhanced the utility of this overall program since it now allows the user to rapidly locate all products from specific or similar starting materials. This is in addition to locating all compounds having the same functional group and rings (other than phenyl) as well as locating specific and similar structures. It is also of value in organizing composition of matter patents. The possibility of further developing this method to a reaction